Risk Stratification of Gastrointestinal Stromal Tumors by Nanostring Gene Expression Profiling

Klaudia Nowak
University of Alberta
Kim Formenti
University of Alberta Hospital
Jingyang Huang
University of Alberta Hospital
Gilbert Bigras
University of Alberta
Quincy Chu
University of Alberta
Benjamin A. Adam
University of Alberta
Iyare Izevbaye (izevbaye@ualberta.ca)
University of Alberta Hospital https://orcid.org/0000-0003-2906-6620

Research Article

Keywords: gene expression profiling, gastrointestinal stromal tumours, risk stratification

Posted Date: October 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1015559/v1

Version of Record: A version of this preprint was published at Journal of Cancer Research and Clinical Oncology on January 28th, 2022. See the published version at https://doi.org/10.1007/s00432-022-03924-3.

Risk Stratification of Gastrointestinal Stromal Tumors by Nanostring Gene Expression Profiling

Klaudia Nowak[1], Kim Formenti[1], Jingyang Huang[1], Gilbert Bigras[2], Quincy Chu[3], Benjamin A. Adam[1], Iyare Izevbaye[1]

[1] Department of Laboratory Medicine and Pathology, 8440 112 Street NW, University of Alberta, Edmonton, AB, Canada T6G 2B7

[2]Department of Laboratory Medicine and Pathology, and [3]Department of Oncology, University of Alberta and Cross Cancer Institute, 11560 University Avenue, Edmonton, AB, Canada T6G 1Z2

Corresponding author: Iyare Izevbaye, Department of Laboratory Medicine and Pathology, 8440 112 Street NW, University of Alberta, Edmonton, AB, Canada T6G2B7, Tel: 780-407-8025 Fax: 780-407-8599; Email: izevbaye@ualberta.ca; ORID iD0000-0003-2906-6620

ABSTRACT

Purpose

The risk assessment classification schemes for gastrointestinal stromal tumors (GIST) include tumor site, size, mitotic count and variably tumour rupture. Heterogeneity in high risk GIST poses limitations for current classification schemes.

This study aims to demonstrate the clinical utility of risk stratification by gene expression profiling (GEP) using Nanostring technology.

Method

Fifty-six GIST cases were analyzed using a 231 gene expression panel. GEP results were correlated with clinical and pathological data. The prognostic performance was assessed in 34 patients with available survival data using ROC curves, Kaplan Meier Survival curves and compared with traditional risk assessment schemes.

Volcano plot analysis identified seven genes with significantly higher expression (FDR < .05) in high risk than in non-high risk tumors, namely, TYMS, CDC2, TOP2A, CCNA2, E2F1, PCNA, and BIRC5. Together, these transcripts exhibited significantly higher expression in high risk tumors than in intermediate (P < .01), low (P < .001), and very low (P = 0.01) risk tumors. Receiver operating characteristic curve analysis demonstrated area under the curve (AUC) to be 0.858 for the separation of high risk and non-high risk tumors. Kaplan-Meier survival analysis demonstrated improved risk stratification (log-rank test P < .001) compared to the current risk assessment classification (P = 0.231).

Conclusion

In addition to current clinical and histology-based risk classification for patients with GIST, gene expression may offer complementary prognostic information.

Key Points:

A seven-gene signature using gene expression profiling by nanostring provides additional risk prediction information to traditional risk stratification in gastrointestinal stromal tumours. We identified cell cycle regulation as a key biological pathway of these genes and used immunohistochemistry for Ki67 and PCNA biomarkers to confirm our results. Keywords: gene expression profiling, gastrointestinal stromal tumours, risk stratification

Statements and Declarations

Funding was provided by Alberta Health Services, Edmonton

Conflicts of interest/Competing interests All authors declare no conflict of interests

Availability of data and material Expression data available as supplemental data

Code availability Not applicable

Ethics approval Ethics approval was obtained from the Ethics Board of the Universities of Alberta and Calgary (HREB.CC-17-0375)

Consent to participate NA

Consent for publication All authors give their consent for publication.

Manuscript Headings

Introduction

Patient and Methods

Results

Discussion

INTRODUCTION

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal tumors of the gastrointestinal tract. GISTs show equal distribution across geographic regions and ethnic groups, and have slightly male (55%) or possibly no gender predisposition and annual global incidence of 10-15 cases per million.[1, 2]

GISTs arise from interstitial cells of Cajal. New evidence suggests that PDGFRA-mutant GISTs are derived from telocytes rather than typical ICC cells.[3]

Approximately 50-70% originate in the stomach, followed by 20–30% from the jejunoileum, and few from the colon and rectum (5–15%), duodenum (5%), and esophagus (< 5%). Primary pancreatic, omental, and mesenteric GISTs have been reported but are very rare. [4]

Detection of mutations, e.g., *KIT, PDGFRA*, enables guided therapeutic decision-making because of the availability of kinase inhibitors that have been demonstrated to show tumor responsiveness. The genetic progression of pre-malignant sub-centimeter GISTs to malignant GISTs is a consequence of a stepwise accumulation of mutations in chromosome 14q, 22q, 1p, and 15q, along with cell cycle dysregulating events and dystrophin inactivation.[5-9] Over 75% of all GISTs involve a *KIT* mutation[10-13], whereas approximately 80% of *KIT*-negative GISTs harbor *PDGFRA* mutations.[14] Wild-type GISTs may also harbor mutations in the succinate dehydrogenase (SDH) complex (or epigenetic silencing of the SDHC promoter)[15-18], *NF1* gene, BRAF p.V600E[17] or may possess *FGFR1* and *NTRK3* gene fusions [19].

MYC-associated factor X (*MAX*) genomic inactivating mutations have been demonstrated to be driver events, enabling GIST progression via loss of MAX expression and consequent p16 silencing and cell cycle progression.[20] Dysegulation of CD133 has been reported to be associated with gastric location and poor prognosis in GIST.[21, 22] Furthermore, inactivation of dystrophin, encoded by *DMD* on Xp21.1, has been shown to occur as a late event in GIST progression and is present in more than 90% of metastatic GISTs.[23] Genetic alterations in *PTCH1, GLI1, SMO, CCND2, EGFR*, and silencing of *CDKN2A, CDKN2C, SMARCB1, PTEN, ANXA1, ETV1,*

TWIST1, PTPN18[24], *TP53, MEN1, FGFR1, CHD4, CTNND2* [25], *SETD2, CBL , ARID1A, BCOR, APC*[19], and *IGFR1* also support tumour growth. Copy number changes detected include 12p amplification in *KRAS, FIF1B, PPM1A, NF2*, and 14p and 22p deletions. [26]

