Analysis of prognostic biomarkers in different types of tumors

Synopsis of Ph.D. thesis

Ádám Nagy

Semmelweis University

Doctoral School of Pathological Sciences

Supervisor:
Balázs Győrffy M.D., D.Sc.

Official reviewers:
Ágnes Tantos Ph.D.
István Hritz M.D., Ph.D.

Head of the Final Examination Committee:
Alán Alpár M.D., D.Sc.

Members of the Final Examination Committee:
Ádám Vannay M.D., Ph.D.
Dávid Szüts Ph.D.
Nándor Gábor Than M.D., Ph.D.

Budapest

2020

1. Introduction

In my doctoral dissertation I deal with the study of prognostic biomarkers of several malignant tumors.

The first biomarker to be examined was the KRAS mutation in non-small cell lung cancer, the role of which in survival as well as in shaping the therapeutic response is disputed. The prognostic effect of KRAS mutations was first described in colon cancer, however, these results could not be supported by subsequent studies.

The second topic involved the NPM1 gene in the acute myeloid leukemia as the most frequently mutated gene. It plays a key role in cell cytoplasmic protein alteration contributing to leukemogenesis processes. The NPM1 mutations are specific mainly for myeloid cells, because the aberrant cytoplasmic localization of the proteins cannot be observed in lymphoid cells from bone marrow or peripheral blood. The NPM1 somatic mutations often co-occur with “internal tandem duplications” (ITDs) in the FLT3 gene and with DNMT3A, IDH1, and IDH2 mutations. It is important to highlight that NPM1 mutations are prominent prognostic markers in acute myeloid leukemia as well as in the selection of the optimal treatment strategy, however, the molecular background of this is still little known.

In hepatocellular carcinoma, a number of biomarkers have been identified that allow for more accurate molecular classification of the tumor, predicting expected survival and response to therapy. The low reproducibility of the prognostic value of previously published microRNA and mRNA-based biomarkers indicates the need to develop a more complex assay method in tumors of diverse etiology.

Many genes play a role in the development of the tumors, which were classified by previous studies into eight major functional groups, also known as tumor characteristics, such as: cell division, tumor suppressors inhibition, angiogenesis stimulation, genomic instability sustainment, reprogramming the energy metabolism, metastasis induction, cell death resistance formation and maintaining DNA replication. However, there is currently no study investigating the changes in gene expression associated with the above mentioned eight tumor factors in different tumor types and their effect on patient survival.

2. Objectives

The objectives of my PhD work are as follows:

2.1. to investigate the prognostic effect of the transcriptomic pattern associated with somatic KRAS mutations in non-small cell lung cancer.

2.2. to examine the effect of NPM1 mutations on gene expression and survival of patients in acute myeloid leukemia and verify these changes with clinical patient samples.

2.3. to validate the prognostic role of previously published mRNA and microRNA based prognostic biomarkers using several independent gene expression database of liver hepatocellular carcinoma.

2.4. to investigate the relationship between tumor factor related gene expression and survival, and to investigate the prognostic effect of transcriptomic patterns determined by the eight major tumor characteristics in different tumor types.

3. Methods

3.1. Processing of mRNA expression data

The mRNA expression data were downloaded from two publicly available databases: the complete RNA sequencing data were obtained from The Cancer Genome Atlas (TCGA) and the microarray gene expression data from the National Center for Biotechnology Information Gene Expression Omnibus (GEO NCBI).

Data from a total of 28 tumor types (with at least 100 samples available) were downloaded and processed from the TCGA program. Normalization of the raw data was performed using the DESeq algorithm, which is based on the negative binomial distribution.

The microarray gene expression data were used in case of the non-small cell lung cancer, acute myeloid leukemia and liver hepatocellular carcinoma. The gene expression data was normalized with the MAS5 algorithm which is part of the "affy" R Bioconductor package.

3.2. Processing of miRNA expression data

Similarly to the mRNA expression data, the microRNA expression data were obtained from the TCGA and GEO NCBI databases. A microRNA expression database linked to clinical data was established for hepatocellular carcinoma.

In case of the hepatocellular carcinoma microRNA expression data originating from the Illumina HiSeq 2000 platform, the data has been normalized with the "reads per million mirna mapped" (RPM) method.

