Integrative clinical transcriptome analysis reveals *TMPRSS2-ERG* dependency of prognostic biomarkers in prostate adenocarcinoma

Julia S. Gerke[1], Martin F. Orth[1], Yuri Tolkach[2], Laura Romero-Pérez[1], Fabienne S. Wehweck[1], Stefanie Stein[1], Julian Musa[1], Maximilian M.L. Knott[1,3], Tilman L.B. Hölting[1], Jing Li[1], Giuseppina Sannino[1], Aruna Marchetto[1], Shunya Ohmura[1], Florencia Cidre-Aranaz[1], Martina Müller-Nurasyid[4,5,6], Konstantin Strauch[7], Christian Stief[8], Glen Kristiansen[2], Thomas Kirchner[3,9,10], Alexander Buchner[6], Thomas G.P. Grünewald[1,3,9,10,§]

1 Max-Eder Research Group for Pediatric Sarcoma Biology, Institute of Pathology, Faculty of Medicine, LMU Munich, Munich, Germany
2 Institute of Pathology, University Hospital Bonn, Bonn, Germany
3 Institute of Pathology, Faculty of Medicine, LMU Munich, Munich, Germany
4 Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
5 Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Munich, Germany
6 Department of Internal Medicine I (Cardiology), Hospital of the LMU Munich, Munich, Germany
7 Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany
Urologic Clinic und Polyclinic, Clinical Center of the University of Munich, Munich, Germany
9 German Cancer Consortium (DKTK), partner site Munich, Germany
10 German Cancer Research Center (DKFZ), Heidelberg, Germany

§ address for correspondence:
Thomas Grünewald, MD, PhD
Max-Eder Research Group for Pediatric Sarcoma Biology
Institute of Pathology, Faculty of Medicine, LMU Munich
Thalkirchner Str. 36, 80337 Munich, Germany
Phone 0049-89-2180-73716
Fax 0049-89-2180-73604
Web www.lmu.de/sarkombiologie
Email thomas.gruenewald@med.uni-muenchen.de

KEYWORDS
Prostate adenocarcinoma, TMPRSS2-ERG, metastasis, prognostic biomarker, personalized medicine

ABBREVIATIONS
CPE: consensus purity estimation; EFS: event-free survival; ESTIMATE: estimation of stromal and immune cells in malignant tumors using expression data; FDR: false discovery rate; GEO: gene expression omnibus; GGG: Gleason Grading Group; GSEA: gene set enrichment analysis; IHC: immunohistochemistry; M0: tumor stage, indicating no distant metastases: N0: tumor stage, indicating no involvement of regional lymph nodes; NES: normalized enrichment score; PCa: prostate adenocarcinoma; rGL-pos/neg: ranked gene list based on T2E-positive/negative PCa samples; topGL-pos/neg: list of most frequent genes involved in top 20 gene-signatures based on T2E-positive/negative PCa samples; RNA-Seq: RNA sequencing; SCAN: single channel array normalization; T2E: TMPRSS2-ERG fusion oncogene; TCGA: The Cancer Genome Atlas; TCGA-PRAD: prostate adenocarcinoma study of TCGA; TNM-classification of malignant tumors describing the stages of a solid tumor (T = size of primary tumor; N = metastasis to regional lymph nodes; M = distant metastases); TMA: tissue microarray

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ijc.32792

ARTICLE CATEGORY

Tumor Markers and Signatures

NOVELTY AND IMPACT

In prostate adenocarcinoma (PCa), risk-prediction is challenging. Around 50% of PCa are characterized by TMPRSS2-ERG (T2E)-fusions defining two molecular subtypes (T2E-postive/negative). However, current prognostic tests do not consider these subtypes, which may compromise their accuracy. By integration of clinical and transcriptomic data from multiple studies, we show that the prognostic value of biomarkers critically depends on the T2E-status, and identify five biomarkers exclusively for T2E-negative PCa, which has strong implications for the development of new prognostic tests.

ABSTRACT

In prostate adenocarcinoma (PCa), distinction between indolent and aggressive disease is challenging. Around 50% of PCa are characterized by TMPRSS2-ERG (T2E)-fusion oncoproteins defining two molecular subtypes (T2E-positve/negative). However, current prognostic tests do not differ between both molecular subtypes, which might affect outcome prediction. To investigate gene-signatures associated with metastasis in T2E-positive and -negative PCa independently, we integrated tumor transcriptomes and clinicopathological data of two cohorts (total *n*=783), and analyzed metastasis-associated gene-signatures regarding the T2E-status.

Here, we show that the prognostic value of biomarkers in PCa critically depends on the T2E-status. Using gene-set enrichment analyses, we uncovered that metastatic T2E-positive and -negative PCa are characterized by distinct gene-signatures. In addition, by testing genes shared by several functional gene-signatures for their association with event-free survival in a validation cohort (*n*=272), we identified five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS)* –three of which are included in commercially available prognostic tests– whose high expression was significantly associated with worse outcome exclusively in T2E-negative PCa. Among these genes, *RRM2* and *TYMS* were validated by immunohistochemistry in another validation cohort (*n*=135), and several of them proved to add prognostic information to current clinicopathological predictors, such as Gleason score, exclusively for T2E-negative patients. No prognostic biomarkers were identified exclusively for T2E-positive tumors.

Collectively, our study discovers that the T2E-status, which is *per se* not a strong prognostic biomarker, crucially determines the prognostic value of other biomarkers. Our data suggest that the molecular subtype needs to be considered when applying prognostic biomarkers for outcome prediction in PCa.

INTRODUCTION

Prostate adenocarcinoma (PCa) is the second most common cancer in men worldwide, which is often detected in early stages due to regular screening [1]. Although most patients exhibit a slowly growing indolent tumor that can be treated with active surveillance [1], 15-20% of patients develop an aggressive tumor requiring intense treatment, which is associated with significant adverse effects [2,3]. However, it remains difficult to discriminate indolent from aggressive PCa [4], wherefore 23-42% of men are 'overtreated' leading to unnecessary therapy-associated morbidity that may affect quality of life and life expectancy [1,5,6]. Further, overtreatment constitutes a significant socioeconomic and healthcare burden in the Western world [5]. Thus, novel strategies to discriminate aggressive from indolent disease are urgently required.

Around 50% of PCa are characterized by chromosomal rearrangements generating chimeric oncogenes through fusion of *TMPRSS2* with *ERG*, the latter belonging to the ETS family of transcription factors [7]. TMPRSS2-ERG (T2E) acts as an aberrant transcription factor with oncogenic properties [7]. Prior studies proved that T2E-positive and -negative PCa constitute molecularly distinct PCa-subtypes [8,9], which may exploit different gene-signatures or pathways to promote PCa malignancy.

A recent study highlighted the importance of certain gene-signatures for progression of PCa and suggested several genes as potential biomarkers [10]. Yet, the impact of molecular alterations such as T2E on these gene-signatures was not specifically considered.

Here, we combined transcriptome profiles and clinicopathological data of two discovery cohorts, and explored gene-signatures and their associated genes involved in metastasis depending on the T2E-status. We identified five prognostic biomarkers specifically suitable for T2E-negative PCa, which were validated in two additional cohorts. Going beyond prior studies [8–10], we show that the T2E-status critically determines the nature of distinct metastasis-associated gene-signatures, and strongly impacts on prognostic biomarkers.