A number of current schemes for risk assessment in newly diagnosed primary GIST include three factors: mitotic index, tumor size, and tumor location.[16, 27, 28] A study using these factors showed that the median overall survival of all patients with diagnosed GIST was 11.7 years.[29] Some models for prediction of risk of recurrence, such as the National Institutes of Health (NIH) classification, consider tumor size and mitotic count alone.[30] Studies using this criteria estimate the risk of recurrence in high risk GISTs at 38.5%,[31] 46%,[32] and 62.5%.[33] Alternative methods incorporate other criteria, such as tumor site and tumor rupture.[27, 34] A study using this method showed that as many as 76.9% patients with the highest risk showed recurrence during follow-up.[34] These risk prediction models have been proposed as instruments for determining the potential benefits of adjuvant therapy.[35] However, prognostic heterogeneity in high and intermediate risk patients, especially in the NIH scheme, is one of their limitations.[35]

Gene expression profiling (GEP) is a powerful tool for assessing risk recurrence in several cancers.[36] Numerous GEP Studies in GIST have contributed to our knowledge of the pathobiology of GIST. Microarray analysis have shown distinctive profiles for gastric and intestinal GIST[37]. AURKA was identified as a marker for high risk GIST by bioinformatics analysis of publicly available GEP data and integration with clinicopathologic studies.[38] CD133 was identified to be a marker for exon 11 mutation, gastric origin and poor prognosis by GEP.[21] Hierarchical clustering of GEP results could be used to further subclassify GIST into 2 groups with distinctive clinical behavior and metastatic risk[39]. Most of these studies used DNA microarray technique, a method that has not be easily translated into routine clinical laboratory workflow.

Novel platforms, e.g., Nanostring, allowing reliable and reproducible gene expression analysis of archival formalin-fixed paraffin-embedded (FFPE) samples are now available.[40-42]

In this study, we hypothesized that gene expression signatures can be used to accurately classify patients with GIST into risk categories using Nanostring, a highly clinically reproducible GEP technique that is amenable to a routine clinical laboratory. The study furthers the goal of precision medicine by identifying and applying molecular determinants of GIST biological behavior to risk stratification.

PATIENTS AND METHODS

Study Cohort and Clinico-pathological Data

This study was approved by the Ethics Review Board at the University of Alberta (HREBA.CC-17-0375). The laboratory information system at the University of Alberta was searched for all cases with an original diagnosis of GIST between January 2010 and December 2016. In total, 56 GIST cases with residual FFPE material were identified and retrieved. In all cases, the initial diagnosis of GIST had been confirmed using CD117 and/or DOG1 immunohistochemistry. Pathological features, including tumor size, site, histological features, mitotic rate, risk assessment according to the modified NIH classification criteria, and patient tumor-node-metastasis (TNM) stage were extracted from the original pathology reports. All cases were reviewed and pathologic assessment confirmed by a soft tissue pathologist. Clinical follow-up data, which was available in the local cancer registry for 36 patients, was also retrieved. These clinical data included age at diagnosis, gender, post-diagnosis survival, and recurrence.

RNA Isolation

One hematoxylin and eosin (H & E)-stained section and five unstained slides with 10-μm sections were obtained from each archival FFPE block. The H & E-stained sections were reviewed by a pathologist and tumor macrodissection was performed. RNA was isolated from the macrodissected sections using the RNeasy FFPE kit (Qiagen, Germantown, MD, USA). RNA concentration and purity were measured using a NanoDropTM 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

Gene Expression Analysis

A commercial gene expression panel, the NanoString GX Human Cancer panel (NanoString Technologies, Seattle, WA, USA), which includes 231 genes previously associated with human cancer, was used. These included transcripts associated with cell cycle, differentiation, proliferation, apoptosis, angiogenesis, and immune response. The panel was quantified using an nCounter FLEX system (NanoString Technologies, Seattle, WA, USA) as described previously.[43] Raw gene expression counts were quality controlled and normalized using the nSolver analysis software version 4.0 (NanoString Technologies, Seattle, WA, USA) and the positive and negative controls provided by the manufacturer. Data were normalized to five housekeeping genes included in the commercial panel.

Data Analysis

Post-normalization statistical analysis and visualization were performed with R version 3.4.3 (R Foundation for Statistical Computing, Austria) and gene set enrichment analysis (GSEA) version 3.0 (Broad Institute, Inc.

Massachussetts Institute of Technology, Cambridge, Massachussets, USA, and Reagents of the University of California, San Diego, California, USA)[43, 44]. Differences between patient groups were assessed using the Mann-Whitney U-test for continuous data and Fisher's exact test for categorical data. Normalized transcript counts were used for individual gene analysis. Geometric means of normalized transcript counts were used for aggregate gene set analysis. Exploratory analysis was performed using heat maps with unsupervised hierarchical clustering by Euclidean distance. Differential gene expression between high risk and non-high risk tumors was assessed using volcano plot analysis incorporating log2 fold change and Student's t-test P-values, with correction for multiple comparisons using false discovery rate (FDR) adjustment. Receiver operating characteristic (ROC) curve analysis was used to assess individual gene and aggregate gene set performance. Cox proportional hazards regression was used to assess the prognostic significance of clinical, pathological, and molecular data. Kaplan-Meier curves were used to visualize differences in patient survival between patient groups. *P*-values < .05 were considered statistically significant.

Biological pathway analysis was performed using GSEA v3.0. Gene Ontology[45] and Reactome[46] were used as the gene set databases. Parameters were set at 1000 phenotype permutations. FDR < 0.25 were used for pathway discovery. FDR < 0.05 was considered significant.

Immunohistochemistry (IHC) Validation of GE results:

PCNA and Ki67 IHC of tumour was performed and analyzed by using the software QuPath 2.3[47]. The percentage of positive nuclei (PPN) of only malignant cells was computed using a trained cell classification algorithm applied to whole slide images (WSI). Receiver operating characteristic (ROC) curves were constructed to assess performance for IHC risk prediction when correlated with pathological data. Kaplan-Meier survival curves were plotted using available patient survival data to compare prognostic performance with actual patient outcome.

RESULTS

Clinico-pathological Characteristics

Fifty-six patients diagnosed with GIST between 2010 and 2016 were studied (Table 1), among which the clinical follow-up data of 34 patients were available. The patients included 31 males (55%) and 25 females (45%), with median age of 58.5 years (range: 29–89 years). The samples included 50 primary tumors (89%) and six recurrent tumor (10.7%). The sites of involvement included 23 gastric (41.1%), three duodenum (5%), 24 small bowel

(42.9%), three rectum (5%) and metastatic, including liver and omentum (10.7%). The histological patterns included were 38 spindle (68%), five epitheliod (9%), and 13 mixed (23%). The histological grades included three very low risk (5.4%), eleven low risk (19.6%), ten intermediate (18%), and thirty-two high risk (57%), according to modified NIH classification criteria.[30, 35] Fifty-two cases were unifocal (92.8%) and four were multifocal (7.2%). One case (1.8%) was metastatic, whereas 55 were non-metastatic (98%). Thirty-one out of 34 patients with available survival data (91%) were alive at the time of data collection, whereas three died from the disease (9%). The median days of survival post-diagnosis was 969.5 (range 348-1814 days).

