

# A METHYLATION-TO-EXPRESSION FEATURE MODEL FOR GENERATING ACCURATE PROGNOSTIC RISK SCORES AND IDENTIFYING DISEASE TARGETS IN CLEAR CELL KIDNEY CANCER

JEFFREY A. THOMPSON<sup>1</sup> and CARMEN J. MARSIT<sup>2</sup>

<sup>1</sup>*Program in Quantitative Biomedical Science, Geisel Medical School at Dartmouth College, Lebanon, NH 03756, USA*

<sup>2</sup>*Department of Environmental Health, Rollins School of Public Health at Emory University, Atlanta, GA 30322, USA*

*E-mail: carmen.j.marsit@emory.edu*

Many researchers now have available multiple high-dimensional molecular and clinical datasets when studying a disease. As we enter this multi-omic era of data analysis, new approaches that combine different levels of data (e.g. at the genomic and epigenomic levels) are required to fully capitalize on this opportunity. In this work, we outline a new approach to multi-omic data integration, which combines molecular and clinical predictors as part of a single analysis to create a prognostic risk score for clear cell renal cell carcinoma. The approach integrates data in multiple ways and yet creates models that are relatively straightforward to interpret and with a high level of performance. Furthermore, the proposed process of data integration captures relationships in the data that represent highly disease-relevant functions.

*Keywords:* prognostic; survival; cancer; data integration; eQTL; m2eQTL; m2eGene

## 1. Introduction

The recent abundance of large datasets of diverse molecular features have vastly increased our knowledge of cellular processes disrupted in disease; yet, these datasets, taken individually, have frequently failed to reveal useful biomarkers for complex diseases, such as cancer [1, 2].

Despite the clear utility of individual ‘omic’ datasets, such as gene expression, DNA methylation, copy number alteration, etc., in better understanding disease etiology and in some cases providing useful prognostic or predictive value [3], it is equally clear that each of these data types can only capture part of the disease signature in a cell. Therefore, interest has been growing in more holistic methods, which integrate data of different types. As of yet, these approaches have met with mixed success. For example, a study of long-term survival in patients with glioblastoma multiforme (an aggressive form of brain cancer), found that joint regression of different types of data did not improve predictive accuracy [4]. Another study across five different cancer types came to a similar conclusion [5]. Nevertheless, a more nuanced approach, based on integrating separate models built from individual datatypes for ovarian cancer outcomes did show a higher predictive accuracy for integration across datatypes [6].

A recent review of data integration approaches classified them as falling into one of two broad categories: multi-stage and meta-dimensional integration [7]. Multi-stage integration techniques are currently the most developed and wide-spread. These involve using separate analyses of multiple types of data, with the results from one data type used to filter, and presumably increase the power of, another. The most commonly used example of multi-stage integration is expression-quantitative trait loci (eQTL) analysis, wherein single nucleotide

polymorphisms (SNPs) are associated with changes in gene expression, which in turn are associated with disease [8, 9]. Meta-dimensional techniques consist of integrated models, in which all data are used as part of a joint model or analysis, which might involve joint regression, or integration at the level of individual models [10, 11].

Important prognostic information may in some cases be obscured by noise. However, it is much less likely that noise will obscure that information from different types of data for the same features. For example, it may be that repression of a gene promoter through DNA methylation represents a disease state. Nevertheless, that gene's expression may be altered in healthy individuals through alternative regulation. Therefore, it may not be enough to capture the gene expression data alone. Furthermore, one type of data may capture nascent information of disease progression that is not yet apparent in other data types. In some cases, there may not be one superior type of data for predicting prognosis. Finally, there may be informative interactions between data types that are not possible to assess when using only one type of data. For these reasons, we hypothesize that an appropriate multi-omic data-integrated approach will create superior prognostics to those using only a single data type.

In this work, we developed a data integration approach for combining gene expression data with DNA-methylation to create prognostic models for clear cell renal cell carcinoma (the most common form of kidney cancer). Our data integration approach is a hybrid method combining both multi-stage and meta-dimensional elements but results in a model that is easily interpreted by those familiar with traditional statistical approaches. Furthermore, it is amenable to extremely high dimensional data but runs quickly compared to other methods. We demonstrate the viability of our approach in the context of creating prognostic markers for kidney cancer and compare it to two other methods that have proven successful in this context: random survival forests and penalized Cox regression [12–14].

We chose to integrate DNA methylation and gene expression because they have proven to be prognostically useful data sources for a number of cancers [12] and are highly related. DNA methylation controls tissue specific expression of genes. Therefore, if we can exploit this redundant information, we may be able to create a more informative prognostic model. Furthermore, it has long been suspected that aberrant DNA methylation itself is related to carcinogenesis [15], although only recently has evidence begun to mount for a causative role [4, 16]. Given that DNA methylation tends to be a more stable mark than gene expression [17, 18], in certain cases it may be informative where gene expression is not. In cancer, hypermethylation of gene promoters silences tumor suppressors and other genes throughout the genome [19]. Hypomethylation of other regions is associated with genomic instability [20]. Thus, disruption of DNA methylation patterns may be a potentially relevant etiological factor, which could increase the utility of our approach.