METHODS

Microarray and RNA sequencing (RNA-Seq) data

Two publicly available gene expression datasets with matched clinicopathological data were downloaded from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) (**Supplementary Table 1**). The GEO dataset (GSE46691) comprised 545 PCa cases profiled on Affymetrix GeneChip Human Exon 1.0 ST arrays [11]. Microarray signal intensities were normalized using the SCAN algorithm of SCAN.UPC [12] and the 'pd.huex.1.0.st.v2' annotation [13] Bioconductor packages with brainarray chip description files (CDF, huex10sthsentrez, version 21), yielding one optimized probe-set per gene (gene level summarization) [14]. The TCGA PCa dataset (TCGA-PRAD) contains preprocessed RNA-Seq level 3 data of 497 cases [8]. Based on the TNM-classification of tumors, we stratified both datasets in cases with/without metastasis (corresponding to N0M0 versus N>0 and/or M>0). As incidence and aggressiveness may be different in Africans and Afro-Americans compared to Europeans [1], we filtered – if possible – for men with European ancestry, which was carried out via principal component analysis in the TCGA-PRAD-cohort based on common SNPs identified by parallel exome sequencing. This resulted in a final TCGA-PRAD-cohort of 384 cases (**Fig. 1A**).

Determination of the T2E-status

In the TCGA-PRAD-cohort, the T2E-status was inferred by Torres-García *et al.* based on RNA-Seq split-reads [15]. In the Affymetrix dataset (GSE46691), the T2E-status was inferred from *ERG* expression levels, which show high concordance with the T2E-status [16]. Cases were classified as T2E-positve or -negative if their individual *ERG* expression level was above/below the median *ERG* expression. To reduce the number of potentially misclassified cases, we excluded those 10% with *ERG* expression levels between the 45^{th} and 55^{th} percentile (**Supplementary Fig. 1**).

Processing of microarray and RNA-Seq data

In both cohorts, we separately determined cancer purity with the ESTIMATE algorithm [17]. Only those cases with a consensus purity estimation (CPE) of >60% corresponding to TCGA standard (http://cancergenome.nih.gov/cancersselected/biospeccriteria) were kept for downstream analyses (**Supplementary Fig. 2**). Next, we removed cases with <90% gene coverage and those 50% of genes with lowest variance across all samples using the genefilter Bioconductor package [18]. Moreover, transcripts or probesets from both cohorts, which could not be unambiguously annotated with official gene symbols, and genes that were represented in only one cohort were removed. The unity of both cohorts corresponded to 3,068 variably expressed genes for 299 cases from the TCGA-PRAD-cohort and 538 cases from the GSE46691-cohort.

We next stratified both cohorts according to the T2E-status resulting in four sub-cohorts comprising 109 T2E-positive and 190 -negative cases for the TCGA-PRAD-cohort, and 242 T2E-positive and 242 -negative cases for the GSE46691-cohort. We then calculated in each sub-cohort separately the median fold change of each gene between samples with/without metastasis at diagnosis. Subsequently, the mean fold change from the corresponding median fold changes of both cohorts was calculated separately for T2E-positive and -negative cases. This yielded two gene lists comprising the unity of 3,068 genes ranked by their mean fold change in T2E-positive (rGL-pos) and T2E-negative cases (rGL-neg) (**Fig. 1A**).

Gene set enrichment analysis (GSEA)

To identify significantly enriched gene-signatures (normalized enrichment score (NES) >1.6, nominal $P<0.05$ and FDR $q<0.3$) in both preranked lists (rGL-pos and rGL-neg) we employed GSEA (MSigDB v6.2; chemical and genetic perturbations; 1,000 permutations) [19]. To identify common genes across the top 20 significantly enriched gene-signatures (highest NES), we extracted those genes by leading-edge analysis that were involved in >3 gene-

signatures. This approach yielded two new top gene-signature gene lists for T2E-positive and -negative cases (topGL-pos and topGL-neg) (**Fig. 2A**).

For identification of gene-signatures associated with the expression of identified marker genes in T2E-negative cases, GSEA was carried out under the same conditions as described above. For these cases ranked gene lists were generated by calculating for each gene the expression fold gene after stratifying the cohort by their median expression of the given marker gene (*ASPN*, *BGN*, *COL1A1*, *RRM2* or *TYMS*) into a high and low expression subgroup. For each of the resulting five ranked gene lists we compared the identified top 20 gene-signatures from GSEA between their corresponding subgroups (low/high expression of the given marker gene).

Identification of genes significantly associated with metastasis

For all genes in topGL-pos and -neg the significance of differential expression in PCa patients with/without metastasis at diagnosis was determined by Mann-Whitney-U test [20]. All genes were separately tested in both PCa cohorts (TCGA-PRAD and GSE46691). *P* values were not adjusted for multiple comparison (significance for $P<0.05$). Only genes being significantly associated with metastasis in both cohorts were considered for further analyses.

First validation cohort

For validation of survival analyses in the TCGA-PRAD-cohort, we used another GEO dataset (GSE16560) [21] comprising 272 Swedish PCa cases with microarray expression data (6,100 genes) and corresponding clinical information including cancer-specific death and T2E-status (**Supplementary Table 1**).

Survival analysis

Survival analyses were carried out on all samples of the TCGA-PRAD-cohort and in the Swedish validation cohort (GSE16560) for all genes of topGL-pos and topGL-neg using the Kaplan-Meier method and the survival package of R [20,22]. For calculation of event-free survival (EFS; event = death, appearance of a new tumor, metastases, and/or relapse), both cohorts were stratified according to their intratumoral gene expression into quartiles, and *P* values were calculated using a Mantel-Haenszel test by comparing the patient groups with the most extreme gene expressions (highest versus lowest).

To analyze the potential added value of biomarkers in addition to the Gleason score, Kaplan-Meier survival analyses were carried out in the same cohorts (TCGA-PRAD, GSE16560) stratified by a) the T2E-status, b) the Gleason Grading Group (GGG; I-III versus IV/V), and c) the intratumoral gene expression levels of the given gene (low versus high; cut-off = 80^{th} percentile).

Tissue microarrays (TMAs) and immunohistochemistry (IHC)

A well-characterized prostatectomy cohort comprising 135 patients with known T2E-status (**Supplementary Table 2**) diagnosed with PCa at the Institute of Pathology of the University Hospital of Bonn (Germany) was used as a second validation cohort [23]. The TMA cohort was established with ethics approval of the institutional review board of the University Hospital Bonn, which waived the need for written informed consent from the participants [23]. TMAs were constructed from formalin fixed, paraffin embedded archived tissue with up to 5 cores (diameter: 1 mm) of non-necrotic tumor tissue per patient. Antigen retrieval was achieved by ProTaqs IV Antigen-Enhancer (#401602392, Quartett) for RRM2 and ProTaqs IX Antigen-Enhancer (#401603692, Quartett) for TYMS. RRM2 was detected with a specific rabbit-anti-human RRM2 antibody (1:500, 60 min incubation time; HPA056994, Atlas Antibodies; https://www.proteinatlas.org/ENSG00000171848-RRM2/tissue). TYMS was

detected with a specific rabbit-anti-human TYMS antibody (D5B3) (1:100, 60 min incubation time; #9045, Cell Signaling Technology). Both primary antibodies were followed by an anti-rabbit IgG antibody (MP-7401 ImmPress Reagent Kit) and DAB+ chromogen (K3468, Agilent Technologies). Slides were counterstained with hematoxylin Gill's Formula (H-3401, Vector). Evaluation of RRM2 immunoreactivity was possible in all 133/135 patient specimens (98.5%) represented on the TMA; for TYMS, 119/135 patient specimens (88.2%) were evaluable. RRM2 and TYMS immunoreactivities were quantified by an experienced data-blinded uropathologist (YT) as percentage of positive tumor cells (cytoplasmatic expression). The survMisc package for R was used for optimal cut-off selection and Kaplan-Meier survival analyses [20]. The following percentages of positive cells were used as best cut-offs: ≥3% for RRM2, and ≥5.5% for TYMS.