Mutations in *KIT* and *PDGFRA* were analyzed for all samples during initial clinical work-up by sanger sequencing (Table 1). Forty-eight cases (86%) had detectable KIT mutations, whereas *PDGFRA* mutations were detected in only five cases (9%). Three cases (5%) had no detectable *KIT* or *PDGFRA* mutations. These three cases were excluded from GEP modeling due to low number in this group. The detectable *KIT* mutations were located in exon 9 in two case (2%), exon 11 in 43 cases (77%), exon 13 in one case (2%), and exon 17 in two cases (4%). The *PDGFRA* mutations were detected in exon 18 in all five cases. The mutation types included single nucleotide variants, insertion-deletions, duplications, and frame-shift changes.

Differential Gene Expression Analysis

Exploratory analysis with hierarchical clustering identified two specific molecular signatures involving the top 50 differentially expressed genes. These signatures were capable of delineating the high risk and non-high risk groups as well as the low-risk and non-low risk groups (Supplemental Figure 1,2). Volcano plot analysis identified seven genes with statistically significant higher expression (FDR < 0.05) in the high risk group than in the non-high risk groups, namely, TYMS, CDC2, TOP2A, CCNA2, E2F1, PCNA, and BIRC5 (Figure 1). As an aggregate seven-gene set, these transcripts exhibited significantly higher expression in high risk tumors than in intermediate ($P = .002$), low ($P < .001$), and very low ($P = .007$) risk tumors (Figure 2). ROC curve analysis demonstrated an area under the curve (AUC) of 0.858 (95% CI: 0.762-0.954) for the separation of high risk vs. non-risk tumors using the seven-gene set expression (Figure 3). This was superior to that observed for any of the individual genes (AUC, 0.770–0.840).

Survival Analysis

To investigate whether the expression of the seven-gene set was a better predictor of outcome than the current clinico-pathological risk assessment classification, Cox proportional-hazards regression was performed for the 34/56 patients with available clinical follow-up data, among which three died during the study follow-up period. Univariate analysis did not demonstrated any significance in all the characteristics tested. This is likely due to our low sample size. We note, however, that the seven-gene set expression was the only variable very close to statistically significant association of increased risk of death ($P = .05$) (Table 2). None of the other clinical or pathological features currently used for risk stratification, including tumor site, tumor size, and mitotic activity were statistically significant in this study (Table 2). We considered the results of the seven-gene set to be suggestive because it performed better than other characteristics that are known to be significant in other better powered studies. It is noteworthy that exon 11 mutations were present in all three patients who died although the hazard ratio for this variable could not be adequately analyzed owing to small sample size.

Kaplan-Meier survival analysis was performed to further assess the prognostic performance of the seven-gene set expression compared to the current risk assessment classification system. Interestingly, although gene selection for the seven-gene set was trained on the risk assessment labels of the existing classification system, statistically significant prediction of patient survival was only observed with the expression of the seven-gene set (log-rank test, $P < .001$) and not with the current risk assessment classification ($P = .231$) (Figure 4).

Biological Pathway Analysis

To determine the specific biological pathways that determine risk stratification groups, GSEA was performed on the risk-stratified groups. The samples were classified into high, intermediate, and low/very low risk groups. Pairwise comparison of risk groups was performed using GSEA. The pairs were high vs. non-high risk, intermediate vs. non-intermediate risk, and low/very low vs. non-low risk. Each of the paired groups demonstrated differential up- and down-regulation of pathways. The high risk group showed the most significantly divergent differential pathway enrichment from the non-high risk groups (i.e. the intermediate and low/very low risk groups). Four pathways from the Reactome gene sets were enriched in the high risk group compared to non-high risk groups, including cell cycle (FDR = 0.043), cell cycle regulation (FDR = 0.088), mitotic, G1, and S phases (FDR = 0.096), and interleukin signaling (FDR = 0.178). The seven-gene set previously identified using volcano plots of the differential expression analysis were also the most significant genes within the biological pathways identified (Supplemental figures 3,4).

These seven genes were frequently up-regulated in pathways enriched in the high risk group and down-regulated in pathways in the low risk group. Pathways with down-regulated genes in the high risk group included developmental biology and axon guidance, although with FDR values of 0.42 and 0.37, respectively. One hundred and two pathways were enriched in the low risk group compared to the non-low risk group. The top six pathways from the Reactome gene sets were cell fate commitment, regulation of locomotion, regulation of GTPase activity, transmembrane receptor protein kinase activity, transmembrane receptor protein tyrosine kinase activity, and positive regulation of ERK1 and ERK2 cascade signaling (FDR < 0.25). The pathways identified in low risk group can be generally categorized as pathways involved in cell differentiation and maturation and cell structure and function.

Leading edge analysis of these biological pathways was performed to identify the most significant genes within each pathway and the common genes among the pathways. Analysis showed a higher clustering of genes for high risk groups than for low risk groups. In the high risk group, CDK1 was most highly expressed with PCNA, and was involved in all the top four biological pathways noted above (Supplemental Figure 3). Other significantly expressed genes involved in four or more pathways included BRCA1/2, TOP2A, and E2F1. In the low risk group, significantly expressed genes involved in the top four or more pathways included KDR, NTRK3, PDGFRB, EGFR, and ERBB4 (Supplemental Figure 4).

IHC Validation of seven-gene results

Since, the seven-gene sets were identified as playing a role in cell cycle signaling and progression, we chose two important gene markers of these process, namely Ki67 and PCNA for result validation by IHC. We performed image analysis of Ki67 and PCNA for risk groups and constructed ROC curves for Ki67 and PCNA, singly and in aggregate for risk prediction using the PPN score. Of the 34 cases with survival data, Ki67 and PCNA, alone and in combination, showed increased PPN in high risk GISTS in comparison to non- high risk GISTs (Figure 5,6). For Ki67, non high risk GISTs demonstrated a mean of 2.2% (range 0.7-4.5%) versus 14.0% (range 1.6-46.5%) for high risk GISTs ($p=0.002$), For PCNA, non high risk GISTs showed a mean of 13.2% (range 2.7-30.8%) versus 25.9% (range 3.6-83.5%) for high risk GISTs ($p=0.041$). Whereas, a combination of the two IHCs (product of PPNs) showed a mean of 32.2 (range 3.2-138.1) for non high risk GISTs and 570.0 (range 7.6-3881.5) for high risk GISTs ($p=0.021$). ROC curves were plotted for Ki67 and PCNA, singly and in combination to determine

sensitivity, specificity, and to estimate the ideal cut-off for PPN. Ki67 alone was the most sensitive (sens 0.905, NPV 0.818). Whereas, the combined PCNA and Ki67 PPN was the most specific (spec 0.900, PPV 0.941) The optimal diagnostic cut-offs for PPN, as defined by Youden's J-statistic, were 2.782 for Ki67, 7.606 for PCNA, and 33.013 for Ki67 and PCNA combined. Performance of Ki67 and/or PCNA PPN as potential risk prediction surrogates was assessed using Kaplan-Meier survival curves (Figure 7). The probability of survival of high PPN cases was compared with low PPN cases. All three classes (Ki67 alone, PCNA alone, and combined Ki67 and PCNA) revealed poorer survival outcomes in specimens with high PPN in contrast to low PPN (Log-rank test, $p<0.0001$). Of the high PPN pattern, combined Ki67 and PCNA demonstrated the worst overall survival outcomes. PCNA alone was the least predictive of poorer survival outcome.