## 2. Methods

We used M2EFM, and two other approaches, to model overall survival in clear cell renal cell carcinoma. For the main analysis, gene expression and DNA methylation profiles from untreated, resected tumors for patients with clear cell renal cell carcinoma were created by The Cancer Genome Atlas (TCGA) project [21] on the Illumina HiSeq 2000 sequencing and

Illumina Infinium HumanMethylation450 platforms respectively. RNA-seq data normalization was performed by TCGA and normalized data were downloaded from the UCSC Cancer Genomics Browser [22] (Table 1). The RSEM normalized read counts were  $\log_2$  transformed by the UCSC, and we left them in that form. DNA methylation data were obtained from the National Cancer Institute’s Genomic Data Commons. These were functionally normalized using the `minfi` package [23, 24] for the R statistical environment [25].

A separate smaller dataset of methylation profiles (from the same platform) was also used by our method to identify differentially methylated loci between 46 paired tumor and tumor-adjacent normal clear cell kidney cancer samples obtained through the National Center for Biotechnology Information’s Gene Expression Omnibus (GSE61441) [26]. Again, we used functional normalization for these data.

Table 1. Distribution of Samples in TCGA Clear Cell Renal Cell Carcinoma (Clear Cell Kidney Cancer) Data

	RNA-seq (%)	450k (%)	Overlap (%)
Samples w/ overall survival data	525	311	310
Male	341 (64.95)	201 (64.63)	201 (64.84)
Female	184 (35.05)	110 (35.37)	109 (35.16)
Stage I	262 (49.90)	150 (48.23)	150 (48.39)
Stage II	56 (10.67)	30 (9.65)	30 (9.68)
Stage III	126 (24.00)	75 (24.16)	74 (23.87)
Stage IV	81 (15.43)	56 (18.01)	56 (18.06)
Grade 1	12 (2.29)	7 (2.25)	7 (2.26)
Grade 2	228 (43.43)	132 (42.44)	132 (42.58)
Grade 3	202 (38.48)	119 (38.26)	119 (38.39)
Grade 4	75 (14.29)	49 (15.76)	48 (15.48)
Grade X	5 (0.95)	2 (0.64)	2 (0.65)
Missing Grade	3 (0.57)	2 (0.64)	2 (0.65)
Deaths	166 (31.62)	99 (31.83)	98 (31.61)
Mean Age	60.65	61.43	61.48

There was no evidence of significant differences in the distribution of staging or tumor grade for cases in the RNA-seq and DNA-methylation data ( $\chi^2$  test,  $p = 7.77e-01$  and  $p = 9.54e-01$  respectively). For all data types, there were 8 cases missing survival data, with 5 having no clinical annotation at all. The remaining 3 were female, had a mean age of 70.33 years, and contained 2 stage I and 1 stage II tumors. Other than the 5 with no clinical annotation, there were no samples missing on clinical predictors, therefore we decided to remove the 8 samples missing outcomes from the analysis.

Beta values were transformed into M-values [27], and we removed probes on the X or Y chromosomes, containing SNPs [28, 29], or with cross-hybridization issues [30]. Finally, probes with values missing for greater than 50% of samples were removed and the remaining values were imputed using the k-nearest neighbors method, with  $k=10$ , from the `impute` package [31, 32] for R.

## 2.1. M2EFM

We developed a data-integrated modeling approach we call Methylation-to-Expression Feature Model (M2EFM). The basis of this approach is to find loci that are differentially methylated between matched pathologic and non-pathologic data and to associate those loci with significant differences in gene expression in the disease state. The process is analogous to expression quantitative trait loci (eQTL) analysis, except that instead of associating SNPs with changes in gene expression, we associate differentially methylated loci. The loci are then called m2eQTLs (for methylation-to-expression QTLs) and the genes are called m2eGenes.