Data availability

Data of the TCGA-PRAD-cohort [8] were downloaded from the TCGA data portal. Two further cohorts are available at GEO under the accession codes GSE46691 [11] and GSE16560 [21]. The remaining data that support the findings of this study are available from the corresponding author upon reasonable request.

RESULTS

T2E-positive and -negative PCa are characterized by distinct metastasis-associated gene-signatures

T2E-positive and -negative PCa constitute distinct molecular subtypes [8,9]. To decipher molecular differences associated with metastasis in either subtype, we analyzed transcriptome profiles with matched clinicopathological data of two public cohorts (TCGA-PRAD and GSE46691). Multiple filtering steps regarding variance and regulation, and determination of the samples' T2E-fusion status led to a unity of 3,068 variably expressed genes (see Methods). Depending on the T2E-status, we created from this set of genes two gene lists ranked by their expression fold change between patients with/without metastasis (rGL-pos and rGL-neg) (**Fig. 1A**). Metastasis was chosen as a surrogate for PCa aggressiveness, because, contrary to other common PCa related clinical records, information on metastasis was publicly available for both cohorts and usually indicates aggressiveness in PCa [4]. GSEA on rGL-pos and -neg showed no overlap between the top 20 significant metastasis-associated gene-signatures in T2E-positive and -negative cases (**Fig. 1B**, **Supplementary Table 3**).

From those top 20 gene-signatures, we extracted genes involved in >3 of them by leading-edge analysis to create two new 'top gene-signature' gene lists (topGL-pos and -neg, **Fig. 2A**). Accordingly, topGL-pos contained 16 genes of rGL-pos, recurrent in significant gene-signatures of T2E-positive cases (**Supplementary Table 4**), whereas topGL-neg contained 74 genes recurrent in significant gene-signatures of T2E-negative (rGL-neg) cases (**Supplementary Table 5**). Only two genes (*RRM2* and *TYMS*) were shared among T2E-positive and -negative cases, but involved in different gene-signatures (**Fig. 2B**).

Apart from these protein coding genes, we explored our transcriptome data for non-coding genes. In the unity of genes from both discovery datasets, we found only 20 non-coding genes comprising lncRNAs, ncRNAs, and miRNAs. However, only one of these non-coding genes

(*DLEU2*) was represented in a single significantly enriched gene-signature (top 20) for T2E-positive or -negative PCa cases, which precluded a comprehensive evaluation of the role of non-coding genes in prognostication of PCa.

Altogether, these results indicated that T2E-positive and -negative PCa are characterized by distinct metastasis-associated gene-signatures.

Different genes are associated with metastasis in T2E-positive and -negative PCa

Next, we separately tested whether all genes of our top gene-signatures gene lists, topGL-pos and -neg (**Fig. 2B**), were significantly differentially expressed depending on the presence of metastasis in the TCGA-PRAD- and GSE46691-cohorts. In T2E-postive cases (topGL-pos), three genes (*GMNN, TROAP* and *WEE1*) out of 16 were significantly higher expressed (P<0.05) in PCa samples with metastasis. In T2E-negative cases (topGL-neg) 29 of 74 genes were significantly (P<0.05) higher expressed in PCa samples with metastasis. We found no overlap of these significantly differentially expressed and metastasis-associated genes between T2E-positive and -negative cases (**Supplementary Tables 4** and **5**). These results further suggested that – depending on the T2E-status – distinct genes are linked to metastasis in PCa.

Identification of subtype-specific prognostic biomarkers

To test whether the identified metastasis-associated genes were correlated with EFS, we performed Kaplan-Meier analyses in two independent cohorts. The first comprised PCa samples from TCGA-PRAD, the second was derived from another independent microarray-based study (GSE16560, first validation cohort) [21]. We only accepted genes as being associated with EFS if they were significantly (P<0.05) and concordantly associated with EFS in both cohorts. While none of the genes identified in screening of T2E-positive cases (topGL-pos) was consistently associated with EFS in both cohorts, seven genes were

consistently associated with EFS in T2E-negative cases (*APOE*, *ASPN*, *BGN*, *COL1A1*, *LY96*, *RRM2* and *TYMS*). For all seven genes, higher expression levels of the respective gene were associated with shorter EFS (**Fig. 3**). Interestingly, the same biomarkers showed no concordant association with EFS in T2E-positive cases. As displayed in **Table 1**, only five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS*) were associated with metastasis and EFS in both discovery cohorts and the first validation cohort, indicating that these genes could be employed for outcome prediction exclusively in T2E-negative PCa.

To explore whether the association of these genes with outcome of T2E-negative cases might be confounded by additional molecular events such as mutations in the *SPOP* gene (around 10% of PCa cases [8,24]), we re-investigated the TCGA-PRAD-cohort for which the *SPOP* mutation status could be inferred from exome sequencing data [8]. However, removal of the 20 cases harboring *SPOP* mutations from the T2E-negative TCGA-PRAD sub-cohort did not affect the significant associations of *ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS* with clinical outcome (not shown), suggesting that *SPOP* mutations do not affect the validity of these biomarkers for T2E-negative PCa cases. Likewise, we tested whether *TP53* or *PTEN* mutations could have impacted our results in the TCGA-PRAD-cohort (overall mutation frequency of 7% and 2%, respectively). In the T2E-negative subcohort, we identified eleven *TP53*- and two *PTEN*-mutated cases. Removal of these cases from this subcohort did not affect the significant associations of *ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS* with clinical outcome (not shown). These results indicated that neither *TP53* nor *PTEN* mutations could have biased our results.

Comparison of gene-signatures associated with T2E-negative PCa stratified by gene expression

Next, we investigated whether T2E-negative PCa cases with high gene expression of *ASPN*, *BGN*, *COL1A1*, *RRM2* or *TYMS* are enriched in different gene-signatures as determined by

GSEA compared with cases with low expression of the corresponding gene. The overlap of the top 20 gene-signatures (**Supplementary Table 6**) identified by GSEA in subgroups with either high or low expression of *ASPN*, *BGN*, *COL1A1*, *RRM2* or *TYMS* ranged from 15% for *COL1A1* to 45% for *RRM2* (average overlap across all five genes: 38%). These relative low overlaps indicate that T2E-negative PCa tumors with high or low expression of the given marker candidate gene may be driven by largely distinct pathways and as such may differ in their (patho)biology.

Validation of RRM2 and TYMS as prognostic biomarkers for T2E-negative cases by IHC

To confirm the T2E-dependent prognostic value of PCa biomarkers, we stained TMAs containing 135 PCa cases by IHC for RRM2 and TYMS as examples, as for both proteins specific antibodies were available. We separately analyzed the biochemical recurrence (BCR)-free survival of T2E-positive and -negative cases stratifying patients by their percentage of RRM2-positive tumor cells (cut-off ≥3%) as well as TYMS-positive tumor cells (cut-off ≥5.5%). In these analyses, we found that patients with T2E-negative PCa exhibiting a high percentage of RRM2-positive tumor cells had significantly worse BCR-free survival than those with low RRM2-postitivity (P=0.005) (**Fig. 4A**). Likewise, we observed a significantly lower BCR-free survival rate for patients with T2E-negative PCa that presented a high percentage of TYMS-positive tumor cells (P=0.004) (**Fig. 4B**). In contrast, no association of either RRM2- or TYMS-positivity with BCR-free survival was found in T2E-positive cases. These results provided further evidence that the prognostic value of biomarkers in PCa depends on the T2E-status, and suggested that 'pooled' analyses ignoring the T2E-status may obscure outcome prediction.