DISCUSSION

In this study, we aim to evaluate a potential role for gene expression analysis in risk stratification for GIST and to identify biological determinants for aggressive behaviour. Using differentially expressed gene analysis, we identified 8 genes, whose expressions signify poorer outcomes. We demonstrate their utility for stratification in high risk GIST tumours.

Current risk stratification schemes in GIST utilize clinico-pathological criteria, including tumor size, site, mitotic count, and variably tumour rupture to predict clinical outcomes. The limitation of these methods is that a percentage of members in each risk group may show different biological behavior than the average member. Nanostring technology applies the principles of precision medicine to classify risk according to molecular features, which may predict behavior more accurately and can be used in a routine clinical setting.

We examined all GIST cases diagnosed at our institution from 2010 to 2016. Our study cohort showed an epidemiological distribution similar to those of other published cases in terms of median age, gender, site of occurrence, and KIT mutations. In this study, we analyzed various genes involved in cancer-related pathways, including cell cycle, cell differentiation, cell proliferation, apoptosis, angiogenesis, and immune response. Differential gene expression analysis identified molecular signatures for high and low risk groups.

Seven genes, namely, BIRC5, PCNA, E2F1, CCNA2, TOP2A, CDC2, and TYMS, demonstrated statistically significant differential expression (FDR < .05) between high risk and non-high risk tumors. These seven genes are involved in cell cycle control, apoptosis, and DNA replication. BIRC5 is a member of the inhibitor of

apoptosis (IAP) gene family, which encodes negative regulatory proteins that prevent apoptotic cell death. PCNA is a cofactor of DNA polymerase delta, which assists in increasing the processivity of leading strand synthesis during DNA replication. E2F1 is a member of the E2F family of transcription factors and controls the cell cycle and the tumor suppressor protein activity. CCNA2 belongs to the cyclin family and is a regulator of the cell cycle. It mediates transition through the G1/S and G2/M checkpoints. TOP2A encodes a DNA topoisomerase and is involved in chromosome condensation, chromatid separation, and relaxation of DNA torsional stress. TYMS encodes the enzyme thymidylate synthase, which maintains the dTMP pool necessary for DNA replication and repair. CDC2 (CDK1) is a serine/threonine protein kinase and regulates the G1/S and G2/M phase transition.

Many of these pathways have been implicated by previous studies in GIST progression, including cell cycle control.[31] Further analysis using GSEA confirmed a role for additional pathways such as cell cycle control and regulation, mitosis, and interleukin signaling. These pathways showed up-regulation of gene expression in high risk GIST. A larger number of pathways were up-regulated in low risk and down-regulated in non-low risk GIST. These pathways involve genes involved in cell fate commitment, cell differentiation and maturation, and cellular structure and function. This observation supports a biological mechanism of aggressive tumor behavior in GIST, in which a larger tumour fraction is in the mitotic phase of the cell cycle. This is consistent with the higher mitotic count seen in current pathologic assessment of high risk GIST. The genes identified by our study were largely different from 25 genes previously identified as prognostic/ predictive markers for soft tissue sarcomas, except for the single overlap of CDK1.[48]

We observe that the expression of the seven-gene set may be a better predictor of overall survival than traditional clinico-pathological features used for risk stratification, including age, gender, site of occurrence, tumor size, and mitotic rate especially for high risk GIST. However, we noted that this analysis is limited by the small number of deaths and overall small number of samples tested in this series, especially for the low risk group. For instance, all patients who died harbored an exon 11 mutation, but the mathematical requirement for a non-zero denominator in calculating the hazard ratio limited its assessment.

We validated our results with immunohistochemistry and image analysis, demonstrating that assessing markers of pathways involved by the seven-gene set was a useful predictor of disease risk especially in aggregage.

Nevertheless, despite the limitations of this study, these results suggest that compared to the current clinical and histology-based risk assessment classification for patients with GIST, gene expression may offer significant but

complementary prognostic performance. Nanostring technology is simple, cost-effective and amenable to FFPE tissue, with established use in routine clinical GEP assays, which up till now has been hindered by inherent limitations of other techniques. Further validation studies are warranted to confirm the clinical utility of this molecular risk stratification approach.

Acknowledgments

We acknowledge our use of the gene set enrichment analysis, (GSEA) software, and Molecular Signature Database (MSigDB).

Figure legends

Figure 1 Volcano plot identifies seven differentially expressed genes (FDR < 0.05) in high vs. non-high risk groups; Significantly higher seven-gene set expression in the high risk group vs. all other groups. (FDR – false discovery rate)

Figure 2. Gene expression box plot showing significantly higher expression of aggregate 7-gene set in high risk tumors vs. all other risk groups (Mann-Whitney U-test)

Figure 3: Receiver operating characteristic curves for all differentially expressed genes comparing high risk vs. non-high risk status. Aggregate 7-gene set has best performance (AUC=0.858). AUC – Area under the curve

Figure 4: Kaplan Meier survival curves show superior prognostic performance with aggregate seven-gene-set compared to current risk assessment classification for prediction of patient survival.

Figure 5 PPN correlates with clinicopathologic risk assessment. Vertical axis shows PPN values of high risk vs non-high risk GIST groups. (PPN – Positive Percentage Nuclei by image analysis.)

Figure 6 ROC curves show Ki67 alone was the most sensitive (sensitivity: 0.905, NPV 0.818).
Combined PCNA and Ki67 was the most specific (spec 0.900, PPV 0.941) The optimal diagnostic cut-offs for PPN were 2.782 for Ki67, 7.606 for PCNA, and 33.013 for Ki67 and PCNA combined. PPN – Percentage Positive, Nuclei NPV – Negative Predictive Value; PPV – Positive Predictive Value

Figure 7 Kaplan Meier survival curves for IHC Risk groups: Blue – low staining, red – high staining. Lower panel shows numerical value of survival time. High staining specimen show poorer survival.

Supplemental Figure 1: Heat map by hierarchical clustering showing gene expression profiling of (a) high risk (grey) vs. non-high risk (yellow) GIST.(b) low risk (yellow) vs. non-low risk (grey) GIST Columns show GIST cases, rows show gene; color intensity shows expression levels; blue: downregulation, red: upregulation. (GIST - Gastrointestinal Tumour)

Supplemental Figure 2: Gene expression heat map with clustering by Euclidean distance showing minimal clustering of risk groups with inclusion of all 231 genes.

