From the
Pathological Institute
of the Ludwig-Maximilians-University Munich

Dissertation
zum Erwerb des Doctor of Philosophy (Ph.D.)
an der Medizinischen Fakultät
der Ludwig-Maximilians-Universität München

Assessment of the contribution of germline variation and somatic mutations to prostate cancer progression and prognostication

vorgelegt von
Julia Sophia Gerke
aus München, Deutschland

München, 2020

First supervisor: PD Dr. Dr. med. Thomas Grünewald
Second supervisor: Prof. Dr. Konstantin Strauch

Dean: Prof. Dr. med. dent. Reinhard Hickel

Date of oral defense: 14.12.2020

For my parents, M & W

Contents

Summary

Prostate cancer (PCa) is the second most common cancer for men worldwide. Nevertheless, prevention, treatment, and patient health care improved the probability for men to survive PCa. However, current screenings and imprecise prognostic tests lead to a high number of overdiagnosed and overtreated patients accompanied by adverse effects and health care burden. Therefore, more specific prognostic and predictive tests are necessary to distinguish between benign and aggressive tumor to be able to adapt therapy accordingly.

As a heterogeneous disease, PCa forms several subtypes. Its most prevalent form is characterized by the *TMPRSS2-ERG* (T2E) gene fusion, which appears in around 50% of PCa. Differentiating etiopathology, progression driving pathways, and disease outcome of PCa subtypes make their treatment challenging. However, many established PCa test do not distinguish between subtypes, so far, possibly affecting outcome prediction and treatment decisions.

Recent advances in high throughput technologies allow the extensive generation of genomic and transcriptomic data and pioneered personalized medicine empowering clinical diagnostics and risk prediction. Consequently, identifying suitable biomarkers via genome-wide association studies (GWAS) and gene expression analysis is crucial for the development of gene and variant based prognostic as well as predictive tests.

This thesis assessed the contribution of somatic mutations and germline variation in PCa progression. It emphasized the molecular differences between PCa subtypes by investigating their gene-signatures associated with metastasis, affirming the importance of considering the T2E-status of a patient in research studies as well as in clinical settings. Additionally, subtype specific biomarkers were identified showing that their prognostic value decisively depends on the T2E-status. Furthermore, a risk locus (7q31.33) was detected that harbored five potential tag single nucleotide polymor-

phisms (SNPs) associated with aggressive, high-grade PCa and partially altering *GRM8* expression, thus, implying an influence of germline variants on PCa progression.

These findings were obtained from a meta-analysis of multiple GWAS and a gene expression analysis of PCa subtypes:

GWAS testing for germline association with Gleason grade (GG) were performed on three PCa cohorts including 1,098 Central European men. While two of the cohorts were whole exon/genome sequenced (TCGA-PRAD, ICGC-CA), the third, own cohort ($PCGA^{LMU}$) was genotyped on a microarray. The results were combined in a meta-analysis. One genome-wide significant SNP (rs12537032) was found at the risk locus chr7q31.33, which is associated with aggressive PCa (high GG) and comprises the *GRM8* gene. With fine-mapping, five tag SNPs (rs12537032, rs74999840, rs1910298267, rs76326523, and rs73451279) from two independent signals on the same haploblock could be detected, which were in linkage disequilibrium with each other, respectively. The risk/minor allele of these tag SNPs increased the risk for high-grade PCa up to a factor of three. Located in intronic regions of *GRM8* only the risk allele G of rs73451279 was associated with lower gene expression (via eQTL). Variant effect prediction indicated that the risk allele may induce nonsense-mediated mRNA decay, affecting PCa progression. In survival analysis, however, neither the genotype nor *GRM8* expression was associated with worse PCa outcome. Conditional analysis showed that rs73451279 might not be the only functional variant at this locus affecting mechanisms that drive the development of high-grade PCa.

Based on comprehensive transcriptomic and matched clinicopathological data of two discovery cohorts (n = 783), the gene expression profiles of T2E-positive and -negative PCa were independently investigated to compare metastasis associated gene-signatures regarding their T2E-status. With gene set enrichment analyses distinct gene-signatures characterizing T2E-positive and -negative tumors were detected. Genes frequently involved in these functional gene-signatures were further investigated in a validation cohort (n = 272). Beside being associated with metastasis, five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) were also identified to be associated with event-free survival. In exclusively T2E-negative tumors, their overexpression was significantly associated with worse outcome. *RRM2* and *TYMS* were both evaluated in another validation cohort (n = 135) by immunohistochemistry. Moreover, some of these subtype spe-

cific biomarkers did enhance the Gleason score as current clinicopathological predictor in T2E-negative PCa. In contrast, these observations did not apply for T2E-positive tumors for which no biomarker could be found in this study. Taken together, these findings strongly suggest considering molecular subtypes in combination with the application of prognostic biomarkers to improve outcome prediction in PCa.