Subtype-specific biomarkers add prognostic information to Gleason grading

One of the most widely used predictors for patient outcome in PCa is the established Gleason grading system which reforms the Gleason score into five new Gleason Grading Groups (GGG; I-V) [25] that proved to be of high prognostic significance in large cohorts [26,27]. However, risk-prediction for individual PCa patients based on Gleason grading still remains limited [28,29].

To test whether our identified biomarkers may add prognostic information to the Gleason grading, we performed compared Kaplan-Meier analyses for which we stratified both cohorts (TCGA-PRAD, GSE16560) by the T2E-status and subsequently by the GGG (I-III vs IV/V). As expected, we observed in both cohorts a significant (P<0.002) association of worse EFS with high GGG (IV/V) regardless of the T2E-status (**Fig. 5A**). We next explored whether further subgrouping by the potential subtype-specific biomarkers would add prognostic information to the GGG. As displayed in **Fig. 5B**, high *RRM2* and *TYMS* expression was associated with significantly worse outcome in both GGG-low (I-III) and GGG-high (IV/V) patients if PCa tumors were T2E-negative. Strikingly, this additive prognostic effect was entirely absent in both cohorts in T2E-postive cases (**Fig. 5B**). Less strong effects were observed for *ASPN*, *BGN*, and *COL1A1*, which showed either statistical trends or reached statistical significance only in one cohort (**Fig. 5B**). A summary of the results is given in **Supplementary Table 7**.

Taken together, these results indicated that at least two genes (*RRM2*, *TYMS*) of our five biomarker candidates can add prognostic information to routine Gleason grading for T2E-negative patients.

DISCUSSION

Prior studies showed that T2E-positive PCa are associated with specific germline susceptibility variants and epigenetic profiles providing evidence that T2E-positive and -negative PCa constitute distinct molecular and perhaps clinical subtypes [8,9]. We hypothesized that differentially expressed genes involved in distinct gene-signatures may be associated with tumor progression in T2E-positive and -negative PCa, and that prognostic biomarkers may be only relevant in the context of a specific molecular subtype.

To explore such molecular differences, we analyzed PCa transcriptomes and matched clinical data of two large cohorts (TCGA-PRAD and GSE46691). Applying several filtering steps and enrichment analyses, we identified the top 20 metastasis-associated gene-signatures for T2E-positive and -negative cases. Strikingly, these gene-signatures showed no overlap, emphasizing that T2E-positive and -negative PCa are distinct molecular subtypes that take different routes on disease progression [8,9]. From these subtype-specific gene-signatures, we extracted overrepresented genes (topGL-pos and -neg) of which five (*ASPN*, *BGN*, *COL1A1*, *RRM2*, *TYMS*) proved to be of high value for risk-prediction exclusively in T2E-negative PCa. These results imply that biomarkers for risk-prediction in PCa should be employed dependent on the PCa-subtype to maximize their prognostic power.

For example, Asporin (*ASPN*) and Biglycan (*BGN*) [30] are both known to be associated with PCa progression [31] and poor prognosis [32]. Our results confirm these previous observations but highlight that they have only prognostic value for T2E-negative cases. Jacobsen *et al.* additionally reported that *BGN* expression may be related to the presence of the T2E-fusion [32]. However, our results showed that in T2E-positive PCa, *BGN* is not involved in the top gene-signatures associated with metastasis, unlike in T2E-negative PCa.

The protein product of the *COL1A1* gene (collagen type I alpha 1), which is a major constituent of the extracellular matrix and connective tissues [30] has hitherto not been

reported to be linked with outcome of PCa patients rendering *COL1A1* a novel potential biomarker for T2E-negative PCa.

RRM2 (ribonuclease reductase regulatory subunit M2) plays a role in DNA synthesis [30], and its overexpression can promote tumor progression [33]. In fact, a study not distinguishing molecular PCa-subtypes suggested that *RRM2* overexpression may be associated with PCa progression [10]. Our findings made on the mRNA and protein level are in line with these findings with the important refinement that RRM2 has strong prognostic power in T2E-negative cases, while having no prognostic value in T2E-positive cases as confirmed in four independent PCa cohorts.

Similar observations were made for TYMS (thymidylate synthetase), which is involved in DNA replication and repair [30] and reported to correlate with worse outcome in PCa [34]. We observed that T2E-negative patients had significantly higher risk for short EFS with high *TYMS* expression – an effect that was absent in T2E-positive cases.

In another pathway analysis focusing only on T2E-negative cases, we identified different gene-signatures for cases with high and low expression of *ASPN*, *BGN*, *COL1A1*, *RRM2* or *TYMS*. The limited average overlap of only 38% between the top 20 gene-signatures in cases with high or low expression of the given marker gene may indicate that these PCa tumors differ in their (patho)biology.

In accordance with our finding that the T2E-status, which is *per se* not a strong prognostic biomarker, is crucially determining the prognostic value of other biomarkers, it has been shown that the proliferation marker Ki-67 is especially prognostic in T2E-negative cases [35,36].

A common clinicopathological marker used in the routine clinical setting for PCa risk-prediction is the Gleason score or the recently established Gleason grading system which reforms the Gleason score into five new Gleason Grading Groups (GGG; I-V) [25]. In our comparative survival analyses, the GGG outperforms the identified subtype-specific

biomarkers. However, two subtype-specific biomarkers (*RRM2, TYMS*) proved to add further prognostic information exclusively for T2E-negative cases. Whether the other three biomarkers may have additional prognostic value has to be tested in larger cohorts. Yet, the availability of suitable anti-RRM2, anti-TYMS and anti-ERG antibodies enables a rapid translation of our findings in the clinic through the detection of the T2E-status, the RRM2 and TYMS expression levels by IHC, in conjunction with Gleason grading on routine histology.

Besides T2E-positive PCa, there are emerging additional molecular PCa subtypes characterized by rare *ETS* translocations or mutations in putative driver genes such as *SPOP*, *FOXA1* and *IDH1* [8]. In our analyses, mutations in *SPOP*, which constitutes after T2E-fusions the second most frequent mutated gene in PCa (around 10%) [8,24], had no impact on the validity of *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS* for outcome prediction in T2E-negative cases. However, whether less frequently occurring mutations in other genes such as *FOXA1* and *IDH1* (mutation frequencies: 1.7 and 0.3% in the TCGA-PRAD-cohort, respectively) impact on biomarker prediction remains to be determined in future studies with larger sample size.

Additionally, we investigated common cancer driving mutations in *TP53* and *PTEN*, which are known to be enriched in PCa [24]. With an overall frequency of 7% in the TCGA-PRAD-cohort, *TP53* was equally distributed in T2E-positive and negative cases and did not bias our results of the survival analyses. Similarly, the number of *PTEN*-mutated cases (overall frequency of 2%) was negligible in the TCGA-PRAD-cohort.