The microarray based hepatocellular carcinoma microRNA expression data were used without renormalization.

3.3. Processing of mutation data

Complete exom sequencing data from the tumor and its associated normal sample of non-small cell lung cancer patients were downloaded from the GDC TCGA database. The MuTect algorithm was used with default setting parameters to identify somatic mutations. The GRCh37 (hg19) version human genome was used as a reference sequence.

Processed copy number variation (CNV) data of non-small cell lung cancer patients were also downloaded from the GDC TCGA database. In the case of gene amplification, the average of the segments is above 0.2, and in the case of gene deletion below -0.2.

Mutation detection was based on mutation annotated format (MAF) files from TCGA, containing identified, processed mutations in case of acute myeloid leukemia. The results from MuTect2, MUSE, VarScan and SomaticSniper variant identification algorithms were used and integrated. The “maftools” R Bioconductor software package was used to aggregate and visualize the data.

3.4. Statistical computations

The effect of somatic mutations on gene expression was assessed by the Mann-Whitney test. Based on the somatic mutation status of the selected gene, the patients' RNA sequencing data were divided into two groups: one group of patients carried the gene mutation and the other group did not. Using the Mann-Whitney test, we examined which genes show a significant change between the mutant and wild-type patient groups. For each gene, the change in gene expression (fold change) between the two groups of patients was determined by dividing the median gene expression values of the mutant and wild-type patient groups. It is important to note that in my work we investigated the effect of copy number differences on gene expression in non-small cell lung cancer, the transcriptomic effect of NPM1 mutations in acute myeloid leukemia, and the association between gene expression and vascular invasion in hepatocellular carcinoma using the Mann-Whitney test.

The relationship between gene expression and survival was examined using a Cox regression model and Kaplan-Meier survival analysis. The “survival” R software package was used for Cox regression analysis. Hazard ratio (HR), 95% confidence intervals (CI), and log-rank P-values were calculated in the analysis. The “survplot” R software package was used in the survival analysis to create Kaplan-Meier plots.

3.5. Clinical samples

The acute myeloid leukemia clinical samples were obtained from the 1st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary were utilized in the in vitro validation of NPM1 mutation associated genes.

We applied Sanger sequencing to detect somatic mutations in NPM1 gene. Quantitative polymerase chain reaction (qPCR) was used for the gene expression quantification. DNA was isolated from peripheral blood and bone marrow samples using the High Pure PCR Template Preparation Kit following the manufacturer's protocol. DNA concentration was measured by UV spectrophotometry.

4. Results

4.1. The transcriptomic effect of KRAS mutation for survival in non-small cell lung cancer

The genomic and transcriptomic data of non-small cell lung cancer patients were derived from the TCGA and NCBI GEO databases. In case of the TCGA dataset, only those patients were selected who also have whole exome, transcriptome, and copy number alteration data. Thus, the total sample count was 555 patients. The NCBI GEO database contained microarray gene expression data of 2347 non-small cell lung cancer patients. In the survival analysis, the microarray gene expression data was used due to a larger number of samples for more reliable results.

First, I examined the impact of KRAS mutations *per se* on overall survival and I found that KRAS mutation status was not significantly correlated with overall survival (HR=1.02; *P*=0.95). I obtained similar results when I investigated the effect of KRAS gene expression on overall survival (HR=1,1; *P*=0,43).

In the Mann-Whitney analysis, I compared the expression of 11,500 genes between KRAS mutant and wild-type groups using the RNA sequencing data of non-small cell lung cancer patients. The five strongest genes correlation to KRAS status in the Mann-Whitney test include the FOXRED2, PEX3, KRAS, TOP1 and ABL2 genes. The mean expression of these genes was used as a surrogate signature of KRAS mutation status in the gene chip database. When using the median expression as a cutoff, I achieved high association with OS (HR=2.4; *P*=1.24E–12). To validate the approach of running a nonparametric analysis using the five strongest genes, I have also run the analysis using the second (HR=1,9; *P*=4,7E-08) and the third set (HR=2,5; *P*=4,4E-14) of five best genes associated with KRAS mutation and I obtained very similar results to the set of the first five genes.