The approach consists of five primary steps (summarized in Fig. 1):

- (1) **Filtering probes and genes for variability.** Gene expression values were filtered to remove very low variability genes (usually genes with no expression) by removing genes with a median absolute deviation of .05 or less, leaving 16907 genes. Methylation probes were filtered to remove those with a median absolute deviation of less than 0.8 (after transformation to M-values). This left 27700 probes for the kidney cancer data.
- (2) **Identifying differentially methylated loci.** Differential methylation was identified using the empirical Bayes method from the `limma` package [33] for R. We used 46 paired tumor and tumor-adjacent normal samples from a separate dataset than used in the rest of the analysis. This initial step was used to identify which loci to focus on. We passed the 500 CpG loci with the lowest adjusted p-values (Benjamini-Hochberg) for differential methylation on to the the next step.
- (3) **Identifying methylation-to-expression quantitative trait loci (m2eQTLs).** m2eQTL analysis involves associating methylation levels at the loci identified in the previous step with gene expression levels genome-wide. In terms of an eQTL analysis, the proportion of methylated alleles for a particular loci is equivalent to the genotype at a single nucleotide polymorphism (SNP), although it is a continuous, rather than discrete value. Identification of m2eQTLs was performed using the `MatrixEQTL` package [34] for R, which builds linear models to test association in a computationally efficient manner. In this way, the M-value of probes in the training data that were found to be differentially methylated in the first step were tested for their association with gene expression patterns in both *cis* and *trans* in a manner analogous to that used in typical eQTL analysis. An m2eQTL was defined to act in *cis* if it was associated with a gene within 10000bp, otherwise it was defined to act in *trans*. The top 150 *cis* and *trans*-m2eQTLs (by effect size) and their associated m2eGenes were passed on to the next step. This number was simply chosen to identify around 200 relevant genes and may not be optimal.
- (4) **Building integrated models from m2eQTLs and m2eGenes.** From the previous results we built a joint regression model across both probes and genes involved in the m2eQTLs. Given that these were bound to have collinearity, to prevent overfitting we used Cox regression with Ridge penalty [35]. The linear predictor from the Cox model was used as a molecular risk score for all training samples (see Supplementary File 1, <http://dx.doi.org/10.5061/dryad.b1t61>).
- (5) **Integrating clinical variables.** M2EFM uses a second regression to integrate clinical variables. For this step, we performed an unpenalized Cox regression on the molecular risk

score from the previous step and the values of clinical variables. This allows the hazards in the model to be more interpretable and keeps the clinical covariates from being penalized. In a typical Cox proportional hazards model, there is a rule of thumb that there should be no more than about 10 events in the data per variable in the model. Each training dataset in our data will have about 69 events (depending on the split of the data), meaning the model should have only about 7 variables. Clinical variables used for cancer prognosis vary but can include TNM staging, tumor grade, AJCC stage, patient sex, and age at diagnosis. We tried a few alternative clinical models on the training data only and picked the one with the highest discrimination (measured by concordance index, Table S1). Although the results were close for TNM staging and AJCC stage (the difference was significant at  $p = 1.04e-05$ ), TNM staging would add 17 variables to the model and AJCC stage only 4, so our final model includes patient age at diagnosis, sex, tumor stage, and risk score. Although this is 8 variables, relaxing the rule to 9 events per variable has been shown to be acceptable [36] and can moreover be judged to some degree from our results.

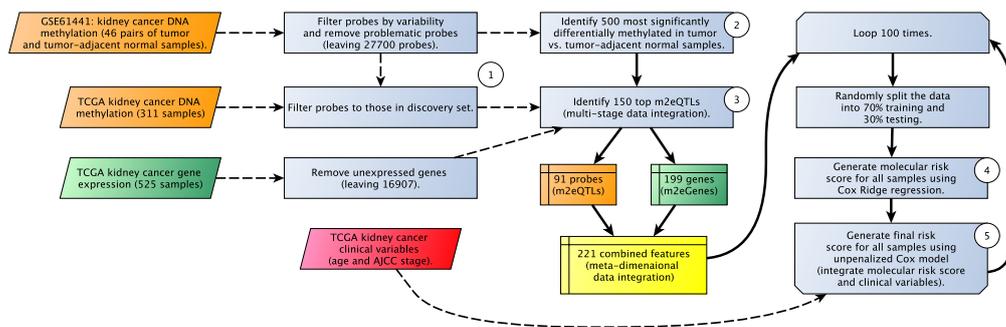


Fig. 1. Workflow for M2EFM analysis of clear cell renal cell carcinoma.

## 2.2. Experimental Design

We built 100 different M2EFM models of overall survival in clear cell kidney cancer for 100 different random splits of the data, using 70% training and 30% testing data sets. This process was repeated for different combinations of data (clinical variables only, gene expression only, methylation only, expression and clinical variables, and methylation and clinical variables).

The results of our approach were compared with two other methods that have previously been shown to successfully integrate molecular and clinical data to generate prognostic markers: penalized Cox regression and random survival forest [12] (although that work did not attempt molecular data integration). We used Cox Ridge regression, rather than LASSO (which was used in [12]), because it generally has better predictive performance. The model was built using the `glmnet` package [37] for R, and the lambda parameter was found using 10-fold cross validation for each split of the data. The random survival forest was built using the `randomForestSRC` package [38] for R. The run time of the random forest prevented cross validation of the parameters, so these were left at the defaults, as in [12]. The performance

of the models was evaluated using concordance or C-index, a commonly used measure of discrimination in prognostic models. The C-index is a measure of how likely it is, in any given pair of individuals, that the individual with the higher risk score has the event first.

