Two microRNA Signatures for Malignancy and Immune Infiltration Predict Overall Survival in Advanced Epithelial Ovarian Cancer

Ilya Korsunsky[1], Janaki Parameswaran[2], Iuliana Shapira[3], John Lovecchio[4,5], Andrew Menzin[4,5], Jill Whyte[4,5], Lisa Dos Santos[4,5], Sharon Liang[4,5] , Tawfiqul Bhuiya[4,5], Mary Keogh[1], Houman Khalili[1], Cassandra Pond[1], Anthony Liew[1], Andrew Shih[1], Peter K. Gregersen[1,5] , and Annette T. Lee[*1,5]

[1]Feinstein Institute for Medical Research, Manhasset, NY USA
[2]Yale University, New Haven, CT USA
[3]SUNY Downstate Medical Center, Brooklyn, NY USA
[4]Northwell Health, Manhasset, NY USA
[5]Hofstra-Northwell School of Medicine, Hempstead, NY USA
[*]Correspondence and requests for materials should be addressed to ATL (ALee@northwell.edu)

Methods

Validation on TCGA data. Direct validation of our mathematical prognostic model was limited by the lack of public datasets with matched expression data assayed with the same platforms we used. To overcome this limitation, we developed a validation strategy that preserved the qualitative, biological hypothesis from our model and fit the quantitative details to an external testing dataset. In this way, we were less constrained by incompatible platforms and focused on the biological contributions of our findings.

The full details of this pipeline are described in the supplementary below.

This strategy consisted of four steps. First, we identified the microRNAs and enriched pathways from the training module we chose to validate. Second, we performed an unsupervised, outcome agnostic CCA on the validation dataset. Third, we filtered the resulting validation set modules with a functional gene enrichment analysis. Finally, we tested the performance of the selected validation module with standard survival analysis tools. Below, we describe these steps in more detail.

1) *Feature selection.* From the module we have selected to validate, we identified the non-zero weight microRNA and the pathways enriched by the non-zero weight mRNA. We did not select the non-zero weight mRNA directly, because the pathways are more likely to replicate across studies. That is, we needed to account for the possibility that different genes may target the same pathways to affect the same biological functions. Thus, we selected all the genes identified in the enriched pathways, even if they have zero weight in our training module.
2) *Unsupervised, semi-sparse CCA.* A key component of our primary analysis was the microRNA/mRNA correlation structure, as is it used to weigh the contributions of individual features. We retained this aspect in the validation pipeline by performing CCA to identify new, highly correlated microRNA and mRNA components. Since our focus in

this analysis was to test the predictive ability of our module, we could not include the outcome information into the CCA. Thus, we performed an unsupervised CCA to avoid artificially biasing the model towards the survival outcome. Finally, we assigned an appropriate sparsity penalty to the mRNA matrix so that the number of genes in the resulting testing modules roughly matched the number of genes in the training module.

3) *Enrichment Filtering.* Although we started with all genes pertaining to some set of pathways, the sparsity in step 2 ensures that each validation module only contains a portion of the original genes. In fact, it is possible for the validation modules to have no enrichment or enrichment in pathways unrelated to those chosen in step 1. Thus, we performed an additional enrichment on the validation module genes to make sure that the selected genes enrich for similar pathways as our original module.
4) *Survival Outcome.* At this point, we had identified preliminary modules within the validation set that contained the same microRNAs and similar pathways as the module we were testing. Now we could test how well these modules were associated with the survival outcome, using two standard approaches. First, we performed a univariate Cox proportional hazards regression on each component of the module. We used the likelihood-ratio test to gauge the significance of the resulting model. Second, we stratified the subjects into equally sized groups based on the module component's median, defining a high risk and a low risk group. We tested the differential survival outcome of these two groups using the log-rank test.

Further Discussion on the Implication of Different microRNA Expression Platforms. In this study, we used a qPCR assay to measure microRNA expression, as opposed to the TCGA, which used a microarray platform. Qualitatively, these platforms should agree on gross expression levels. That is, highly expressed microRNA should be highly quantified in both and vice versa. However, the sensitivity to edge cases (e.g. low molecular copy number), different types of technical noise, as well as the relative scale between the numerical outputs can be different between the two platforms. The last point refers to areas of linearity (i.e. twice the assay ouput reflects twice the true molecular expression number) and non-linearity in the expression values. The primary implication of these difference is that there is no mathematical function that can transform a quantitative expression level in one platform into a quantitative expression level in the other. Thus, a mathematical model designed to read in numbers from one assay will fail if it is fed numbers from another assay. For this reason, we relearned an unsupervised model for the TCGA validation cohort, instead of attempting to reuse the model learned with our EOC cohort.