I further analyzed the prognostic effect of KRAS amplification and KRAS deletion associated transcriptomic signature and I found that only the KRAS deletion associated signature showed significant correlation with survival (HR=2,3; *P*=1,8E–11).

4.2. Examining the transcriptomic effect of NPM1 mutations in acute myeloid leukemia

In the first part of the analysis, I built a database containing four acute myeloid leukemia patient cohorts with independent gene expression, mutation, and clinical data. These datasets are the GSE6891 (460 patients), TCGA (116 patients), GSE1159 (247 patients), and the Semmelweis (169 patients) groups.

Using the first training cohort (GSE6891), I identified 85 genes that expression was significantly altered between the NPM1 mutant and wild-type patient groups. In the second (TCGA) and third (GSE1159) training cohorts, 49 of the previously identified 85 genes reached statistical significance.

For qPCR measurement only those genes were selected which showed a significant gene expression change and a fold change over 2.0 or below 0.5 in each training set (n=32). Correlation to survival was used as an additional filter (n=27), and the pipeline of gene selection for qPCR measurement.

The best performing genes discriminating NPM1 mutant and wild-type samples were HOXA5, HOXB5, HOXA10, PBX3, MEIS1, and ITM2A. Of these, ITM2A was the only downregulated gene. Kaplan-Meier curves show that high expression of these genes was correlated with poor survival. In the case of ITM2A, lower expression was associated with worse outcome.

The most significant genes associated with NPM1 mutations as observed in the training sets was validated by qPCR. The expressions of HOXA5, HOXA10, HOXB5, MEIS1 and PBX3 genes were significantly higher while the expression of the ITM2A gene was significantly lower in the NPM1 mutant patient cohort. Finally, the survival analysis provided a significant association between the expression of the HOXA5, HOXA10, PBX3, and MEIS1 genes and overall survival in the validation cohort.

4.3. Identification of prognostic biomarkers in hepatocellular carcinoma

Hepatocellular carcinoma database contains mRNA and microRNA expression data from the TCGA and NCBI GEO repositories. The micro-RNA expression database contains three GEO datasets – GSE31384 (166 patients), GSE10694 (156 patients), GSE6857 (481 patients) – and the TCGA dataset (421 patients). According to the mRNA expression database TCGA contains 372 patients, GSE9843 contains 91 patients and GSE20017 dataset contains 135 patients.

4.3.1. miRNA based biomarkers

In the first part of the analysis, I performed a search using the „hepatocellular", „carcinoma" and „miRNA" keywords, where I identified 173 previously published prognostic miRNAs in hepatocellular carcinoma.

Of the 173 biomarker candidates, the expression of 55 miRNAs showed significant correlation with overall survival in the TCGA dataset.

I also examined the prognostic value of the 173 miRNAs in the GSE31384 dataset and I identified 29 miRNAs that the expression was significantly associated with overall survival. Important to highlight that the expression of these miRNAs showed significant association with overall survival in TCGA dataset as well.

Of the 173 survival associated miRNAs, 113 had altered expression compared to normal tissues. Among these, the most significant miRNAs were the hsa-miR-199a, hsa-miR-34a, hsa-miR-106b, hsa-miR-222 and the hsa-miR-221.

4.3.2. mRNA based biomarkers

The PubMed search for mRNA based HCC prognostic biomarkers resulted 318 biomarkers, of which 180 genes showed significant association with overall survival in Asian patients. Out of 318 biomarkers, 128 were associated with overall survival in White/Caucasian patients.

Out of the 318 biomarker candidates, 82 biomarker candidates were shared by both ethnic groups. Among the shared biomarkers, 72 were originally described in Asian and 10 in White/Caucasian subjects.

In addition, we also investigated the prognostic significance of the 318 previously published biomarkers in the pooled dataset (including all types of ethnicity), 178 genes showed significant association with overall survival.

4.4. Examining the prognostic effect of tumor associated genes in different types of tumors

The complete dataset of RNA-seq samples with follow-up comprised 9,663 specimens from 26 distinct tumor types. Across the entire database, the median follow-up for overall survival was 24.3 months, and for relapse-free survival, it was 23.8 months.

