

Gerke *et al.*

Biomarker-based outcome prediction in prostate adenocarcinoma depends on the *TMPRSS2-ERG* status

Julia S. Gerke¹, Martin F. Orth¹, Yuri Tolkach², Laura Romero-Pérez¹, Fabienne Wehweck¹, Stefanie Stein¹, Julian Musa¹, Maximilian M. L. Knott^{1,3}, Tilman L. B. Hölting¹, Jing Li¹, Giuseppina Sannino¹, Aruna Marchetto¹, Shunya Ohmura¹, Florencia Cidre-Aranaz¹, Martina Müller-Nurasyid^{4,5}, Konstantin Strauch^{4,5}, Christian Stief⁶, Glen Kristiansen², Thomas Kirchner^{3,7,8}, Alexander Buchner⁶, Thomas G. P. Grünewald^{1,3,7,8,§}

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 Urologic Clinic und Polyclinic, Clinical Center of the University of Munich, Munich, Germany

7 German Cancer Consortium (DKTK), partner site Munich, Germany

8 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

Word count: 2479

Display items: 4 Figures, 1 Table, 2 Supplementary Figures, 4 Supplementary Tables

Keywords Prostate carcinoma, *TMPRSS2-ERG*, metastasis, RNA-Seq, microarray, gene-signature, biomarkers

Gerke *et al.*

ABSTRACT

Background: Prostate adenocarcinoma (PCa) with/without the *TMPRSS2-ERG* (T2E)-fusion represent distinct molecular subtypes.

Objective: To investigate gene-signatures associated with metastasis in T2E-positive and -negative PCa, and to identify and validate subtype-specific prognostic biomarkers.

Design, setting and participants: Gene expression and clinicopathological data of two discovery PCa cohorts (total $n=783$) were separately analyzed regarding the T2E-status. Selected subtype-specific biomarkers were validated in two additional cohorts (total $n=405$).

Outcome measurements and statistical analysis: From both discovery cohorts, we generated two gene lists ranked by their differential intratumoral expression in patients with/without metastases stratified by T2E-status, which were subjected to gene set enrichment and leading-edge analyses. The resulting top 20 gene-signatures of both gene lists associated with metastasis were analyzed for overlaps between T2E-positive and -negative cases. Genes shared by several functional gene-signatures were tested for their association with event-free survival using the Kaplan-Meier method in a validation cohort. Immunohistochemistry was performed in another validation cohort.

Results and limitations: Metastatic T2E-positive and -negative PCa are characterized by different gene-signatures. Five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2* and *TYMS*) were identified whose high expression was significantly associated with worse outcome exclusively in T2E-negative PCa. This was validated in an independent cohort for all genes and additionally for *RRM2* by immunohistochemistry in a separate validation cohort. No prognostic biomarkers were identified exclusively for T2E-positive tumors.

Conclusions: Our study demonstrates that the prognostic value of biomarkers critically depends on the molecular subtype, i.e. the T2E-status, which should be considered when screening for and applying novel prognostic biomarkers for outcome prediction in PCa.

Patient summary: Outcome prediction for PCa is complex. The results of this study highlight that the validity of prognostic biomarkers depends on the molecular subtype, specifically the presence/absence of T2E. The reported new subtype-specific biomarkers exemplify that biomarker-based outcome prediction in PCa should consider the T2E-status.

Gerke *et al.*

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 creating 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 [8–10], 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 [11]. 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 was validated in two additional cohorts. Going beyond prior studies [8,10,11], we show that the T2E-status critically determines the nature of distinct metastasis-associated gene-signatures, and strongly impacts on prognostic biomarkers.

MATERIALS AND 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). The GEO dataset (GSE46691) comprised 545 PCa cases profiled on Affymetrix GeneChip Human Exon 1.0 ST arrays [12]. Microarray signal intensities were normalized using

Gerke *et al.*

the SCAN algorithm of SCAN.UPC [13] and the ‘pd.huex.1.0.st.v2’ annotation [14] Bioconductor packages with brainarray chip description files (CDF, huex10sthsentrez, version 21), yielding one optimized probe-set per gene (gene level summarization) [15]. The TCGA PCa dataset (TCGA-PRAD) contains preprocessed RNA-Seq level 3 data of 497 cases [10]. 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 (**Figure 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 (<http://www.tumorfusions.org/>) [16]. In the Affymetrix dataset (GSE46691), the T2E-status was inferred from *ERG* expression levels, which show high concordance with the T2E-status [17]. Cases were classified as T2E-positive 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 45th and 55th percentile (**Supplementary Figure 1**).