Introduction

1.1 | Prostate cancer on a social level

Prostate cancer (PCa) is the second most common cancer in men worldwide [1, 2] and the most common cancer for males in Europe [3]. Despite having the third-highest mortality rate of cancer death in European men [3], PCa is no longer a death sentence due to intense, regular screening [4]. Less than 20% of PCa patients develop aggressive tumors, which are often lethal and require a fast and intense treatment [5], such as radical prostatectomy and radiation therapy. These treatments come along with significant adverse effects [6, 7]. The remaining patients mostly exhibit a slowly growing tumor, which can be safely treated with active surveillance [8]. These men can live with their tumor without ever showing any PCa symptoms. So far, it remains difficult to discriminate indolent from aggressive PCa [9], which is why 23 to 42% of men are over diagnosed and receive unnecessary therapy with significant morbidity [2,10,11]. Overtreatment tends to be a major issue affecting men's life quality and expectancy [12,13], which spurs debates on regular PCa screening, including prostate specific antigen (PSA) measurement [14, 15]. Apart from the adverse effects of current PCa treatment regimens for individual patients, the enormous number of PCa incidences and the numerous overtreatments tend to be a significant socioeconomic and health care burden in the Western world [2, 16]. In fact, compared to radical prostatectomy, treatment of patients with low-risk PCa by active surveillance can reduce therapy costs by 35% [17]. Thus,

novel approaches to discriminate aggressive from indolent forms of the disease are urgently required.

1.1.1 | Screening, diagnosis, and prognosis

In Europe, there is no standardized screening for PCa. Even the advised age for starting regular screening varies among countries. In Germany, the cancer information service provided by the German Cancer Research Center (DKFZ) recommends starting regular screening for men without genetic predisposition from 45 years [18]. This normally includes digital rectal examination (DRE) and/or blood-based PSA screening. Both are controversial because scientists and physicians disagree on their efficacy and sensitivity [14,15,19,20] and since they require further testing after initial PCa diagnosis. However, the complexity of pathogenesis makes it difficult to determine a patient's disease outcome. Hence, overtreatment is a major issue for many PCa patients.

Since the rise of biomarker discovery, multiple diagnostic and prognostic tests have been developed based on the patients blood, urine or tumor tissue. Even though some have been approved by the United States Food and Drug Administration (FDA) or Clinical Laboratory Improvement Amendments (CLIA), none has yet predicted PCa progression clearly enough to replace the current standard procedure and facilitate appropriate treatment choices [21]. A prognostic test to reliably differentiate aggressive from benign PCa still has to be found. Some biomarker-based tests available in the field are briefly described below. The emerging number of tests and their differences in timing and purpose of application complicate choosing the appropriate test. For screening, the PSA, PHI, and 4Kscore test have established themselves as indicators for a potential PCa onset for which more information is needed, such as via biopsy. Other diagnostic tests, such as the ExoDx Prostate IntelliScore (EPI), assays of *TMPRSS2-ERG* (T2E), or *PCA3*, attempt reliable recommendations regarding re-biopsy, prostatectomy, or active surveillance. Prognostic tests, such as Decipher, Prolaris, and Oncotype Dx, enable PCa risk stratification for feasible therapy approaches after PCa diagnosis as well as treatment decisions after surgery and after care [21].

The variety of tests for PCa outcome prediction makes a decision for physicians burdensome. Alam *et al.* compared three genomic tests (Decipher, Prolaris, Oncotype Dx), which are described in this thesis below, pairwise to each other and found that

they differed notably in their prognostic outcome [22]. This makes it even harder for physicians to choose the right test for their respective patient. Hence, knowing whether certain PCa tests are more suitable for specific patient subgroups is important.

Prostate Specific Antigen (PSA)

Potential PCa onset can be diagnosed by measuring increased (> 4 ng/ml) or continuous rising PSA levels from blood serum [11, 18]. PSAs are androgen regulated serine proteases encoded by Kallikrein 3 (KLK3) and produced in the prostate gland [11]. PSAs exist in the serum in different forms. While most are bound to protease inhibitors and refered to as 'total PSA', the molecular, unbound form of PSA is called 'free PSA' [23]. Some alternative blood tests, such as the Prostate Health Index (PHI) or the 4Kscore test, are among others mainly based on total and free PSA, but these methods prevailed as screening tests for PCa rather than predictors for tumor aggressiveness [21].

For years, the PSA test was the primary method for PCa screening [23], but it has received more and more criticism due to its exorbitant sensitivity and lack of specificity, which leads to overdiagnosis [11, 14, 24].

Decipher

The Decipher Prostate Cancer Test developed by GenomeDx Biosciences in cooperation with the Mayo Clinic is a genomic test relying on RNA expression extracted from primary PCa [25, 26]. The predictive test uses a genomic classifier modeled by the gene expression of 22 markers with a random forest machine learning algorithm [25]. The Decipher test is exerted after surgery to predict biochemical recurrence in men with high-risk pathology [25, 27]. Its predictive score can assist in therapy to find an appropriate treatment for the individual patient, which can relate to active surveillance, radiotherapy, or even adjusting the timing of a treatment. Multiple studies [26, 28–30] have validated the Decipher test as good predictor of aggressive tumors in high-risk PCa patients.

