Prognostic values of cross-omics screening for cancer survival

Dimitrieva Slavica and Rehrauer Hubert

Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Introduction

Large-scale molecular profiling of cancers offers a great potential to advance our understanding of the development and progression of this disease. Systematic cancer genomics projects, like The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), have applied high-throughput genome analysis techniques to generate genomics, transcriptomics, epigenomics and clinical data for several cancers. These data can be informative for multiple aspects ranging from discovering of new markers for more accurate cancer diagnosis and prognosis, to development of new therapeutics and personalized treatments.

The overall goal of our study, as a response to one of the CAMDA 2015 Challenges, is to gain novel biological insights into three less well studied cancers: Lung Adenocarcinoma (LUAD), Kidney Renal Clear Cell Carcinoma (KIRC) and Head and Neck Squamous Cell Carcinoma (HNSC). We performed a systematic analysis of genome-wide molecular datasets provided from the ICGC Data Portal (miRNA, mRNA and protein expression, somatic copy-number variation (CNV) and DNA methylation profiles) to investigate underlying mechanisms of cancer initiation and progression. Cancer is an extremely complex disease and it is of no surprise that previous genomics analyses have revealed extensive tumor heterogeneity[1]. As consequence, the identification of molecular signatures from genomics analyses that can give accurate prediction and prognosis of response to therapy is still a major challenge. In the last few years, extensive efforts have been made to incorporate diverse molecular information for better prognosis and treatment plans[2,3]. However, due to the high cost of large-scale molecular profiling, in practice clinicians are mainly focusing on a small number of selected genes or are using only single-platform genomic data. Therefore, with our study we want to understand how and to what extent different molecular profiling data can be useful in cancer diagnosis and prognosis. Using miRNA and mRNA expression, somatic copy-number variation, DNA methylation and somatic mutation profiles we have identified genes that are frequently altered in each of the selected cancers and are linked to patient survival. Some of the biological markers that we identified have already been reported in previous studies, but few of them are yet to be examined. In addition, we assessed which of the molecular dataset, as a standalone platform is the most informative for patient diagnostic and survival prediction.

Results

Molecular signatures for discrimination between normal and cancer tissues

First, we were interested in finding molecular signatures that can discriminate neoplastic from normal tissue in the selected cancer cohorts. For this purpose, we used a classification approach based on LASSO regression model[4]. In this analysis. only molecular data from normal tissue that is adjacent to primary tumor was used; the molecular data from blood derived normal tissue was not considered in order to avoid building models based on genes that can discriminate between blood and the corresponding solid tissue (lung, kidney or head/neck). The classification performance of the selected models was measured using the AUC ("Area Under Curve") statistic, which can be interpreted as a probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example[5]. The AUCs values of the selected models for discrimination between normal and cancer populations range from 0.95 – 1.00 (see Table 1). Almost perfect performance can be reached easily, which suggest that there are radical molecular changes in cancerous cells compared to normal cells. Interestingly, the best (and perfect) classifier performance was achieved based on DNA methylation data for the LUAD and KIRC cohorts (Table 1). It is a well-known fact that DNA methylation can alter the expression of genes and several recent studies have shown that it also plays a crucial role in the development of nearly all types of cancer[6,7]. In the HNSC cohort, miRNA and mRNA expression data had equal performance with DNA methylation data in discriminating between normal and cancer tissue. With CNV data, we observed the worst performance in each cancer cohort.

Building a model that can discriminate whether a sample comes from a tumor that will go into remission or from one that will progress until the donor's death has proven to be a much more difficult task. For this task the above approach based on LASSO regression gave poor prognostic results (AUC values in range 0.5 – 0.76).