Supplemental figure 3: Biological pathways and their differentially regulated component genes in high risk vs. non high risk

Supplemental Figure 4: Biological pathways and their differentially regulated component genes low vs. non-low risk. Columns represent genes, rows represent pathways. Colored boxes indicate genes implicated in each pathway. Color intensity shows level of expression. Red depicts up-regulation, blue depicts down-regulation.

References

[1] M. Miettinen, L. H. Sobin, and J. Lasota, "Gastrointestinal stromal tumors of the stomach: a clinicopathologic, immunohistochemical, and molecular genetic study of 1765 cases with long-term follow-up," *Am. J. Surg. Pathol.,* vol. 29, no. 1, pp. 52-68, Jan 2005, doi: 10.1097/01.pas.0000146010.92933.de.

[2] K. Soreide, O. M. Sandvik, J. A. Soreide, V. Giljaca, A. Jureckova, and V. R. Bulusu, "Global epidemiology of gastrointestinal stromal tumours (GIST): A systematic review of population-based cohort studies," *Cancer Epidemiol.,* vol. 40, pp. 39-46, Feb 2016, doi: 10.1016/j.canep.2015.10.031.

[3] R. Ricci *et al.*, "Telocytes are the physiological counterpart of inflammatory fibroid polyps and PDGFRA-mutant GISTs," *J. Cell. Mol. Med.,* vol. 22, no. 10, pp. 4856-4862, Oct 2018, doi: 10.1111/jcmm.13748.

[4] X. Zhao and C. Yue, "Gastrointestinal stromal tumor," *J. Gastrointest. Oncol.,* vol. 3, no. 3, pp. 189-208, Sep 2012, doi: 10.3978/j.issn.2078-6891.2012.031.

[5] W. El-Rifai, M. Sarlomo-Rikala, L. C. Andersson, S. Knuutila, and M. Miettinen, "DNA sequence copy number changes in gastrointestinal stromal tumors: tumor progression and prognostic significance," *Cancer Res.,* vol. 60, no. 14, pp. 3899-903, Jul 15 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10919666.

[6] B. Gunawan *et al.*, "An oncogenetic tree model in gastrointestinal stromal tumours (GISTs) identifies different pathways of cytogenetic evolution with prognostic implications," *J. Pathol.,* vol. 211, no. 4, pp. 463-70, Mar 2007, doi: 10.1002/path.2128.

[7] J. Henze *et al.*, "p53 modulation as a therapeutic strategy in gastrointestinal stromal tumors," *PLoS One,* vol. 7, no. 5, p. e37776, 2012, doi: 10.1371/journal.pone.0037776.

[8] M. A. Pantaleo *et al.*, "Dystrophin deregulation is associated with tumor progression in KIT/PDGFRA mutant gastrointestinal stromal tumors," *Clin Sarcoma Res,* vol. 4, p. 9, 2014, doi: 10.1186/2045-3329-4-9.

[9] A. Wozniak *et al.*, "Array CGH analysis in primary gastrointestinal stromal tumors: cytogenetic profile correlates with anatomic site and tumor aggressiveness, irrespective of mutational status," *Genes Chromosomes Cancer,* vol. 46, no. 3, pp. 261-76, Mar 2007, doi: 10.1002/gcc.20408.

[10] J. Andersson *et al.*, "Gastrointestinal stromal tumors with KIT exon 11 deletions are associated with poor prognosis," *Gastroenterology,* vol. 130, no. 6, pp. 1573-81, May 2006, doi: 10.1053/j.gastro.2006.01.043.

[11] G. Gastrointestinal Stromal Tumor Meta-Analysis, "Comparison of two doses of imatinib for the treatment of unresectable or metastatic gastrointestinal stromal tumors: a meta-analysis of 1,640 patients," *J. Clin. Oncol.,* vol. 28, no. 7, pp. 1247-53, Mar 1 2010, doi: 10.1200/JCO.2009.24.2099.

[12] M. C. Heinrich *et al.*, "Molecular correlates of imatinib resistance in gastrointestinal stromal tumors," *J. Clin. Oncol.,* vol. 24, no. 29, pp. 4764-74, Oct 10 2006, doi: 10.1200/JCO.2006.06.2265.

[13] X. H. Liu, C. G. Bai, Q. Xie, F. Feng, Z. Y. Xu, and D. L. Ma, "Prognostic value of KIT mutation in gastrointestinal stromal tumors," *World J. Gastroenterol.,* vol. 11, no. 25, pp. 3948-52, Jul 7 2005, doi: 10.3748/wjg.v11.i25.3948.

[14] J. Lasota and M. Miettinen, "Clinical significance of oncogenic KIT and PDGFRA mutations in gastrointestinal stromal tumours," *Histopathology,* vol. 53, no. 3, pp. 245-66, Sep 2008, doi: 10.1111/j.1365-2559.2008.02977.x.

[15] L. A. Doyle and J. L. Hornick, "Gastrointestinal stromal tumours: from KIT to succinate dehydrogenase," *Histopathology,* vol. 64, no. 1, pp. 53-67, Jan 2014, doi: 10.1111/his.12302.

[16] M. Miettinen, H. Makhlouf, L. H. Sobin, and J. Lasota, "Gastrointestinal stromal tumors of the jejunum and ileum: a clinicopathologic, immunohistochemical, and molecular genetic study of 906 cases before imatinib with long-term follow-up," *Am. J. Surg. Pathol.,* vol. 30, no. 4, pp. 477-89, Apr 2006, doi: 10.1097/00000478-200604000-00008.

[17] M. Nannini, M. Urbini, A. Astolfi, G. Biasco, and M. A. Pantaleo, "The progressive fragmentation of the KIT/PDGFRA wild-type (WT) gastrointestinal stromal tumors (GIST)," *J. Transl. Med.,* vol. 15, no. 1, p. 113, May 23 2017, doi: 10.1186/s12967-017-1212-x.

[18] A. J. Wagner, S. P. Remillard, Y. X. Zhang, L. A. Doyle, S. George, and J. L. Hornick, "Loss of expression of SDHA predicts SDHA mutations in gastrointestinal stromal tumors," *Mod. Pathol.,* vol. 26, no. 2, pp. 289-94, Feb 2013, doi: 10.1038/modpathol.2012.153.

[19] P. Trimboli and S. Sorrenti, "Low value of color flow-doppler in predicting malignancy of thyroid follicular neoplasms," *Diagn Cytopathol,* vol. 37, no. 5, pp. 391-2, May 2009, doi: 10.1002/dc.21050.

[20] I. M. Schaefer *et al.*, "MAX inactivation is an early event in GIST development that regulates p16 and cell proliferation," *Nat Commun,* vol. 8, p. 14674, Mar 8 2017, doi: 10.1038/ncomms14674.