Processing of microarray and RNA-Seq data

In both cohorts, we separately determined cancer purity with the ESTIMATE algorithm [18]. Only 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 Figure 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 [19]. 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

Gerke *et al.*

metastasis at diagnosis. Subsequently, the mean fold change from the two corresponding median fold changes was calculated 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) (**Figure 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) [20].

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 yielding two new top gene-signature gene lists for T2E-positive and -negative cases (topGL-pos and topGL-neg) (**Figure 2a**).

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 [21]. 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) comprising 272 Swedish PCa cases with microarray expression data (6,100 genes) and corresponding clinical information including cancer-specific death and T2E-status [22].

Survival analysis

Survival analyses were carried out in the TCGA-PRAD-cohort and 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 [21,23]. 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).

Gerke *et al.*

Tissue microarrays (TMAs) and immunohistochemistry (IHC)

A well-characterized prostatectomy cohort comprising 133 patients with known T2E-status (**Supplementary Table 1**) diagnosed with PCa at the Institute of Pathology of the University Hospital of Bonn (Germany) was used as a second validation cohort [24]. TMAs were constructed from formalin fixed, paraffin embedded archived tissue with up to 5 cores (diameter: 1 mm) of non-necrotic tumor tissue for each patient. For IHC, RRM2 antigens were treated with ProTags IV Antigen-Enhancer (Quartett, Berlin, Germany). 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>), followed by an anti-rabbit IgG antibody (MP-7401 ImmPress Reagent Kit) and DAB+ chromogen (Agilent Technologies). Slides were counterstained with hematoxylin Gill's Formula (H-3401, Vector). RRM2 immunoreactivity was quantified by an experienced data-blinded uropathologist 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 [21].

RESULTS

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

T2E-positive and -negative PCa constitute distinct molecular subtypes [8–10]. 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) (**Figure 1a**). Metastasis was chosen as a surrogate for PCa aggressiveness, because information on metastasis was 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 (**Figure 1b**, **Supplementary Table 2**).

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, **Figure 2a**). Accordingly, topGL-pos contained 16 genes of rGL-pos, overrepresented in significant gene-

Gerke *et al.*

signatures of T2E-positive cases (**Supplementary Table 3**), whereas topGL-neg contained 74 genes overrepresented in significant gene-signatures of T2E-negative (rGL-neg) cases (**Supplementary Table 4**). Only two genes (*RRM2* and *TYMS*) were shared among T2E-positive and -negative cases, but involved in different gene-signatures (**Figure 2b**).

Collectively, 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 (**Figure 2**), were significantly differentially expressed depending on the presence of metastasis in the TCGA-PRAD- and GSE46691-cohorts. In T2E-positive 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. Again, we found no overlap of these significantly differentially expressed and metastasis-associated genes between T2E-positive and -negative cases (**Supplementary Tables 3 and 4**). 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) [22]. 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) were 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, patients of both cohorts had a higher risk for shorter EFS when their tumors expressed high levels of the respective genes (**Figure 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 all three cohorts tested, indicating that these genes could be employed for outcome prediction exclusively in T2E-negative PCa.

Gerke *et al.*

In synopsis, our results implied that the prognostic value of biomarkers critically depends on the T2E-fusion status.

Validation of RRM2 as prognostic biomarker for T2E-negative cases by IHC

To confirm the T2E-dependent prognostic value of PCa biomarkers, we stained a TMA containing 133 PCa cases by IHC for RRM2 as an example, as for RRM2 a specific antibody was 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 $\geq 3\%$). We found that patients with T2E-negative PCa exhibiting a high percentage of RRM2-positive tumor cells had significantly ($P=0.005$) worse BCR-free survival than those with low RRM2-positivity (**Figure 4**). In contrast, no association of RRM2-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.