Cancer Type	Analyzed Data	AUC	Number of Selected Features
Lung Adenocarcinoma	miRNA expression	0.98	16
	mRNA expression	0.99	24
	CNV	0.95	64
	DNA Methylation	1.00	30
Kidney Renal Clear Cell Carcinoma	miRNA expression	0.97	12
	mRNA expression	0.98	36
	CNV	0.98	76
	DNA Methylation	1	120
Head and Neck Squamous Cell Carcinoma	miRNA expression	0.99	29
	mRNA expression	0.99	33
	CNV	0.93	66
	DNA Methylation	0.99	23

Table 1. Classification performance of the supervised learning models for discrimination between normal and cancer tissues

Molecular biomarkers associated with overall patient survival

To identify molecular signatures linked to patient survival for each cancer cohort, we asked whether low or high levels of a particular measured entity (expression, CNV or methylation) are significantly correlated with patients overall survival. In particular, in each cancer cohort, for a given miRNA, mRNA, protein, CNV and methylation probe, we separated the patients into quartiles based on the measured levels of the particular entity (miRNA/mRNA/protein expression, CNV or methylation values respectively). Then, using a log-rank statistical test we compared the overall survival of the patient group characterized by low levels of the particular measured entity (ie. values below the first quartile) to the survival of the patient group with high levels of that particular measured entity (values above the third quartile) (see Figure 1). The patients were split into training and validation sets and all statistical tests were conveyed on the training datasets. Based on this "quartile" approach, we could identify miRNAs, protein-coding genes, CNV and methylation probes whose extreme measured values were statistically linked to overall patients survival (p-value of log-rank test < 0.05). For further analyses, we kept only those that were significantly associated to the overall survival also in the validation dataset. Next, in each molecular dataset we clustered the selected genes/probes from the "quartile" test using non-negative matrix factorization[8] and selected best representatives from each cluster. To build prognosis models for each molecular dataset and each cancer cohort, we performed a multivariate Cox regression[9] on the selected genes/probes. For each signature, coefficients from a multivariate Cox regression analysis on the training cohort were used to compute a risk on the validation cohort. The accuracy of the prognosis methods was assessed through a concordance index, which is a non-parametric measure that quantifies the fraction of pairs of patients whose predicted survival times are correctly ordered among all pairs that can actually be ordered[10]. The best performing models for each cancer cohort are shown in Table 2. Using only 3 or 4 genes/probes from each molecular dataset, we could achieve concordance correlation coefficient greater than 0.7 in the validation cohorts (see the red bars on Figure 2).

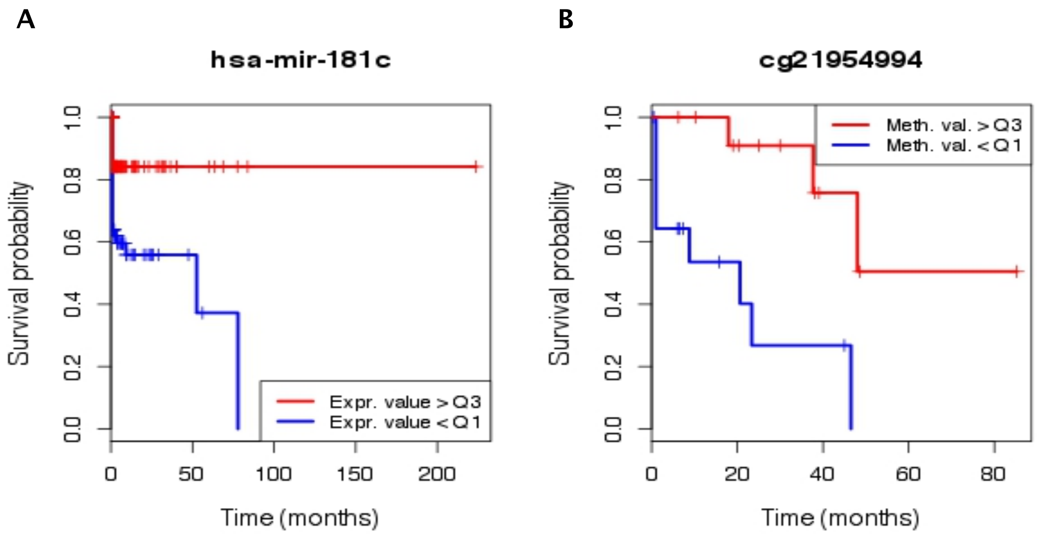

Figure 1. Quartile-based selection of features associated to overall survival. A) Differences in the survival probability between patients with high expression values of “hsa-mir-181c” (>Q3) and patients with low expression values (<Q1). B) Differences in the survival probability between patients with high methylation values (>Q3) of the “cg21954994” methylation probe and patients with low methylation values (<Q1).