[21] G. Arne *et al.*, "Expression profiling of GIST: CD133 is associated with KIT exon 11 mutations, gastric location and poor prognosis," *Int. J. Cancer,* vol. 129, no. 5, pp. 1149-61, Sep 1 2011, doi: 10.1002/ijc.25755.

[22] H. Geddert *et al.*, "Epigenetic Regulation of CD133 in Gastrointestinal Stromal Tumors," *Am. J. Clin. Pathol.,* vol. 147, no. 5, pp. 515-524, May 1 2017, doi: 10.1093/ajcp/aqx028.

[23] Y. Wang *et al.*, "Dystrophin is a tumor suppressor in human cancers with myogenic programs," *Nat. Genet.,* vol. 46, no. 6, pp. 601-6, Jun 2014, doi: 10.1038/ng.2974.

[24] M. Saponara *et al.*, "Molecular characterization of metastatic exon 11 mutant gastrointestinal stromal tumors (GIST) beyond KIT/PDGFRalpha genotype evaluated by next generation sequencing (NGS)," *Oncotarget,* vol. 6, no. 39, pp. 42243-57, Dec 8 2015, doi: 10.18632/oncotarget.6278.

[25] M. A. Pantaleo *et al.*, "Genome-Wide Analysis Identifies MEN1 and MAX Mutations and a Neuroendocrine-Like Molecular Heterogeneity in Quadruple WT GIST," *Mol. Cancer Res.,* vol. 15, no. 5, pp. 553-562, May 2017, doi: 10.1158/1541-7786.MCR-16-0376.

[26] A. Astolfi *et al.*, "A molecular portrait of gastrointestinal stromal tumors: an integrative analysis of gene expression profiling and high-resolution genomic copy number," *Lab. Invest.,* vol. 90, no. 9, pp. 1285-94, Sep 2010, doi: 10.1038/labinvest.2010.110.

[27] M. Miettinen and J. Lasota, "Gastrointestinal stromal tumors: pathology and prognosis at different sites," *Semin. Diagn. Pathol.,* vol. 23, no. 2, pp. 70-83, May 2006, doi: 10.1053/j.semdp.2006.09.001.

[28] M. Schmieder *et al.*, "Comparison of Different Risk Classification Systems in 558 Patients with Gastrointestinal Stromal Tumors after R0-Resection," *Front. Pharmacol.,* vol. 7, p. 504, 2016, doi: 10.3389/fphar.2016.00504.

[29] C. D. Blanke *et al.*, "Long-term results from a randomized phase II trial of standard- versus higher-dose imatinib mesylate for patients with unresectable or metastatic gastrointestinal

stromal tumors expressing KIT," *J. Clin. Oncol.,* vol. 26, no. 4, pp. 620-5, Feb 1 2008, doi: 10.1200/JCO.2007.13.4403.

[30] C. D. Fletcher *et al.*, "Diagnosis of gastrointestinal stromal tumors: A consensus approach," *Hum. Pathol.,* vol. 33, no. 5, pp. 459-65, May 2002, doi: 10.1053/hupa.2002.123545.

[31] N. Nakamura *et al.*, "Prognostic significance of expressions of cell-cycle regulatory proteins in gastrointestinal stromal tumor and the relevance of the risk grade," *Hum. Pathol.,* vol. 36, no. 7, pp. 828-37, Jul 2005, doi: 10.1016/j.humpath.2005.03.012.

[32] G. Tryggvason, H. G. Gislason, M. K. Magnusson, and J. G. Jonasson, "Gastrointestinal stromal tumors in Iceland, 1990-2003: the icelandic GIST study, a population-based incidence and pathologic risk stratification study," *Int. J. Cancer,* vol. 117, no. 2, pp. 289-93, Nov 1 2005, doi: 10.1002/ijc.21167.

[33] B. Nilsson *et al.*, "Gastrointestinal stromal tumors: the incidence, prevalence, clinical course, and prognostication in the preimatinib mesylate era--a population-based study in western Sweden," *Cancer,* vol. 103, no. 4, pp. 821-9, Feb 15 2005, doi: 10.1002/cncr.20862.

[34] H. Joensuu *et al.*, "Risk factors for gastrointestinal stromal tumor recurrence in patients treated with adjuvant imatinib," *Cancer,* vol. 120, no. 15, pp. 2325-33, Aug 1 2014, doi: 10.1002/cncr.28669.

[35] H. Y. Huang *et al.*, "A modification of NIH consensus criteria to better distinguish the highly lethal subset of primary localized gastrointestinal stromal tumors: a subdivision of the original high-risk group on the basis of outcome," *Surgery,* vol. 141, no. 6, pp. 748-56, Jun 2007, doi: 10.1016/j.surg.2007.01.024.

[36] M. Gnant *et al.*, "Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone," *Ann. Oncol.,* vol. 25, no. 2, pp. 339-45, Feb 2014, doi: 10.1093/annonc/mdt494.

[37] R. Hara *et al.*, "Microarray analysis reveals distinct gene set profiles for gastric and intestinal gastrointestinal stromal tumors," *Anticancer Res.,* vol. 35, no. 6, pp. 3289-98, Jun 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26026088.

[38] C. C. Yen *et al.*, "Integrating bioinformatics and clinicopathological research of gastrointestinal stromal tumors: identification of aurora kinase A as a poor risk marker," *Ann. Surg. Oncol.,* vol. 19, no. 11, pp. 3491-9, Oct 2012, doi: 10.1245/s10434-012-2389-0.

[39] K. M. Skubitz, K. Geschwind, W. W. Xu, J. S. Koopmeiners, and A. P. Skubitz, "Gene expression identifies heterogeneity of metastatic behavior among gastrointestinal stromal tumors," *J. Transl. Med.,* vol. 14, p. 51, Feb 13 2016, doi: 10.1186/s12967-016-0802-3.

[40] G. K. Geiss *et al.*, "Direct multiplexed measurement of gene expression with color-coded probe pairs," *Nat. Biotechnol.,* vol. 26, no. 3, pp. 317-25, Mar 2008, doi: 10.1038/nbt1385.

[41] T. Nielsen *et al.*, "Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens," *BMC Cancer,* vol. 14, p. 177, Mar 13 2014, doi: 10.1186/1471-2407-14-177.

[42] P. P. Reis *et al.*, "mRNA transcript quantification in archival samples using multiplexed, color-coded probes," *BMC Biotechnol.,* vol. 11, p. 46, May 9 2011, doi: 10.1186/1472-6750-11-46.

[43] V. K. Mootha *et al.*, "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat. Genet.,* vol. 34, no. 3, pp. 267-73, Jul 2003, doi: 10.1038/ng1180.

[44] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 102, no. 43, pp. 15545-50, Oct 25 2005, doi: 10.1073/pnas.0506580102.

[45] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.,* vol. 25, no. 1, pp. 25-9, May 2000, doi: 10.1038/75556.