DISCUSSION

Prior studies showed that T2E-positive PCa are associated with specific germline susceptibility variants [8] and epigenetic profiles [9] providing evidence that T2E-positive and -negative PCa constitute distinct molecular and perhaps clinical subtypes [10]. 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 for a given 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 [8–10]. 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 exclusively for subtype-specific risk-prediction in T2E-negative PCa. These results imply that biomarkers for risk-prediction in PCa should be employed in the corresponding PCa-subtype to maximize their prognostic power.

Gerke *et al.*

For example, Asporin (*ASPN*) and Biglycan (*BGN*) [25] are both known to be associated with PCa progression [26] and poor prognosis [27]. 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 [27]. However, our study 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* (collagen type I alpha 1), which is a major constituent of the extracellular matrix and connective tissues [25] has hitherto not been reported to be connected to outcome of PCa patients rendering *COL1A1* a novel potential biomarker for T2E-negative PCa.

RRM2 (ribonuclease reductase M2) plays a role in DNA synthesis [25], and its overexpression can promote tumor progression [28]. In fact, a study not distinguishing molecular PCa-subtypes suggested that *RRM2* overexpression may be associated with PCa progression [11]. 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 [25] and reported to correlate with worse outcome in PCa [29]. 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 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 [30,31]. However, Ki-67 (encoded by *MKI67*) was not prioritized in our screen because of discordant results in both discovery cohorts. 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* [10]. Whether these alterations also impact on biomarker prediction remains to be determined in future studies.

CONCLUSIONS

Our findings suggest that the T2E-status should be considered when applying prognostic biomarkers to improve risk stratification of PCa patients.

Gerke *et al.*

AUTHOR CONTRIBUTIONS

Julia S. Gerke and Thomas G. P. Grünewald had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Gerke, Grünewald, Buchner

Acquisition of data: Gerke, Orth, Tolkach, Kristiansen, Grünewald

Analysis and interpretation of data: Gerke, Grünewald, Tolkach

Drafting of the manuscript: Gerke, Grünewald

Critical revision of the manuscript for important intellectual content: Buchner, Strauch, Müller-Nurasyid, Orth, Kirchner, Stief, Tolkach, Kristiansen

Statistical analysis: Gerke, Tolkach

Obtaining funding: Grünewald

Administrative, technical, or material support: Kirchner, Stief, Romero-Pérez, Wehweck, Stein, Musa, Knott, Hölting, Li, Sannino, Marchetto, Ohmura, Cidre-Aranaz, Kristiansen

Supervision: Grünewald

Other (specify): None

FINANCIAL DISCLOSURE

Thomas G. P. Grünewald certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (e.g., employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

FUNDING/SUPPORT AND ROLE OF THE SPONSOR

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 Walter Schulz Foundation, the Friedrich-Baur Foundation, the Wilhelm Sander-Foundation (2016.167.1), the Deutsche Forschungsgemeinschaft (DFG 391665916) and by the German

Gerke *et al.*

Cancer Aid (DKH-111886 and DKH-70112257). The sponsors had no role in study design and interpretation of the results.

REFERENCES

- [1] Sathianathen NJ, Konety BR, Crook J, Saad F, Lawrentschuk N. Landmarks in prostate cancer. *Nature Reviews Urology* 2018;15:627–42. doi:10.1038/s41585-018-0060-7.
- [2] Cancer Research UK. Cancer Research UK, prostate cancer 2018. <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/survival#heading=Four> (accessed January 14, 2018).
- [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* 2010;2:45. doi:10.1186/gm166.
- [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* 2016;25:16–27. doi:10.1158/1055-9965.EPI-15-0578.
- [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. doi:10.1002/cncr.28926.
- [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* 2009;56:275–86. doi:10.1016/J.EURURO.2009.04.036.
- [8] 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. doi:10.1158/1055-9965.EPI-15-1078.
- [9] Geybels MS, Alumkal JJ, Luedeke M, Rinckleb A, Zhao S, Shui IM, et al. Epigenomic profiling of prostate cancer identifies differentially methylated genes in TMPRSS2 : ERG fusion-positive versus fusion- negative tumors. *Clinical Epigenetics* 2015. doi:10.1186/s13148-015-0161-6.
- [10] 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. doi:10.1016/j.cell.2015.10.025.
- [11] 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.
- [12] 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. doi:10.1371/journal.pone.0066855.
- [13] Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows 2012. doi:10.1016/j.ygeno.2012.08.003.
- [14] Carvalho B. pd.huex.1.0.st.v2: Platform Design Info for Affymetrix HuEx-1_0-st-v2 2015.