The recently developed genomic Decipher test for PCa, which was based on one of the studies used here (GSE46691 [11]), enables risk-stratification of PCa patients after surgery by evaluating the expression pattern of 22 genes [11,37], which was confirmed by multiple studies in the clinical setting [38–40]. Interestingly, none of our identified subtype-specific biomarkers is among the 22 Decipher genes, probably because this test does not discriminate between T2E-positive and -negative cases. Similarly, other genomic tests such as Oncotype

DX and Prolaris do not consider the molecular PCa-subtype [41,42], but have a concordance between their tested markers and our identified T2E-negative specific markers. While Prolaris is testing among 32 markers also for *RRM2* [41], Oncotype Dx has tests for 22 transcripts including *BGN* and *COL1A1* [42]. Unfortunately, a direct comparison between the predictive genes of each of these genomic tests and our candidate genes was not possible, because the unity of our variably expressed genes only covered a fraction of the genes necessary for these tests. Thus, it remains to be explored if and how subtype-specific prognostic genes affect the accuracy of such tests when including information on the T2E-status.

Finally, it remains to be determined whether the T2E-status is the only factor influencing the differential expression and/or activity of *ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS* in PCa, or whether other alterations, e.g. on the epigenetic level, may play a role in regulation of these genes.

CONCLUSIONS

Our study exemplifies the power of integrating comprehensive 'omics' and clinical data to identify subtype-specific biomarkers in PCa, and suggests that the T2E-status should be considered when applying prognostic biomarkers to improve risk-prediction of PCa patients in personalized medicine.

Acknowledgements

We thank Mrs. A. Sendelhofert and A. Heier for their expert technical assistance.

Funding

The laboratory of T.G.P.G. is supported by LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative, by grants from the 'Mehr LEBEN für krebskranke Kinder – Bettina-Bräu-Stiftung', the Dr. Leopold und Carmen Ellinger Foundation, the Matthias-Lackas Foundation, the Kind-Philipp Foundation, the Friedrich-Baur Foundation, the Wilhelm Sander-Foundation (2016.167.1), the Dr. Rolf M. Schwiete-Foundation, the Gert und Susanna Mayer-Foundation, the Deutsche Forschungsgemeinschaft (DFG 391665916) and by the German Cancer Aid (DKH-70112257). The sponsors had no role in study design and interpretation of the results.

Authors' contributions

Study concept and design was done by JSG, AB and TGPG. Data was acquired by JSG, MFO and TGPG. JSG and TCPG analyzed and interpreted the data of this study. The prostatectomy cohort established at the University Hospital of Bonn, its immunohistochemistry analysis was conducted by YT and GK. JSG and TGPG wrote the manuscript. AB, KS, MMN, MFO, TK, CS, YT and GK critically revised the manuscript for important intellectual content. Administrative, technical, or material support was given by TK, CS, LRP, FSW, SS, JM, MMLK, TLBH, JL, GS, AM, SO, FCA, GK. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. Sathianathen NJ, Konety BR, Crook J, Saad F, Lawrentschuk N. Landmarks in prostate cancer. Nature Reviews Urology. Nature Publishing Group; 2018;15:627–42.
2. Cancer Research UK. Cancer Research UK, prostate cancer [Internet]. 2018 [cited 2018 Jan 14]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/survival#heading-Four
3. Schapira MM, Lawrence WF, Katz DA, McAuliffe TL, Nattinger AB. Effect of treatment on quality of life among men with clinically localized prostate cancer. Medical care. 2001;39:243–53.
4. Wiklund F. Prostate cancer genomics: can we distinguish between indolent and fatal disease using genetic markers? Genome medicine. BioMed Central; 2010;2:45.
5. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. American Association for Cancer Research; 2016;25:16–27.
6. Daskivich TJ, Lai J, Dick AW, Setodji CM, Hanley JM, Litwin MS, et al. Variation in treatment associated with life expectancy in a population-based cohort of men with early-stage prostate cancer. Cancer. 2014;120:3642–50.
7. Tomlins SA, Bjartell A, Chinnaiyan AM, Jenster G, Nam RK, Rubin MA, et al. ETS Gene Fusions in Prostate Cancer: From Discovery to Daily Clinical Practice. European Urology. Elsevier; 2009;56:275–86.
8. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163:1011–25.
9. Penney KL, Pettersson A, Shui IM, Graff RE, Kraft P, Lis RT, et al. Association of Prostate Cancer Risk Variants with TMPRSS2:ERG Status: Evidence for Distinct Molecular Subtypes. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2016;25:745–9.
10. He Z, Tang F, Lu Z, Huang Y, Lei H, Li Z, et al. Analysis of differentially expressed genes, clinical value and biological pathways in prostate cancer. American journal of translational research. 2018;10:1444–56.
11. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and Validation of a Prostate Cancer Genomic Classifier that Predicts Early Metastasis Following Radical Prostatectomy. PLoS ONE. 2013;8.
12. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. 2012;
13. Carvalho B. pd.huex.1.0.st.v2: Platform Design Info for Affymetrix HuEx-1_0-st-v2. 2015.
14. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic acids research. Oxford University Press; 2005;33:e175.
15. Torres-García W, Zheng S, Sivachenko A, Vegesna R, Wang Q, Yao R, et al. PRADA: pipeline for RNA sequencing data analysis. Bioinformatics (Oxford, England). 2014;30:2224–6.
16. Perner S, Rupp NJ, Braun M, Rubin MA, Moch H, Dietel M, et al. Loss of SLC45A3 protein (prostein) expression in prostate cancer is associated with SLC45A3-ERG gene rearrangement and an unfavorable clinical course. Int J Cancer. 2013;132:807–12.

17. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. Nature Publishing Group; 2013;4:2612.
18. Gentleman R, Carey V, Huber W, Hahne F. genefilter: methods for filtering genes from high-throughput experiments. 2017.
19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. 2005; 102:15545-50
20. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016.
21. Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y, et al. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. BMC medical genomics. 2010;3:8–8.
22. Therneau TM. A Package for Survival Analysis in S. 2015.
23. Uhl B, Gevensleben H, Tolkach Y, Sailer V, Majores M, Jung M, et al. PITX2 DNA Methylation as Biomarker for Individualized Risk Assessment of Prostate Cancer in Core Biopsies. J Mol Diagn. 2017;19:107–14.
24. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J-P, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44:685–9.
25. Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system: Prognostic Gleason grade grouping. BJU International. 2013;111:753–60.
26. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P. Prognostic value of the Gleason score in prostate cancer. BJU Int. 2002;89:538–42.
27. Andrén O, Fall K, Franzén L, Andersson S-O, Johansson J-E, Rubin MA. How well does the Gleason score predict prostate cancer death? A 20-year followup of a population based cohort in Sweden. J Urol. 2006;175:1337–40.
28. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer. 2008;113:3075–99.
29. Capitanio U, Briganti A, Gallina A, Suardi N, Karakiewicz PI, Montorsi F, et al. Predictive models before and after radical prostatectomy. The Prostate. 2010; 70:1371-1378
30. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;44:D733–45.
31. Rochette A, Boufaied N, Scarlata E, Hamel L, Brimo F, Whitaker HC, et al. Asporin is a stromally expressed marker associated with prostate cancer progression. British Journal of Cancer. 2017;116:775–84.
32. Jacobsen F, Kraft J, Schroeder C, Hube-Magg C, Kluth M, Lang DS, et al. Up-regulation of Biglycan is Associated with Poor Prognosis and PTEN Deletion in Patients with Prostate Cancer. Neoplasia. 2017;19:707–15.
33. Zhou BS, Tsai P, Ker R, Tsai J, Ho R, Yu J, et al. Overexpression of transfected human ribonucleotide reductase M2 subunit in human cancer cells enhances their invasive potential. Clinical & experimental metastasis. 1998;16:43–9.
34. Burdelski C, Strauss C, Tsourlakis MC, Kluth M, Hube-Magg C, Melling N, et al. Overexpression of thymidylate synthase (TYMS) is associated with aggressive tumor features and early PSA recurrence in prostate cancer. Oncotarget. 2015;6:8377–87.