[46] A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.,* vol. 46, no. D1, pp. D649-D655, Jan 4 2018, doi: 10.1093/nar/gkx1132.

[47] P. Bankhead *et al.*, "QuPath: Open source software for digital pathology image analysis," *Sci Rep,* vol. 7, no. 1, p. 16878, Dec 4 2017, doi: 10.1038/s41598-017-17204-5.

[48] A. Takahashi *et al.*, "Analysis of gene expression profiles of soft tissue sarcoma using a combination of knowledge-based filtering with integration of multiple statistics," *PLoS One,* vol. 9, no. 9, p. e106801, 2014, doi: 10.1371/journal.pone.0106801.

Figures

Figure 1: Volcano plot showing statistically significant differential expression (FDR<0.05) for 7 out of 231 genes between high risk and non-high risk tumors. All differentially expressed genes have increased expression in high risk vs. non-high risk tumors.

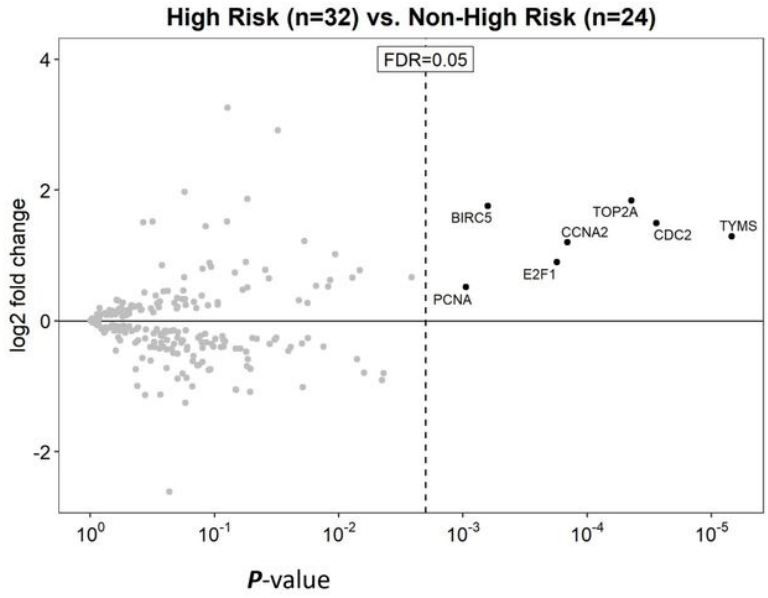

Figure 1

Please See image above for figure legend.

Figure 2: Gene expression box plot showing significantly higher expression of aggregate 7-gene set in high risk tumors vs. all other risk groups (Mann-Whitney U-test).

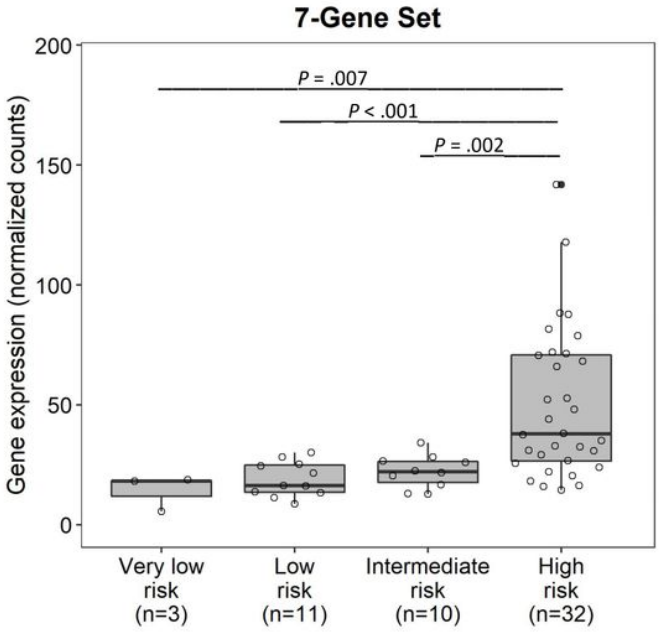

Figure 2

Please See image above for figure legend.

Figure 3: Receiver operating characteristic curves for all differentially expressed genes comparing high risk vs. non-high risk status. Aggregate 7-gene set has best performance (AUC=0.858). AUC – Area under the curve

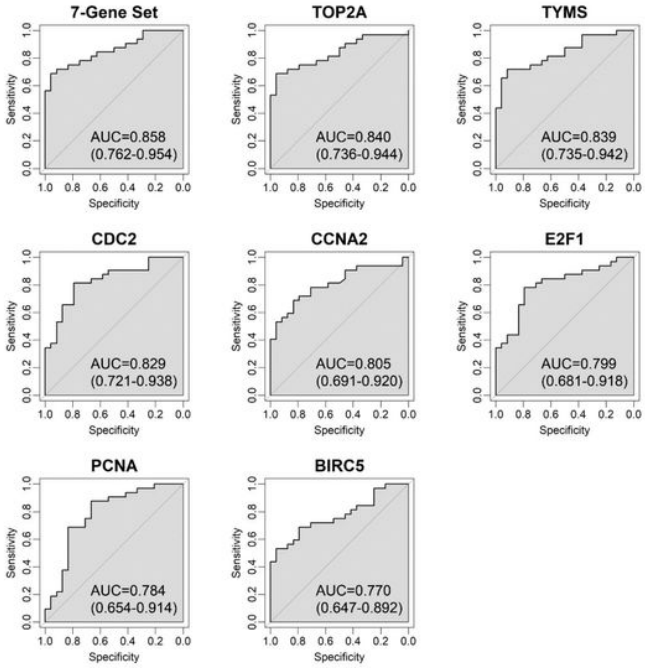

Figure 3

Please See image above for figure legend.

Figure 4: Kaplan-Meier survival curves showing improved risk prediction with 7-gene set expression versus current risk assessment classification (n=34 patients with survival data).

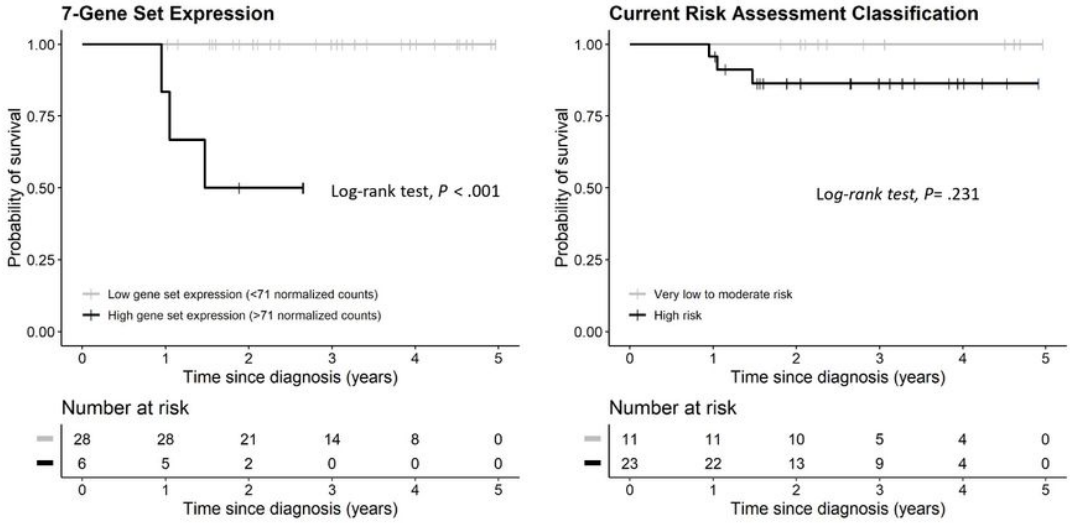

Figure 4

Please See image above for figure legend.