### 2.3. *Functional Analysis Approach*

Although it is not a requirement that the genes used in a prognostic model are functionally related to the disease, models built from functional relationships can reveal important insight into why one patient might have a better prognosis than another, which can lead to improved treatment decisions and a higher probability of model validation. Therefore, we performed a functional analysis of the gene set used in our model. The m2eQTL genes were used to perform a gene set network enrichment analysis using the online tool WEB-based GENE SeT AnaLysis Toolkit (WebGestalt) [39] to identify genes in our gene set that were enriched in sub-networks of protein-protein interactions that were, in turn, enriched for biological functions. We also used it to perform enrichment analysis for GO biological process terms. For both of these analyses we required at least 5 genes to overlap the gene module or pathway.

Our goal with this work was to demonstrate a method by which a biomarker can be identified. We do not identify a specific gene and DNA methylation probe set, in part because an independent validation dataset would be required.

## 3. Results

### 3.1. *M2EFM Prognostics*

The m2eQTL phase of M2EFM identifies differentially methylated loci that are associated with changes in gene expression throughout the genome. An example is shown in Fig. S1.

For the M2EFM-based risk score, the median C-index over 100 random splits of the data of the score from combined clinical and molecular variables (M2EFM Exp+Meth+Clin) reflects the highest prognostic accuracy of any method or data type used at .792. The median C-index of the risk score from clinical variables alone (M2EFM Clin) was .776 and the median C-index of the risk score from molecular variables alone (M2EFM Exp+Meth) was .702 (Fig. 2). The improvement in C-index for the combined clinical and molecular model over the clinical variables alone was significant at  $p = 4.25e-06$  by two tailed Wilcoxon signed-rank test.

The M2EFM expression without methylation models had only slightly lower accuracy than models built using both data types. For these models, the median C-index for the combined clinical and expression models (M2EFM Exp+Clin) was .791 and for the expression only models (M2EFM Exp) was .703. The improvement in C-index for M2EFM Exp+Clin over the clinical variables alone was significant at  $p = 1.50e-08$ .

The M2EFM methylation without expression models were not as accurate as the other M2EFM models. The median C-index for the combined clinical and methylation models (M2EFM Meth+Clin) was .755 and for the methylation only models (M2EFM Meth) was .643. In this case, the clinical variable model had generally stronger C-index values than M2EFM Meth+Clin at  $p = 2.068e-08$ .