35. Goltz D, Montani M, Braun M, Perner S, Wernert N, Jung K, et al. Prognostic relevance of proliferation markers (Ki-67, PHH3) within the cross-relation of ERG translocation and androgen receptor expression in prostate cancer. Pathology. 2015;47:629–36.
36. Karnes RJ, Cheville JC, Ida CM, Sebo TJ, Nair AA, Tang H, et al. The ability of biomarkers to predict systemic progression in men with high-risk prostate cancer treated surgically is dependent on ERG status. Cancer research. 2010;70:8994–9002.
37. Karnes RJ, Bergstralh EJ, Davicioni E, Ghadessi M, Buerki C, Mitra AP, et al. Validation of a Genomic Classifier that Predicts Metastasis Following Radical Prostatectomy in an At Risk Patient Population. Journal of Urology. 2013;190:2047–53.
38. Karnes RJ, Choeurng V, Ross AE, Schaeffer EM, Klein EA, Freedland SJ, et al. Validation of a Genomic Risk Classifier to Predict Prostate Cancer-specific Mortality in Men with Adverse Pathologic Features. European Urology. 2018;73:168–75.
39. Dalela D, Santiago-Jiménez M, Yousefi K, Karnes RJ, Ross AE, Den RB, et al. Genomic Classifier Augments the Role of Pathological Features in Identifying Optimal Candidates for Adjuvant Radiation Therapy in Patients With Prostate Cancer: Development and Internal Validation of a Multivariable Prognostic Model. Journal of Clinical Oncology. 2017;35:1982–90.
40. Klein EA, Haddad Z, Yousefi K, Lam LLC, Wang Q, Choeurng V, et al. Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology. 2016;90:148–52.
41. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. The Lancet Oncology. 2011;12:245–55.
42. Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay – a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genomics. 2013;14:690.

FIGURE LEGENDS

Fig. 1. T2E-positive and -negative PCa are characterized by distinct metastasis-associated gene-signatures

A) Schematic displaying the processing pipeline of the transcriptome data from the TCGA-PRAD- and, GSE46691-cohorts, and the generation of differentially ranked gene lists (rGL-pos and -neg)

B) Venn diagram showing the top 20 significant gene-signatures as identified by GSEA of rGL-pos and -neg.

Fig. 2. T2E-positive and -negative PCa are characterized by distinct metastasis-associated genes

A) Schematic of the analysis pipeline to identify recurrent genes in top metastasis-associated gene-signatures in T2E-positive and -negative cases. LEA, leading-edge analysis.

B) Venn diagram showing the overlap of recurrent genes in top metastasis-associated gene-signatures in T2E-positive and -negative cases.

Fig. 3. The prognostic value of identified biomarkers depends on the T2E-status

Kaplan-Meier survival plots derived from either T2E-positive or -negative samples from TCGA-PRAD- and GSE16560-cohorts for significantly event-free survival (EFS)-associated genes (*APOE*, *ASPN*, *BGN*, *COL1A1*, *LY96*, *RRM2* and *TYMS*) of topGL-neg. Patients were stratified by their quartile intratumoral gene expression levels of the given gene. *P* values were calculated between the lowest (Q1) and highest (Q4) gene expression quartiles using a Mantel-Haenszel test.

Fig. 4. Validation of (A) RRM2 and (B) TYMS as prognostic biomarker for T2E-negative cases by IHC

Top A/B: Representative micrographs of T2E-positive and -negative PCa stained for (A) RRM2 and (B) TYMS by IHC. Scale bars = 50 µm for 10× and 40× magnification, respectively.

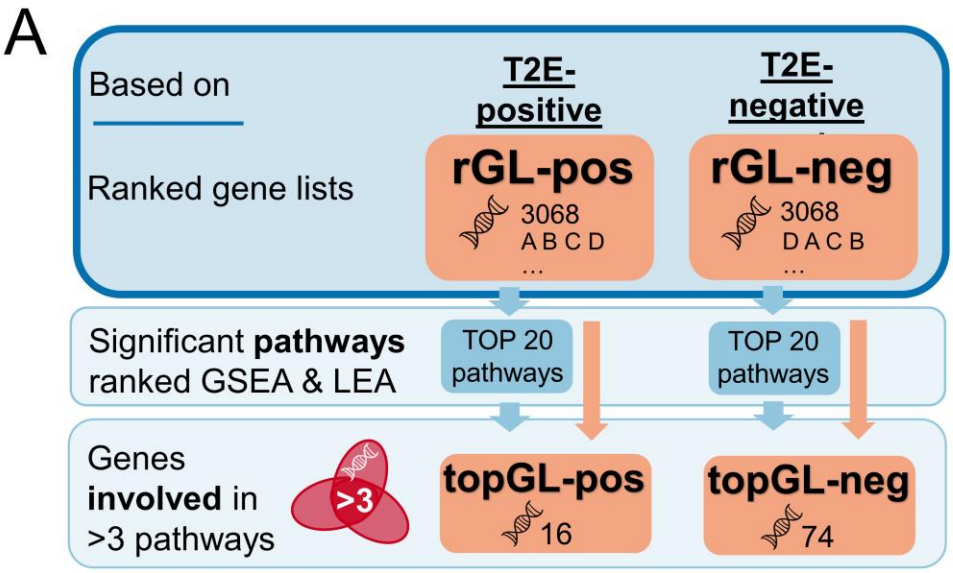
Bottom A/B: Kaplan-Meier analysis of biochemical relapse (BCR)-free survival of T2E-positive and -negative cases stratified by their best cut-off for (A) RRM2-positive tumor cells (≥3%) and (B) TYMS-positive tumor cells (≥5.5%). Mantel-Haenszel test.

Fig. 5. Subtype-specific biomarkers add prognostic information to Gleason grading

A) Kaplan-Meier analysis of EFS for either T2E-positive or -negative cases from TCGA-PRAD- and GSE16560-cohorts. Patients were stratified by their Gleason Grading Group (GGG). Mantel-Haenszel test.

B) Kaplan-Meier analysis of EFS of cases from the TCGA-PRAD- and GSE16560-cohorts stratified by their T2E-status, the GGG (as in A) and by high or low expression (cut-off = 80^{th} percentile) of the indicated biomarker. *P* values (**Supplementary Table 7**) were calculated with a Mantel-Haenszel test between high and low biomarker expression separately for high (IV/V; red color) and low (I-III; blue color) GGG.