Figure 5: PPN assessment vs Clinicopathologic Risk assessment

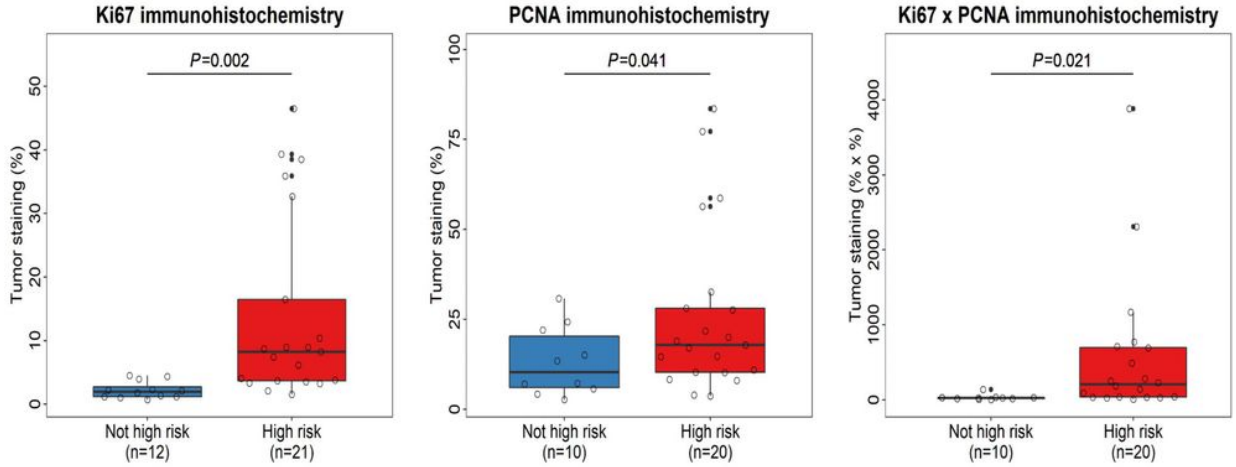

Figure 5

Please See image above for figure legend.

Figure 6: ROC of PPN staining vs Risk stratification groups

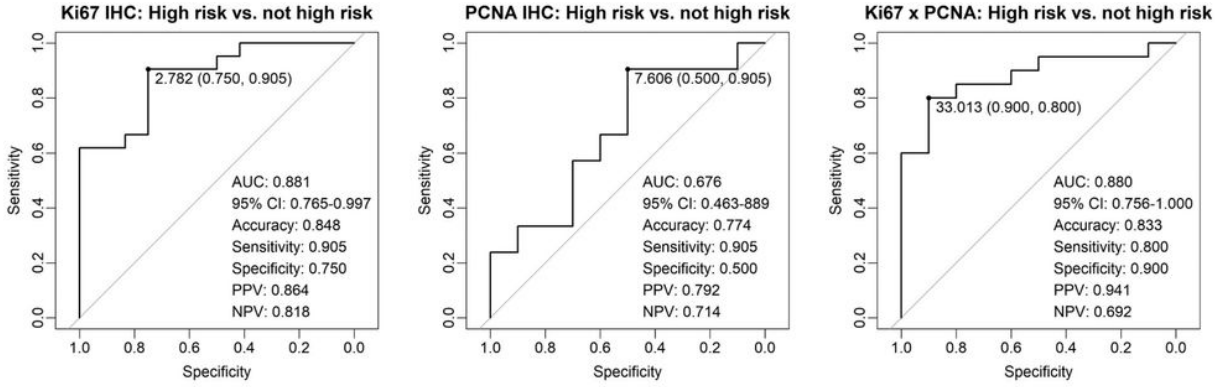

Figure 6

Please See image above for figure legend.

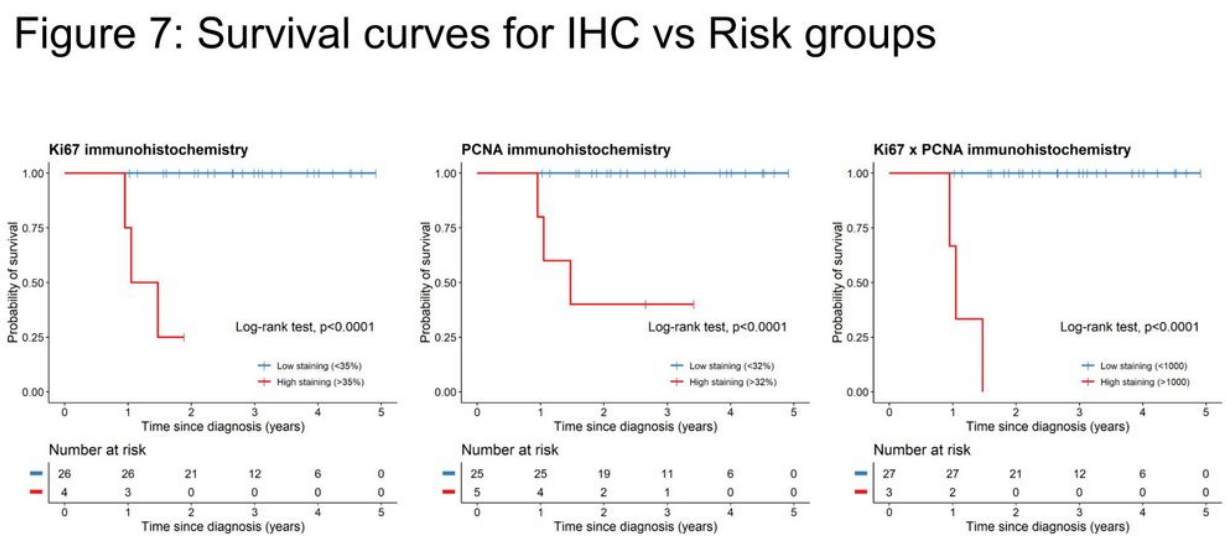

Figure 7

Please See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GISTrawdata.xlsx
- Suppfig1.pdf
- Supplementalfig24.pdf
- Table2.pdf
- Table12.pdf