Cancer Type	Analyzed Data	Survival Concordance Index	Molecular Signatures for Survival Prognosis
Lung Adenocarcinoma	miRNA expression	0.70	hsa-mir-23b; hsa-mir-181c; hsa-mir-1976
	mRNA expression	0.71	ATP8A2; FOXM1; LCN10
	CNV	0.57	HP1BP3; MLLT3; GDPD3; RP11-778D9.13
	Methylation	0.73	cg06602857; cg21954994; cg19213569
Kidney Renal Clear Cell Carcinoma	miRNA expression	0.72	hsa-mir-21; hsa-mir-183; hsa-mir-3942; hsa-let-7b
	mRNA expression	0.69	BARX1; ITPKA; NKX2-5
	CNV	0.65	IFNA5; CDKN2A; RP11-399D6.2
	Methylation	0.77	cg09635053; cg14898260; cg23368159
Head and Neck Squamous Cell Carcinoma	miRNA expression	0.68	hsa-mir-520g; hsa-mir-29b-1; hsa-mir-144; hsa-mir-137
	mRNA expression	0.64	AQP5; CAMKV; SNAP25
	CNV	0.52	RP11-419C19.2; HOXD3; BRIX1
	Methylation	0.67	cg14526044; cg15716405; cg17720011; cg12042587

Table 2. Molecular signatures for cancer survival prognosis and their performance on the validation datasets for each cancer cohort.

Next, we wanted to test whether the molecular profiles that are distinctive for normal and cancer tissues are also correlated with patient survival. Using the selected genes/probes from the normal vs cancer tissue classification, we built multivariate Cox regression prognostic models and assessed their prediction performance through a concordance index (green bars on Figure 2). Our results show that even though one can well discriminate between normal and cancer tissues using selected features, the same features are not necessarily good survival predictors. In fact, only very few genes selected from the normal vs cancer classification appear to be predictive for survival. For example, the miRNA “hsa-mir-21”, an "oncomir” associated with a wide variety of cancers[11], is predictive for survival in KIRC cohort, but it is also selected as a discriminatory feature in the normal vs cancer tissue prediction in the KIRC and LUAD cohorts.

To further assess the power of our selected molecular signatures, we built multivariate Cox regression prognostic models using randomly selected genes/probes. Figure 2 shows that our prognostic markers selected from the different molecular datasets (miRNA, mRNA, CNV, methylation) are largely superior to randomly chosen genes/probes in the three cancer cohorts.

We extended the analyses to include survival prediction based on somatic mutations profiles (SNP data), which we obtained from the TCGA Data Portal. For each gene we split the patients into two groups: patients having a somatic mutation in that particular gene, and patients with no somatic mutations in that gene. If the difference in survival between the two patient groups is significant (p<0.01), we included the corresponding gene in the multivariate Cox model. Again we split the set of patients on training and validation sets. The Cox model built on the training set was used to predict the survival on the validation dataset. For each particular gene, we required that at least 10 patients have a mutation in that gene. The survival prognosis signatures from SNPs data were superior over the signatures from the other datasets in LUAD and HNSC cohorts. Only in the KIRC cohort the signature from the methylation data gave the best performance. Next, we integrated the prediction signatures from the different "-omics" data together with clinical variables (donor age, sex and donor icd10 diagnosis) to build a "multi-omics" Cox survival prediction model. The addition of variables into the model was assessed through a forward model selection procedure (Aikake information criterion) combined with a Cox regression. However, the prognostic performance of this "multi-omics" prediction model has not improved.