Gerke *et al.*

- [15] 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* 2005;33:e175. doi:10.1093/nar/gni179.
- [16] 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. doi:10.1093/bioinformatics/btu169.
- [17] 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. doi:10.1002/ijc.27733.
- [18] 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* 2013;4:2612. doi:10.1038/ncomms3612.
- [19] Gentleman R, Carey V, Huber W, Hahne F. *genefilter: methods for filtering genes from high-throughput experiments* 2017.
- [20] 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.
- [21] R Core Team. *R: A Language and Environment for Statistical Computing* 2016.
- [22] 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. doi:10.1186/1755-8794-3-8.
- [23] Therneau TM. *A Package for Survival Analysis in S* 2015.
- [24] 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. *The Journal of Molecular Diagnostics* 2017;19:107–14. doi:10.1016/j.jmoldx.2016.08.008.
- [25] 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. doi:10.1093/nar/gkv1189.
- [26] 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. doi:10.1038/bjc.2017.15.
- [27] 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. doi:10.1016/j.neo.2017.06.003.
- [28] 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.
- [29] 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. doi:10.18632/oncotarget.3107.
- [30] 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. doi:10.1097/PAT.0000000000000320.
- [31] 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. doi:10.1158/0008-5472.CAN-10-1358.

Gerke *et al.*

FIGURES AND FIGURE LEGENDS

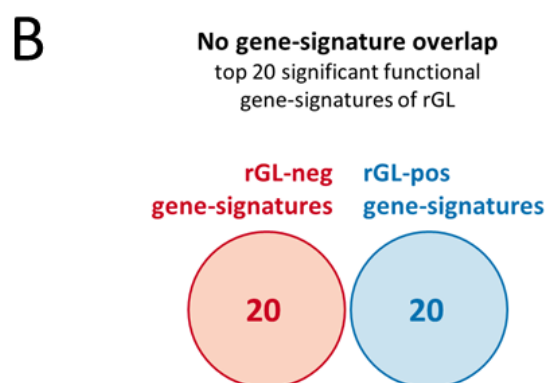
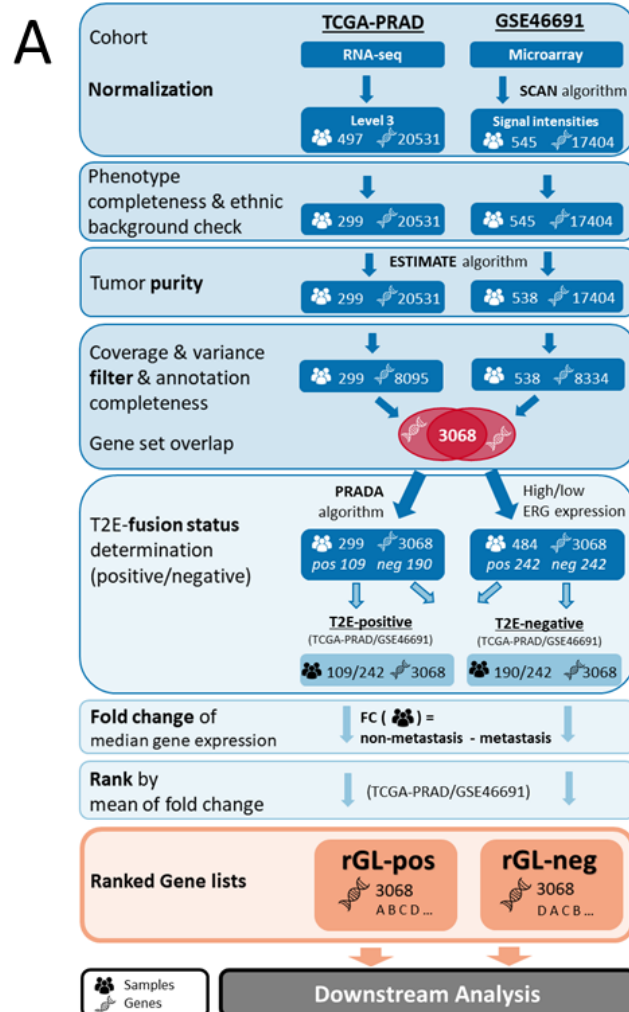


Figure 1. T2E-positive and -negative PCa are characterized by distinct metastasis-associated gene-signatures

A) Pipeline overview depicting the processing transcriptome data from the TCGA-PRAD- and, GSE46691-cohorts, and 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.