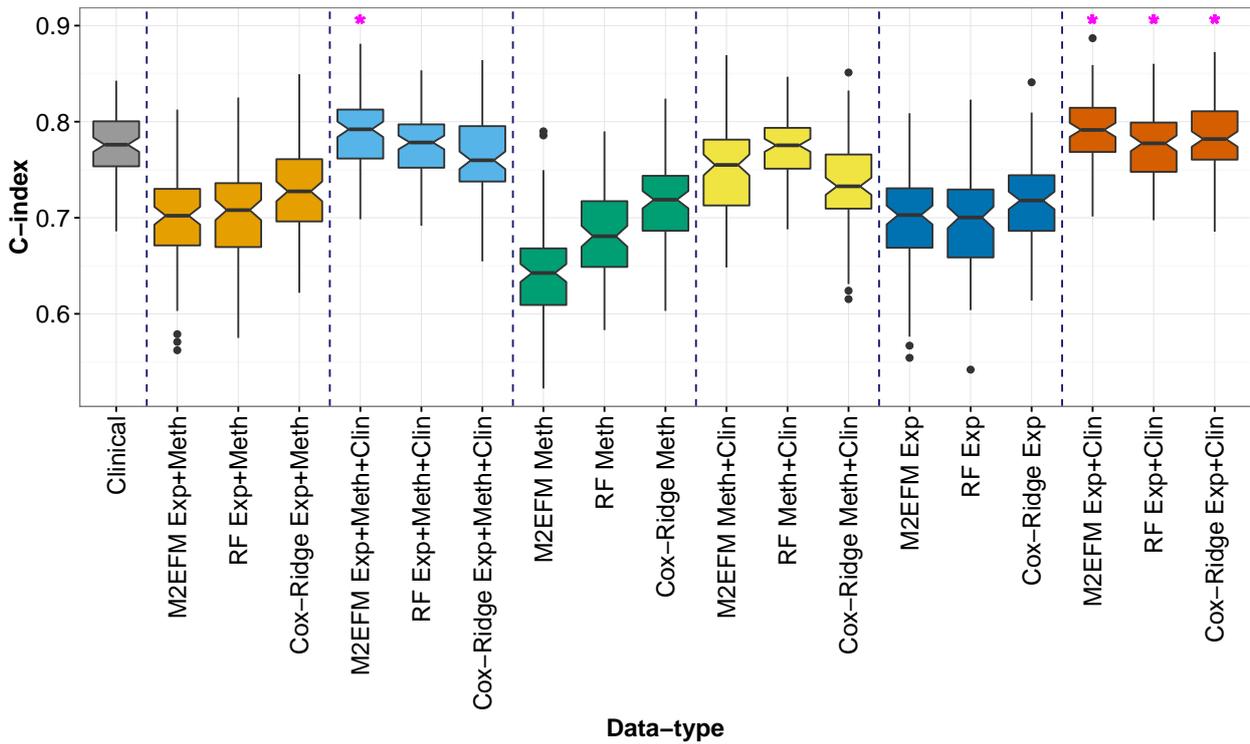


Fig. 2. C-index across 100 random splits into training and testing data of the various approaches. If one method was significantly better than another than the notches in the box plots will not overlap. For convenience, if a method resulted in significantly better results than clinical data alone, it is marked with “\*”.

### 3.2. Random Survival Forest Prognostics

Random survival forest was not as effective at exploiting the integrated expression and methylation data as our guided M2EFM approach. The median C-index for the combined clinical and molecular features (RF Exp+Meth+Clin) over the same 100 random splits of the data was .776 and the median C-index of the models built from the molecular data alone (RF Exp+Meth) was .696. The addition of the molecular data using random survival forest models was no more discriminatory than the clinical variables alone.

The performance of the expression without methylation random survival forest model was similar to the model with both data types. The median C-index for RF Exp+Clin model was .777, which was very slightly but significantly stronger than the clinical only model ( $p = 7.16e-03$ ) while the median C-index for the RF Exp model was .694.

The performance of the methylation without expression random survival forest model was slightly worse than when both data types were used. The median C-index for the RF Meth+Clin model was .776, but the RF Meth model was a significant improvement over M2EFM Meth ( $p = 8.05e-13$ ) and had a median C-index of .682.

### 3.3. Cox-Ridge Prognostics

Cox regression with ridge penalty [40] outperformed M2EFM when it came to the molecular data alone, but its molecular risk score was less independent of the clinical variables, thus

its accuracy for the full model was less than that of M2EFM. The median C-index for the combined clinical and molecular features (Cox-Ridge Exp+Meth+Clin) of the same 100 random splits of the data was .760 and the median C-index of the models built from molecular data alone (Cox-Ridge Exp+Meth) was .727 and was improved over M2EFM Exp+Meth ( $p = 3.10e-05$ ).

The performance of Cox-Ridge Exp+Clin model was slightly worse than the M2EFM Exp+Clin model ( $p = 9.14e-13$ ) with a median C-index of .782. Again, the performance of the molecular data only model, Cox-Ridge Exp, was somewhat better than M2EFM Exp ( $p = 2.35e-06$ ) with a median C-index of .718.

Finally, the Cox-Ridge Meth+Clin model did not perform as well as the M2EFM model. It achieved a median C-index of .735, which was significantly worse than the M2EFM Meth+Clin model ( $p = 6.37e-15$ ). Nevertheless, the Cox-Ridge Meth model, with a median C-index of .705, performed better than the M2EFM Meth model ( $p = 1.62e-12$ ).

### 3.4. *Comparison to Yuan et al.*

A direct comparison of our approach to that used in [12] on the same data was not possible, because the data they deposited included only the pre-filtered DNA methylation values, which did not include the same probes we identified in our discovery set. Nevertheless, we attempted to run our method on this subset of probes (which necessarily created different models than those used above). The highest mean C-index of any method listed in [12] on the kidney cancer data as .767 for a model including microRNA and clinical variables. On the same data (normalized by Yuan et al.), we achieved a mean C-index of .775 for the M2EFM Meth+Exp+Clin model and a mean C-index of .773 for the M2EFM Exp+Clin.

### 3.5. *Functional Analysis*

#### 3.5.1. *Gene Set Network Enrichment*

Next we performed gene set network enrichment analysis using the online tool WebGestalt, requiring a minimum of 5 genes to overlap a gene module. All significant results (after multiple testing correction) are shown in Table 2. The full list of genes found in each pathway is given in Supplementary File 2. This approach revealed enrichment for gene modules associated with immune response, proliferation, and other functions. As an example, a portion of the largest sub-network our model was enriched in (which is enriched for the JAK-STAT Cascade) is shown in Fig. 3 (visualized using Cytoscape [41]). The genes from our gene set are shown in green and are highly connected nodes in the network.