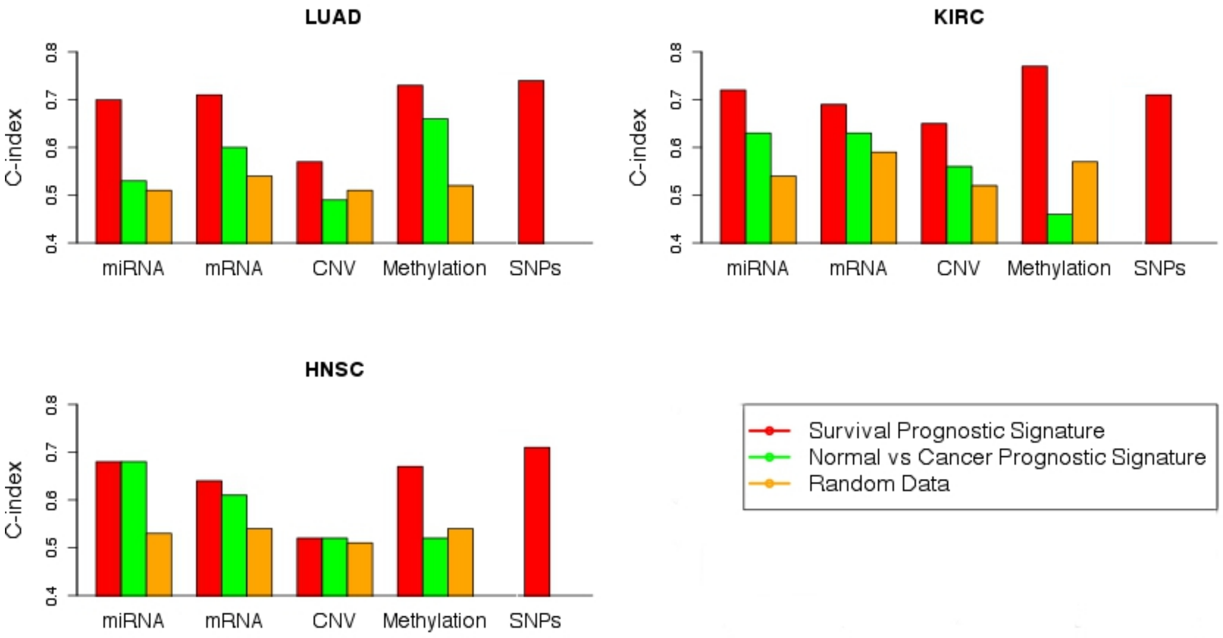

Figure 2. Performance assessment of several prognosis signatures on the validation datasets in A) Lung Adenocarcinoma, B) Kidney Renal Clear Cell Carcinoma and C) Head and Neck Squamous Cell Carcinoma. Red: Survival prognostic using molecular signatures listed in Table 2. Green: Survival prognostic using molecular signatures from normal-cancer classification. Orange: Survival prognostic using randomly chosen molecular data.

Discussion

In this work we evaluated patient survival prediction from different molecular data types and described potential prognostic signatures across three cancer types. Currently, only a few gene expression signatures are routinely used in the clinical practice for these three cancers[12]. In LUAD and HNSC cancer cohorts, somatic mutation profiles (SNP data) appear to be the most informative resources for prognostics, while DNA methylation profiles are the most informative in the KIRC cohort. Using a quartile-based selection we identified features that are prognostic for at least a subset of patients. This approach inherently supports heterogeneity, in contrast to classification methods. Some of the prognostic signatures that we identified are well studied in the literature: eg. the FOXM1 gene has been shown to promote tumor metastasis in non-small cell lung cancer patients and is associated with chemotherapy resistance[13,14]. But we also identified prognostic signatures that have not been reported as linked to cancer progression. For example, the gene ATP8A2, member of aminophospholipid transporter family, is associated with several diseases, but not with cancer. However, another gene from the same family, ATP11A, was recently identified as a predictive marker for metastasis in colorectal cancer[15].

The fact that we can relatively easily discriminate normal from tumor tissue suggests that cancer consistently alters the molecular machinery. However, cancer malignancy is heterogeneously defined within cancer type, and as a consequence molecular signatures do not perfectly predict survival. Different molecular data types have different predictive values in cancer types, which suggests that cancer malignancy relies on different mechanisms across cancers. Our analyses do not necessarily identify the cancer causal changes; they rather identify molecular markers that are affected by causal changes and are associated with survival. They offer new prospects for further investigations of cancer pathogenesis.