Gerke *et al.*

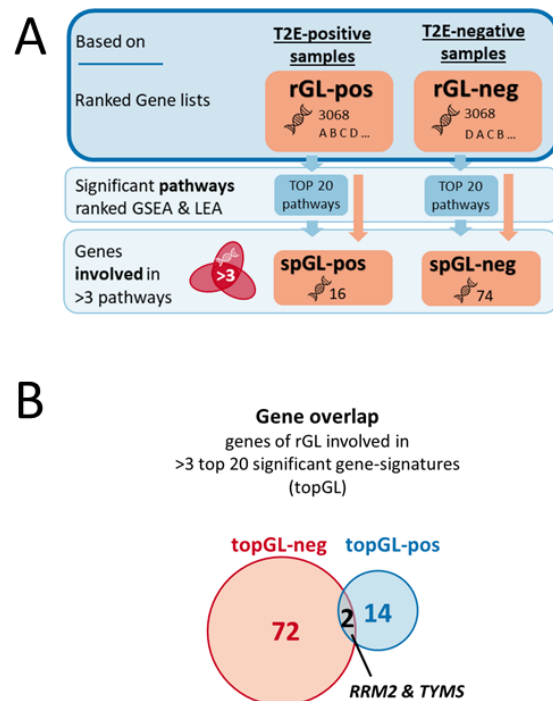


Figure 2. T2E-positive and -negative PCa are characterized by distinct metastasis-associated genes

A) Pipeline overview depicting the identification of overrepresented genes in top metastasis-associated gene-signatures in T2E-positive and -negative cases.