#### 3.5.2. *Biological Process Enrichment*

We further tested the straight enrichment for biological process terms in the Gene Ontology using our gene set (without network enrichment), again requiring a minimum of 5 genes to overlap a pathway. The results in Table 3 show the top 5 most enriched GO terms, with a clear enrichment for immune system related genes. The full list of genes enriched in each pathway is given in Supplementary File 3.

Table 2. Protein Interaction Network Module Enrichment

Pathway	Observed	Expected	Adj. p
T Cell Costimulation	7	.33	2.69e-07
Regulation of Defense Response to Virus by Host	11	1.18	2.69e-07
JAK-STAT Cascade Involved in Growth Hormone Signaling Pathway	34	19.64	3.80e-03
Complement Activation	6	1.82	3.43e-02

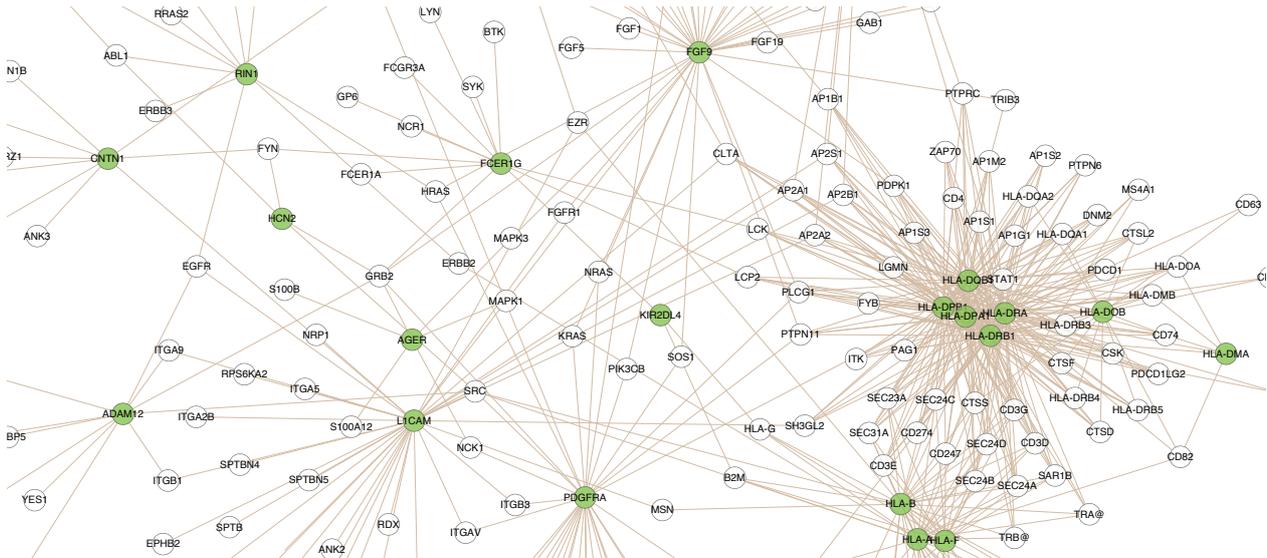


Fig. 3. Portion of the gene module enriched for the JAK-STAT cascade. Genes from our gene set are shown in green.

Table 3. Enriched for GO Biological Process

Pathway	Observed	Expected	Adj. p
Antigen Processing and Presentation of Exogenous Antigen	13	2.04	1.60e-05
Antigen Processing and Presentation of Exogenous Peptide Antigen	13	2.01	1.60e-05
Antigen Processing and Presentation	15	2.6	1.60e-05
Response to Interferon-Gamma	11	1.30	1.60e-05
Cellular Response to Interferon-Gamma	10	1.07	1.60e-05
Exogen	13	2.15	2.09e-05
Interferon-Gamma-Mediated Signaling Pathway	9	.89	2.09e-05
Antigen Processing and Presentation of Peptide Antigen	13	2.23	2.87e-05
Immune Response	32	12.24	2.95e-05

#### 4. Discussion

All of the approaches described show it is possible to attain a meaningful level of prognostic discrimination using a joint regression on both gene expression and DNA-methylation values, if collinearity is properly accounted for. However, our approach, which first identifies dysregu-

lation of DNA methylation in cancer, then associates that dysregulation to differences in gene expression, and finally builds prognostic markers from genes and CpG loci that are associated with this loss of regulation, was able to build models with a higher level of prognostic discrimination than either a random survival forest approach or Cox regression with Ridge penalty as well as a model built from traditional clinical variables. The results for our joint molecular regression with M2EFM were about the same as using expression data alone, leaving it unclear if this form of meta-dimensional integration is helpful on top of the multi-stage integration, which selected the features, thus more work on this part of the approach is needed.