Figure 2 Gerke *et al.*



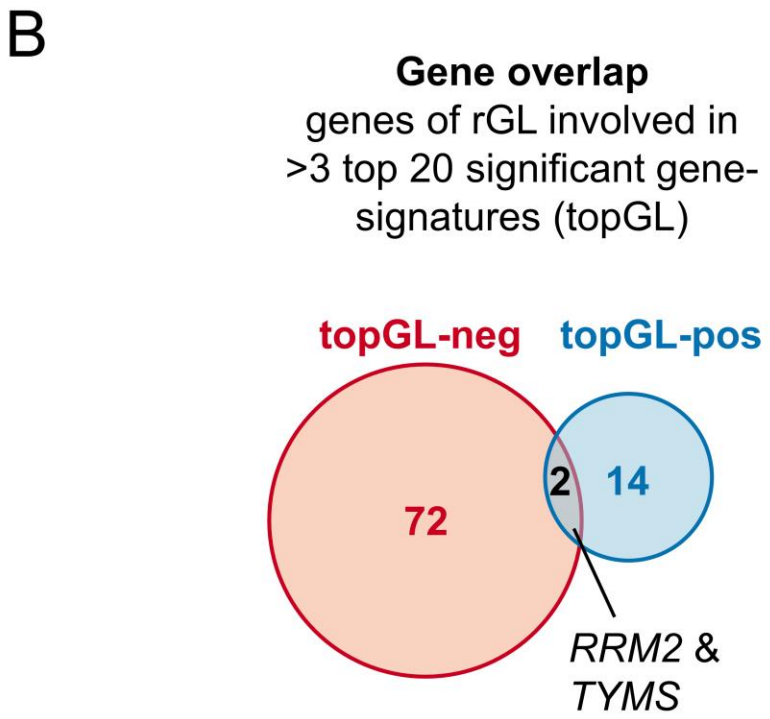


Figure 3 Gerke *et al.*

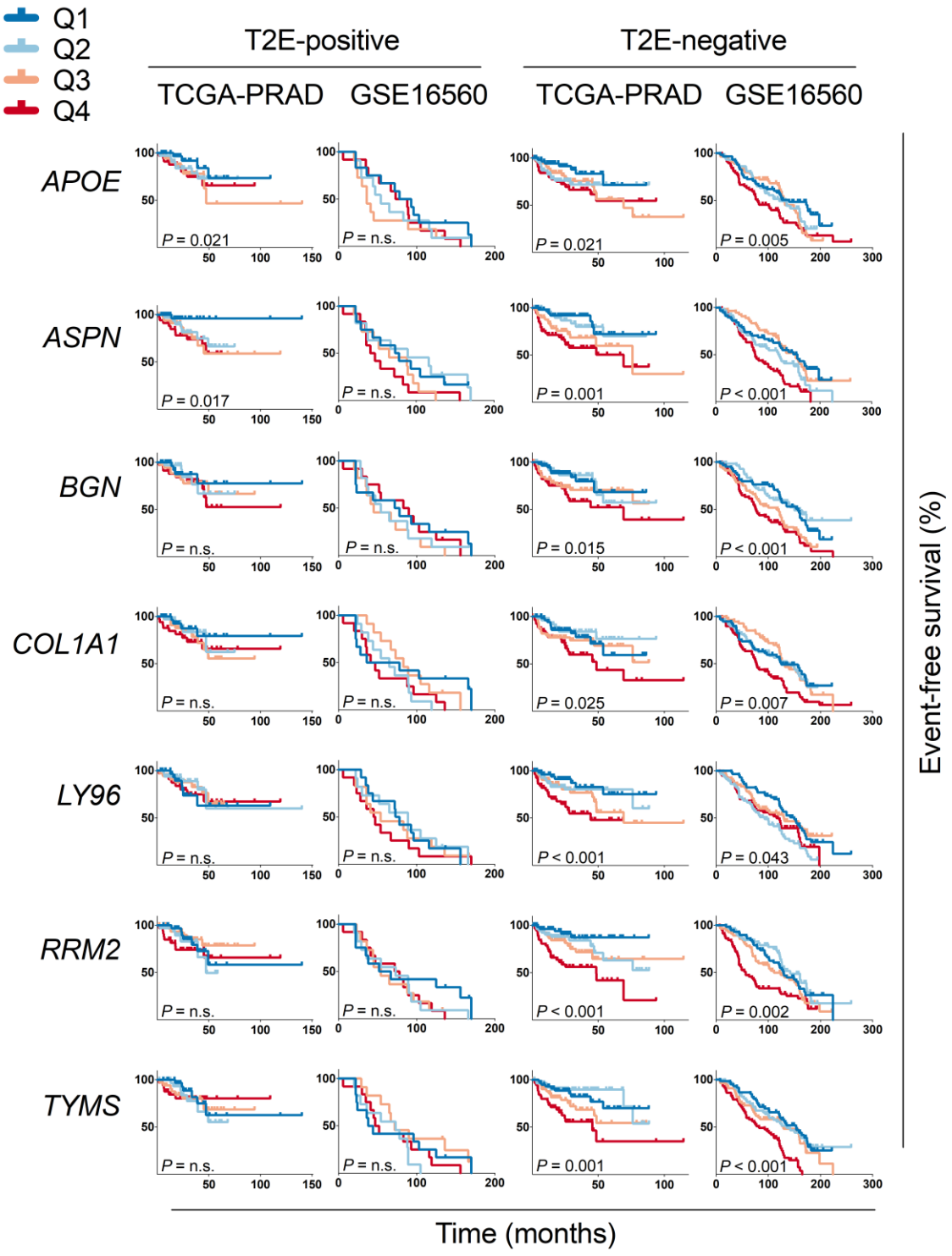


Figure 4 Gerke *et al.*

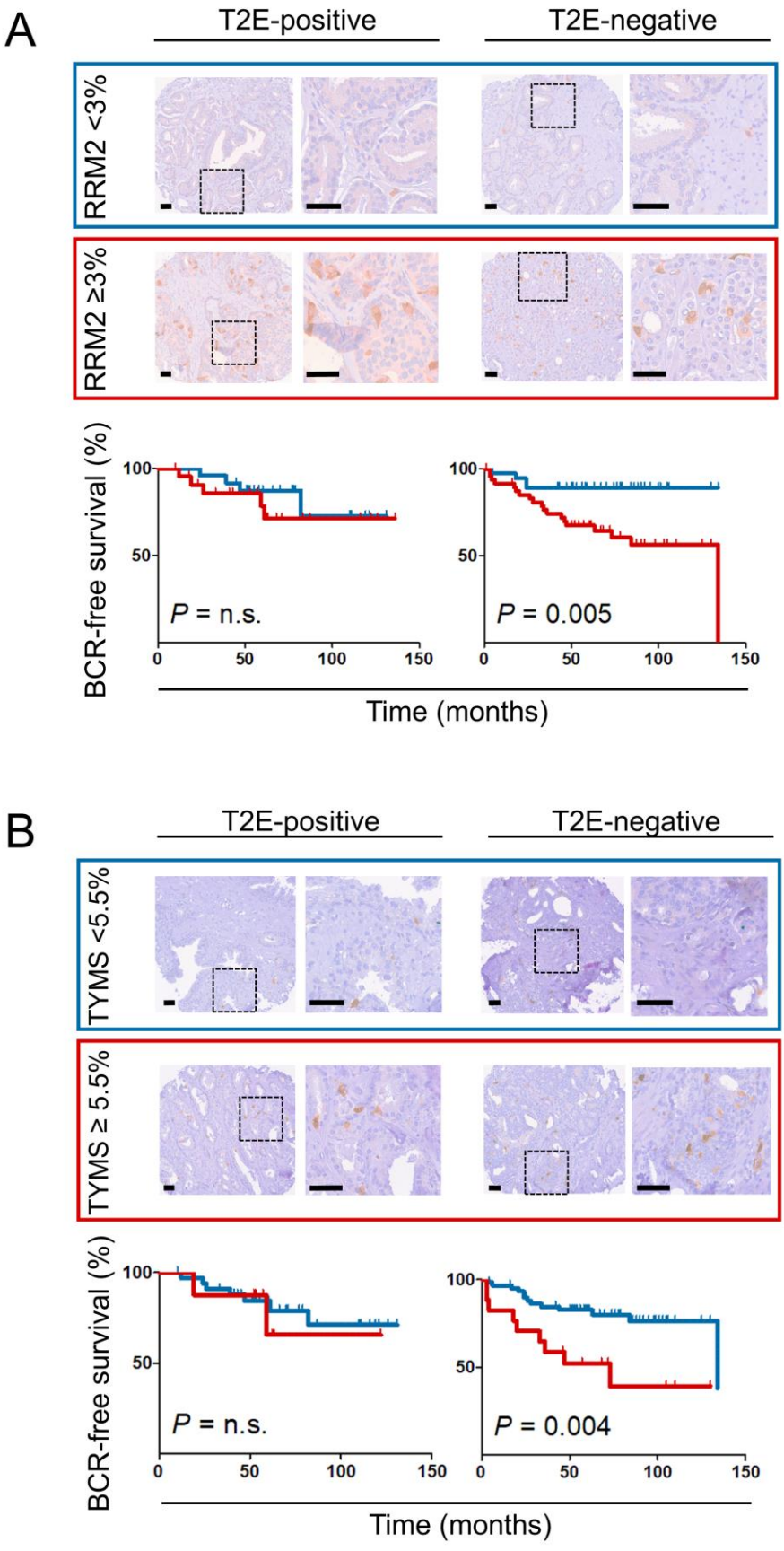


Figure 5 Gerke *et al.*

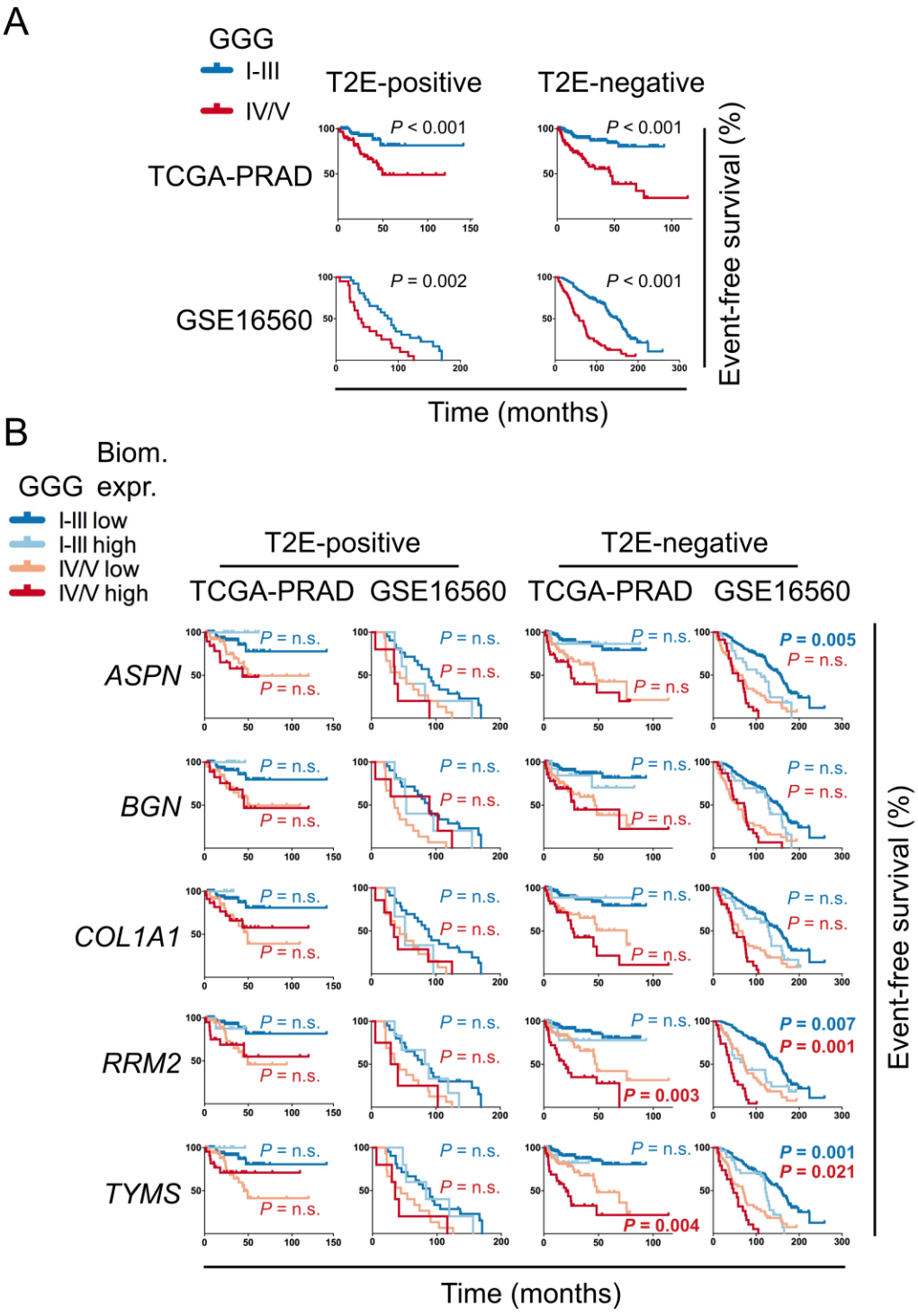


TABLE LEGEND

Table 1. Result summary of genes in topGL-neg and topGL-pos that passed ≥1 of our tests (association test and survival analysis) for all cohorts, as well as those two genes (*RRM2*, *TYMS*) which were included in both gene lists (topGL-pos and -neg) (significant genes

extracted from **Supplementary Tables 3 and 4**). Genes being significant in all tests are highlighted in bold font.

Table 1. Result summary of genes in topGL-neg and topGL-pos that passed ≥1 of our tests (association test and survival analysis) for all cohorts, as well as those two genes (*RRM2*, *TYMS*) which were included in