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

Gerke *et al.*

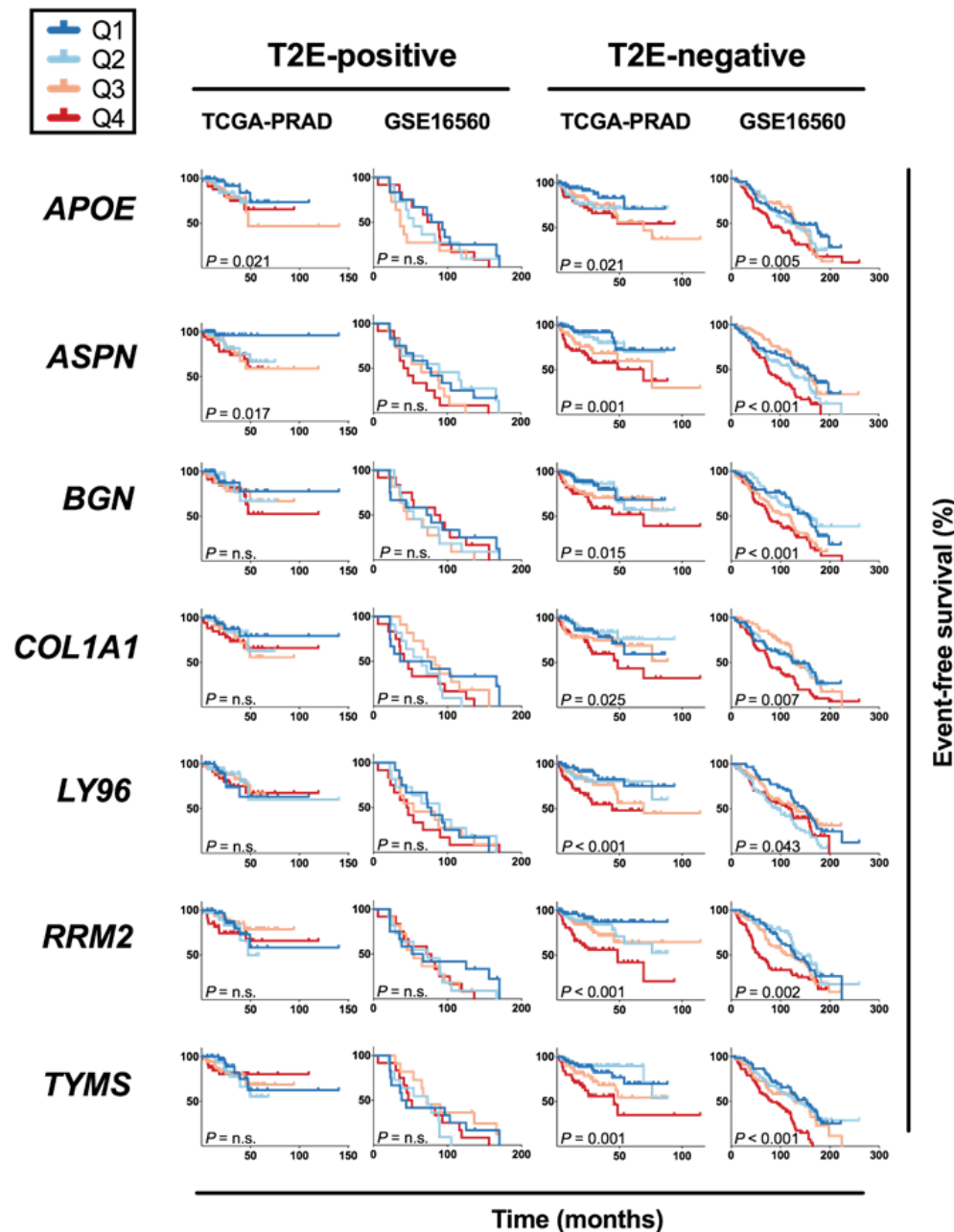


Figure 3. The prognostic value of identified biomarkers depends on the T2E-status

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.

Gerke *et al.*

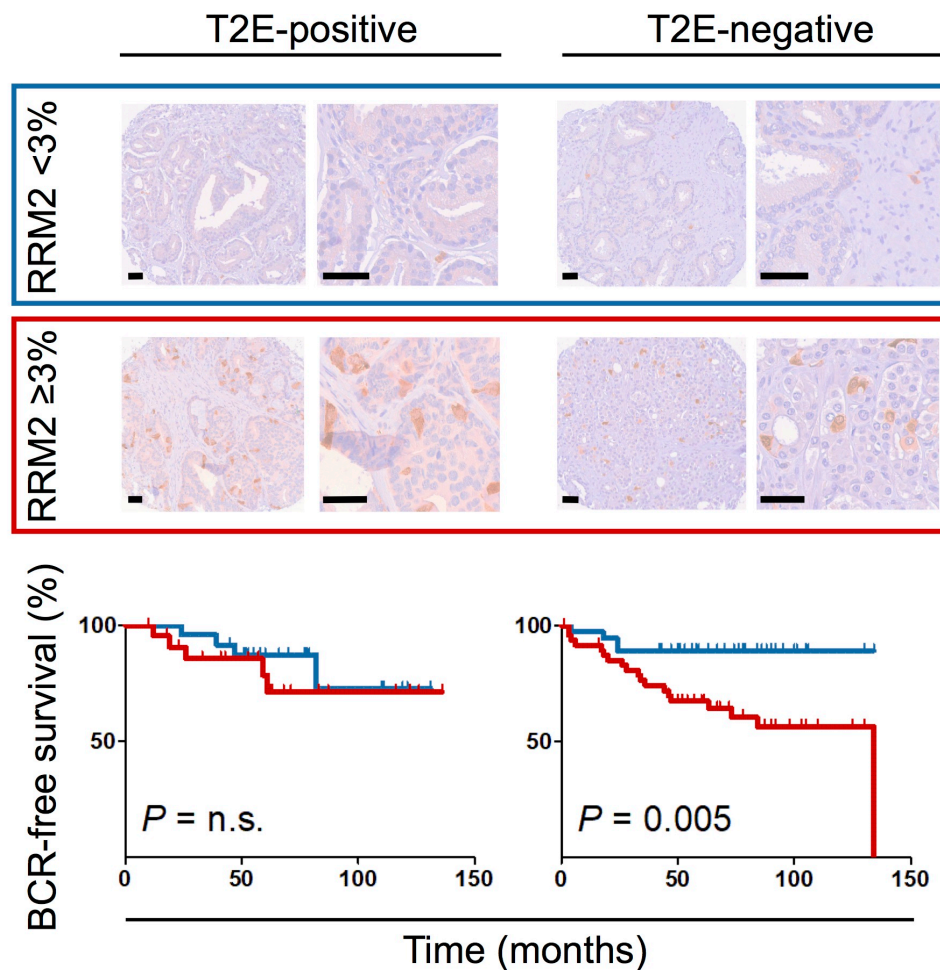
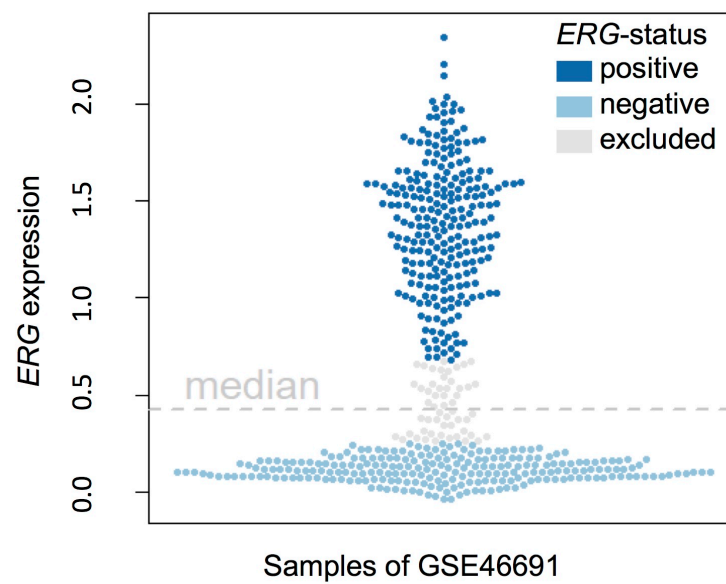