Methods

Data

We used preprocessed mRNA expression (mRNA-seq), miRNA expression, protein expression, somatic CNV (all them downloaded from the ICGC Data Portal, release 17) and DNA methylation data (ICGC, release 18). The LUAD dataset contains molecular profiles of 473 patients, KIRC dataset contains molecular profiles of 515 patients, and HNSC 422 patients. The data comes from 3 tissue types: primary tumor solid tissue, normal tissue adjacent to primary and normal blood derived tissue. Expression data are the most commonly and consistently available ICGC data type. Training and validation sets were created from each cancer cohort in a ratio 2:1, meaning that two-thirds of the corresponding data set was used for building the models and one-third of it for validating the models. No bias in tumor stage, age, overall survival, or gender distribution was observed between the training and validation sets.

Identification of prognostic signatures

For each molecular profile (i.e. for each miRNA, mRNA and protein) in the training dataset two groups of patients were constructed based on expression levels of the miRNA, mRNA or protein respectively: lower than the 25% quartile and higher than the 75% quartile. A log-rank test was then applied to determine if the difference in terms of overall survival between the two groups was significant (p-value < 0.05). Clustering of significant survival-associated genes (probes) was performed through a non-negative matrix factorization (NMF) with ranks tested from 2 to 6. Representative genes (probes) for each cluster were selected based on their basis coefficient. All possible combinations of representative genes (probes), such that to have only one representative per cluster, were tested to obtain the signature. A multivariate Cox regression analysis on miRNA expression values was used to compute a risk for each combination. For each signature, coefficients from a multivariate Cox regression analysis on the training cohort were used to compute a risk on the validation cohort. Performance was assessed through a concordance index (c-index). To test the significance of a particular molecular signature, we selected random genes (probes) from the ICGC datasets and trained a Cox model using these genes (probes). The number of the randomly selected genes in each test was equal to the size of the particular molecular signature. Sampling was performed over 1000 iterations to obtain an average C-index and its standard deviation.

References

1. Vogelstein B et al.. (2013). Cancer Genome Landscapes. Science. vol. 339 no. 6127 pp. 1546-1558
2. Gerlinger M et al.. (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat. Genet. 2014/02/04 ed. Nature Publishing Group; 2014; 46:1–12.
3. The Cancer Genome Atlas Research Network. 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550
4. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1
5. Fawcett, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters 27 (8): 861–874.
6. Craig, JM; Wong, NC (editor) (2011). Epigenetics: A Reference Manual. Caister Academic Press. ISBN 978-1-904455-88-2.
7. Kulis M1, Esteller M. 2010. DNA methylation and cancer. Adv Genet. 70:27-56.
8. Brunet, J-P et al. (2004). Metagenes and molecular pattern discovery using matrix factorization. PNAS. USA 101(12)
9. Cox, D. R.; Oakes, D. (1984). Analysis of Survival Data. New York: Chapman & Hall. ISBN 041224490X
10. Lawrence I-Kuei Lin (1989). "A concordance correlation coefficient to evaluate reproducibility". Biometrics (International Biometric Society) 45 (1): 255–268.
11. Zheng J, Xue H, Wang T et al. (2011). "miR-21 downregulates the tumor suppressor P12(CDK2AP1) and Stimulates Cell Proliferation and Invasion". J. Cell. Biochem. 112 (3): 872–80.
12. Yuan Y et al. 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types.
13. Nuo Xu e al. 2013. FoxM1 Is Associated with Poor Prognosis of Non-Small Cell Lung Cancer Patients through Promoting Tumor Metastasis. Plos ONE. DOI: 10.1371/journal.pone.0059412
14. Wang et al. 2013. FoxM1 expression is significantly associated with cisplatin-based chemotherapy resistance and poor prognosis in advanced non-small cell lung cancer patients. Lung Cancer. 2013 Feb;79(2):173-9
15. Miyoshi, N.et al. ATP11A is a novel predictive marker for metachronous metastasis of colorectal cancer. Oncol. Rep 2010, 23, 505–510.