Cox regression analysis was performed using the RNA-seq expression of 671 cancer hallmark genes. I computed the proportion of significant genes in each cancer hallmark and in each tumor type. Hierarchical clustering was performed to correlate different tumor types and cancer hallmark-associated genes. In this analysis, genes associated with invasion and metastasis activation, genome instability, sustained proliferative signaling and cellular energetics deregulation clustered into separate cohorts. The top five tumors that contained the highest proportion of established cancer hallmark genes significantly associated with overall survival were kidney renal clear cell carcinoma, low grade glioma, melanoma, thymoma, and liver cancer.

The expression signature of hallmark features was determined for each tumor types, and the prognostic effect of these signatures was investigated in different types of cancer. Of the eight hallmark feature signatures, the sustaining proliferating signaling, genome instability, cellular energetics deregulation, invasion and metastasis activation and resisting cell death showed significant association with overall survival in at least five tumor types.

In at least ten tumor types, there were 39 genes whose expression was associated with overall survival. I pinpointed the genes with the highest prognostic power in each cancer hallmark feature: BRCA1 associated with genome instability in low grade glioma, CDK1 linked to cell death resistance in kidney papillary carcinoma, the E2F1 tumor suppressor in cervical cancer, EREG enabling replicative immortality in cervical cancer, FBP1 participating in the deregulation of cellular energetics in kidney renal clear cell carcinoma, MYC activating invasion and metastasis in bladder cancer, RUNX1 sustaining proliferative signaling in glioma and SERPINE1 playing a role in inducing angiogenesis in glioma.

5. Conclusions

5.1. KRAS mutations have no direct effect on survival in non-small cell lung cancer, however, the gene expression pattern associated with KRAS mutations has strong prognostic power.

5.2. Gene expression changes were identified which were associated with NPM1 mutations in the training set and were confirmed in two additional patient groups as well as in clinical samples. The HOXA5, HOXA10, HOXB5, PBX3, MEIS1, and ITM2A genes showed the most significant gene expression differences between the NPM1 mutant and wild-type patient groups, which was also confirmed in clinical specimens.

5.3. Out of the previously published 173 miRNAs, the expression of 29 miRNAs showed significant association with overall survival using multiple independent hepatocellular carcinoma miRNA expression datasets. Less than half of the previously published mRNA based biomarkers retained their prognostic properties in survival analysis in hepatocellular carcinoma.

5.4. Signatures constructed by using cancer hallmark genes showed tumor type-specific correlations with survival. Individual cancer hallmark genes showing prognostic significance in multiple types of cancers were also uncovered.

6. Bibliography of the candidate's publications

6.1. Publications related to the thesis:

- **Nagy Á**, Pongor LS, Szabó A, Santarpia M, Győrffy B. (2017) KRAS driven expression signature has prognostic power superior to mutation status in nonsmall cell lung cancer. Int J Cancer, 140: 930-937. **IF=7,36**

- **Nagy Á**, Ősz Á, Budczies J, Krizsán S, Szombath G, Demeter J, Bödör C, Győrffy B. (2019) Elevated HOX gene expression in acute myeloid leukemia is associated with NPM1 mutations and poor survival. J Adv Res, 20: 105-116. **IF=6,992**

- **Nagy Á**, Lánczky A, Menyhárt O, Győrffy B. (2018) Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. Sci Rep, 8: 9227. **IF=4,011**

- Menyhárt O, **Nagy Á**, Győrffy B. (2018) Determining consistent prognostic biomarkers of overall survival and vascular invasion in hepatocellular carcinoma. R Soc Open Sci, 5: 181006. **IF=2,515**

- **Nagy Á**, Munkácsy G, Győrffy B. (2020) Pancancer survival analysis of cancer hallmark genes. Under publication.

6.2. Publications related to the thesis

- **Nagy Á**, Győrffy B. (2016) Internet-based opportunities in breast cancer diagnostics and research. Magy Onkol, 60: 273-280.

- Lánczky A, **Nagy Á**, Bottai G, Munkácsy G, Szabó A, Santarpia L, Győrffy B. (2016) miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. Breast Cancer Res Treat, 160: 439-446. **IF=3,626**

- Szász AM, Lánczky A, **Nagy Á**, Förster S, Hark K, Green JE., Boussioutas A, Busuttil R, Szabó A, Győrffy B. (2016) Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. Oncotarget, 7: 49322-49333. **IF=5,168**