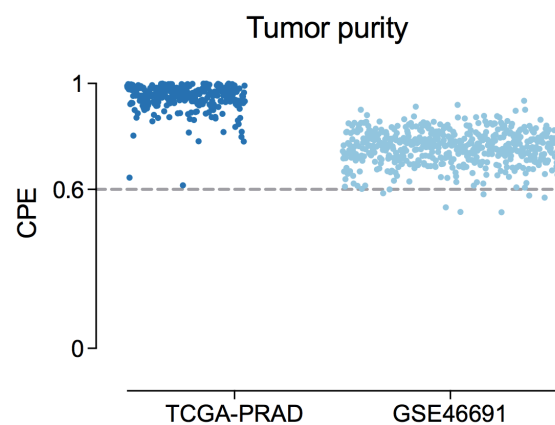
Figure 4. Validation of RRM2 as prognostic biomarker for T2E-negative cases by IHC
Top: Representative micrographs of T2E-positive and -negative PCa stained for RRM2 by IHC. Scale bars = 50μM for 10× and 40× magnification, respectively.
Bottom: Kaplan-Meier analysis of biochemical relapse (BCR)-free survival of T2E-positive and -negative cases stratified by their median percentage of RRM2-positive tumor cells (cut-off ≥3%). Mantel-Haenszel test.

Gerke *et al.*



Supplementary Figure 1

T2E-fusion status in samples of the GSE46691-cohort defined by median *ERG* expression (0.42). Samples ranging with their *ERG* expression between the 45th and 55th percentile (grey) were excluded.



Supplementary Figure 2

Consensus Purity Estimation (CPE) calculated with ESTIMATE algorithm of all samples of the TCGA-PRAD- and GSE46691-cohorts.

Gerke *et al.*

TABLE AND TABLE LEGEND

		GSE46691	TCGA			GSE16560	
Genelist	Gene	<i>P</i> value (metastasis)	<i>P</i> value (metastasis)	<i>P</i> value (EFS)	Expression level associated with long EFS	<i>P</i> 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
	ASP	< 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

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 significant in all tests are highlighted in red color.