The median C-index of .792 achieved by our M2EFM Exp+Meth+Clin model was the most accurate predictor of overall survival achieved by any approach in this study. This result was achieved through three data integrations, including different types of molecular data, as well as clinical variables. Notably, we showed that M2EFM's combination of a molecular risk score with clinical variables was a significant improvement over the clinical variables alone. Furthermore, our m2eQTL analysis identified 199 genes with high relevance to clear cell kidney cancer, without *a priori* knowledge of those genes' association to the disease. In fact, one of the top results from our gene set network enrichment analysis was for the JAK-STAT Cascade pathway, which is a known factor in kidney cancer progression [42]. That we identified this pathway by associating differentially methylated CpGs with differences in gene expression may suggest a role for dysregulation of methylation in the development of the disease, although caution in this interpretation is warranted, due to the cross-sectional nature of our study. An additional limitation was our lack of an independent dataset containing samples with gene expression and DNA methylation profiles as well as clinical data for validation.

The high enrichment we observed for genes involved in the immune system may indicate the utility of our approach in identifying survival differences based on dysregulation of immune functions. Given that immunotherapy has emerged over the last several years as an important component of kidney cancer treatment [43] and the pressing need for biomarkers that can identify the patients that will benefit from treatment [43], further development of this approach may be warranted in this regard. Another interesting result was our identification of *CA9*, which is currently of interest as a possible serum biomarker for kidney cancer [44], as a potential target for radioimaging [45], and as a potential therapeutic target [46]. Taken together, our results suggest that our approach is able to identify functionally relevant, and not just prognostic, genes. This is promising in terms of eventual validation of our approach.

Most of our results were better than those in a recent study including kidney cancer prognostics [12], but in a couple of cases, either the random forest or the Cox-Ridge approach did not perform as well as the methods in that work. However, they used fewer samples in that study and included inferred cancer subtypes from non-negative matrix factorization (NMF), in addition to gene and probe level measurements. Using only the DNA methylation and gene expression data from that study, which handicapped our method in discovery, M2EFM still showed slightly higher discrimination than any other approach. However, our goal was to develop a method based primarily on feature selection, rather than transformative dimensionality reduction techniques, in order to reduce the complexity of the models. Although interpretability is still limited by our use of Cox Ridge regression in generating the molecular

risk score, it is over a limited number of genes that appear to be functionally related, mitigating this issue. It is notable that our m2eQTL-based approach creates models that outperform those using NMF, through a motivated feature selection technique that selects for putative regulatory relationships. We also note that Cox-Ridge in most cases outperformed the Cox-LASSO approach used in [12], and in some subsets of the data performed slightly better than M2EFM for prognostic accuracy. However, this accuracy comes at the cost of interpretability. The Cox-Ridge models contain thousands of genes or probes, telling us little in terms of the function of prognostic genes and creating unwieldy biomarkers in terms of real world use.

## 5. Conclusions

We developed a new data-integrated approach to modeling cancer prognostics and applied it to clear cell renal cell carcinoma data. M2EFM uses both a multi-stage data integration that links changes in methylation between tumor and normal tissues to levels of gene expression, and a meta-dimensional data integration that combines DNA methylation and gene expression values as part of a joint regression for outcome prediction. M2EFM was shown to identify not only prognostic, but functionally relevant features that may be associated with therapeutic response and that were highly connected in relevant protein-protein interaction networks.

## 6. Acknowledgements

We would like to acknowledge our funding sources: NIH-NIMH R01MH094609, NIH-NIEHS R01ES022223, and NIH-NIEHS P01 ES022832/EPA RD83544201 and a Hopeman award from the Norris Cotton Cancer Center.

## References

- [1] M. Huang, A. Shen, J. Ding and M. Geng, *Trends Pharmacol Sci* **35**, 41 (2014).
- [2] S. E. Kern, *Cancer Res* **72**, 6097 (2012).
- [3] A. S. Coates, E. P. Winer, A. Goldhirsch, R. D. Gelber, M. Gnant, M. Piccart-Gebhart, B. Thürlimann, H.-J. Senn, F. André, J. Baselga *et al.*, *Ann Oncol* **26**, 1533 (2015).
- [4] J. Lu, M. C. Cowperthwaite, M. G. Burnett and M. Shpak, *PLoS ONE* **11**, p. e0154313 (2016).
- [5] L. Xu, L. Fengji, L. Changning, Z. Liangcai, L. Yinghui, L. Yu, C. Shanguang and X. Jianghui, *PLoS ONE* **10**, p. e0142433 (2015).
- [6] D. Kim, R. Li, S. M. Dudek and M. D. Ritchie, *BioData Min* **6**, p. 23 (2013).
- [7] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim, *Nat Rev Genet* **16**, 85 (2015).
- [8] A. C. Nica and E. T. Dermitzakis, *Phil Trans R Soc B* **368**, p. 20120362 (2013).
- [9] R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. De Haan, A. I. Su *et al.*, *PLoS Genet* **4**, p. e1000232 (2008).
- [10] E. R. Holzinger, S. M. Dudek, A. T. Frase, S. A. Pendergrass and M. D. Ritchie, *Bioinformatics*, p. btt572 (2013).
- [11] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine and C. Sander, *PLoS ONE* **6**, p. e24709 (2011).
- [12] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao *et al.*, *Nat Biotechnol* **32**, 644 (2014).
- [13] G. Ambler, S. Seaman and R. Omar, *Stat Med* **31**, 1150 (2012).
- [14] F. R. Datema, A. Moya, P. Krause, T. Bäck, L. Willmes, T. Langeveld, B. de Jong, J. Robert and H. M. Blom, *Head Neck-J Sci Spec* **34**, 50 (2012).