both gene lists (topGL-pos and -neg) (significant genes extracted from **Supplementary Tables 3 and 4**). Genes being significant in all tests are highlighted in bold font.

		GSE46691	TCGA			GSE16560	
Gene list	**Gene**	***P* value (metastasis)**	***P* value (metastasis)**	***P* value (EFS)**	**Expression level associated with long EFS**	***P* value (EFS)**	**Expression level associated with long EFS**
topGL-pos	*GMNN*	<0.001	0.005	n.s.	low	n.s.	high
	RRM2	0.005	n.s.	n.s.	high	n.s.	low
	TROAP	0.021	0.032	n.s.	low	n.s.	high
	TYMS	<0.001	n.s.	n.s.	high	n.s.	low
	WEE1	<0.001	0.002	n.s.	low	n.s.	high
topGL-neg	*APOE*	n.s.	0.011	0.021	low	0.005	low
	ASPN	<0.001	<0.001	0.001	low	<0.001	low
	BGN	0.003	<0.001	0.015	low	<0.001	low
	COL1A1	<0.001	<0.001	0.025	low	0.007	low
	RRM2	0.044	<0.001	<0.001	low	0.002	low
	LY96	n.s.	<0.001	0.001	low	0.043	low
	TYMS	0.009	0.018	0.001	low	<0.001	low