- [15] M. Ehrlich *et al.*, *Oncogene* **21**, 5400 (2002).
- [16] D.-H. Yu, R. A. Waterland, P. Zhang, D. Schady, M.-H. Chen, Y. Guan, M. Gadkari and L. Shen, *J Clin Invest* **124**, 3708 (2014).
- [17] J.-P. Issa, *J Clin Oncol* **30**, 2566 (2012).
- [18] P. W. Laird, *Nat Rev Cancer* **3**, 253 (2003).
- [19] M. Esteller *et al.*, *Oncogene* **21**, 5427 (2002).
- [20] K. L. Sheaffer, E. N. Elliott and K. H. Kaestner, *Cancer Prev Res (Phila)* **9**, 534 (2016).
- [21] Cancer Genome Atlas Research Network *et al.*, *Nature* **499**, 43 (2013).
- [22] M. S. Cline, B. Craft, T. Swatloski, M. Goldman, S. Ma, D. Haussler and J. Zhu, *Sci Rep* **3** (2013).
- [23] J.-P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood and K. D. Hansen, *Genome Biol* **15**, p. 1 (2014).
- [24] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry, *Bioinformatics* **30**, 1363 (2014).
- [25] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2016).
- [26] J.-H. Wei, A. Haddad, K.-J. Wu, H.-W. Zhao, P. Kapur, Z.-L. Zhang, L.-Y. Zhao, Z.-H. Chen, Y.-Y. Zhou, J.-C. Zhou *et al.*, *Nat Commun* **6** (2015).
- [27] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou and S. M. Lin, *BMC Bioinformatics* **11**, p. 587 (2010).
- [28] 1000 Genomes Project Consortium *et al.*, *Nature* **491**, 56 (2012).
- [29] L. Butcher, *Illumina450ProbeVariants.db: Annotation Package combining variant data from 1000 Genomes Project for Illumina HumanMethylation450 Bead Chip probes*, (2013). R package version 1.1.1.
- [30] Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson and R. Weksberg, *Epigenetics* **8**, 203 (2013).
- [31] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, *Bioinformatics* **17**, 520 (2001).
- [32] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, Imputing missing data for gene expression arrays (1999).
- [33] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, *Nucleic Acids Res* **43**, p. e47 (2015).
- [34] A. A. Shabalina, *Bioinformatics* **28**, 1353 (2012).
- [35] A. E. Hoerl and R. W. Kennard, *Technometrics* **12**, 55 (1970).
- [36] E. Vittinghoff and C. E. McCulloch, *Am J Epidemiol* **165**, 710 (2007).
- [37] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, *J Stat Softw* **39**, p. 1 (2011).
- [38] H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, *Ann Appl Stat*, 841 (2008).
- [39] J. Wang, D. Duncan, Z. Shi and B. Zhang, *Nucleic Acids Res* **41**, W77 (2013).
- [40] H. Zou and T. Hastie, *J R Stat Soc Series B Stat Methodol* **67**, 301 (2005).
- [41] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res* **13**, 2498 (2003).
- [42] S. Li, S. J. Priceman, H. Xin, W. Zhang, J. Deng, Y. Liu, J. Huang, W. Zhu, M. Chen, W. Hu *et al.*, *PLoS ONE* **8**, p. e81657 (2013).
- [43] M. W. Ball, M. E. Allaf and C. G. Drake, *Discov Med* **21**, 305 (2016).
- [44] M. Takacova, M. Bartosova, L. Skvarkova, M. Zatovicova, I. Vidlickova, L. Csaderova, M. Barathova, J. Breza, P. Bujdak, J. Pastorek *et al.*, *Oncology Lett* **5**, 191 (2013).
- [45] P.-C. Lv, J. Roy, K. S. Putt and P. S. Low, *Mol Pharm* **13**, 1618 (2016).
- [46] J. Tostain, G. Li, A. Gentil-Perret and M. Gigante, *Eur J Cancer* **46**, 3141 (2010).