Prolaris

The Prolaris test designed by Myriad Genetic Laboratiories uses expression profiles of 31 genes involved in cell cycle progression (CCP) [31]. The test measures gene expression levels using RT-PCR with tumor tissue extracted via radical prostatectomy, but it also works with tissue taken by needle biopsy [31, 32]. Normalized against 15 house keeping genes, these genes were tested for association against recurrence via Cox proportional hazards regression and partial likelihood ratio tests and configure the predefined CCP score [33]. The Prolaris test enables risk prediction of recurrence after radical prostatectomy for men with both low- and high-risk PCa [31, 32].

Oncotype Dx

The Oncotype Dx Prostate Cancer Assay established by Genomic Health, is a multi-gene RT-PCR expression assay for prostate tissue specifically taken from needle biopsy specimens. The test covers 12 cancer-related genes from four different biological pathways and five reference genes. Using algorithms, these genes are compound to generate the Genomic Prostate Score (GPS) [34]. The Oncotype Dx test has been validated in clinical setting to relative reliably predict cancer aggressiveness [35, 36].

ExoDx Prostate IntelliScore (EPI)

A novel urine based test, the ExoDx Prostate IntelliScore (EPI), was recently developed by McKiernan *et al.* [37, 38]. Using an exosome gene expression assay, mRNA of three genes (*ERG*, *PCA3*, *SPDEF*) are measured to predict high-grade PCa. The noninvasive EPI test applies to men with elevated PSA levels at initial biopsy [37, 38]. An initial study [39] delivered promising results. Approved in 2019, the EPI test is only available in the United States, so far [40]. It remains to be seen if the EPI test prevails and establishes itself in daily clinical routines.

1.1.2 | Tumor grading and staging

Although prognostic biomarker tests became popular, PCa is still evaluated by histopathological examination of tissue specimen. Based on needle biopsy or prostatectomy specimen, the current stage of the tumor can be determined [41, 42]. While a needle biopsy

is highly selective, the large scale post-operative tumor staging is much more reliable. Tumor encapsulation and spreading to other parts of the patients body also play a role in the staging process. For PCa, two common tumor staging systems have been established: the Gleason grading system and the TNM-staging system, which are both concomitantly used to describe the current tumor stage and as a rough prognostic indicator for disease progression [41–43]. While the Gleason grading system is PCa specific, the TNM-staging applies to malignant solid tumors in general and is adapted to PCa accordingly [41, 42, 44]. All biomarker-based tests were developed using at least one of these systems to distinguish between aggressive and benign tumors [25, 32, 34, 37].

Gleason grading system

PCa is graded by the Gleason score, which describes the differentiation of cancer cells [41]. The Gleason grading system was developed by Donald Gleason in 1966 [45], but is still in use today. The Gleason score is the sum of the two most common histologic pattern scores, which range from 1 to 5 each, in multiple PCa core biopsies. A higher score represents a poorly differentiated PCa, often referred to as high-grade PCa. Nowadays, the Gleason score differs from the original system [46], starting with a lowest grade of 6 [47], which is prone to misdiagnosis of low grade PCa [46]. Another problem is the discrimination between Gleason score 7 deriving from either Gleason pattern 3+4 or 4+3 [46], which show significant differences in their biologic behavior, prognosis, and pathological specification [48, 49]. Therefore, the Gleason grading group system based on the Gleason score was recently introduced [50]. It reevaluates and regroups the classical Gleason score into five new groups (Gleason grade group (GGG) I-V, Table 1.1). The intention was to improve grade stratification, reduce the grading groups to 5 instead of multiple Gleason pattern combinations, and reduce misdiagnosis in low-grade PCa to accurately depict prostate cancer biology [46].

TNM-staging system

The TNM-staging system introduced by The American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) is composed of three categories. Thereupon, PCa can be classified based on primary tumor (T), regional lymph node infiltration (N), and distant metastases (M) [51]. The T category derives

Table 1.1: Stages of cancer differentiation as defined by Gleason score, Gleason grading group, and corresponding histologic pattern [45, 46].

Grade Group	Gleason Score	Histologic Gleason Pattern
I	6	3+3
II	7	3+4
III	7	4+3
IV	8	3+5, 5+3, 4+4
V	9, 10	4+5, 5+4, 5+5

from pathological staging determined after the surgical removal of the prostate. It describes the size and extent of the primary tumor. Without evidence of primary tumor, it is denoted as T0. T1 and T2 classify a PCa, which is still located inside the prostate gland. In contrast, T3 and T4 stand for large PCa not encapsulated of the gland anymore but already spread to nearby tissue [42, 44, 52]. The N category specifies whether the lymph nodes are already affected (N1) or not (N0). The M category refers to the existence of metastasis. If the cancer has spread it is documented as M1, otherwise it is M0 [52].

1.2 | Prostate cancer on the molecular level

1.2.1 | Subtypes

Localized PCa is a distinctly heterogeneous disease, but early genomic driver events in oncogenisis enable classification of PCa into multiple molecular subtypes. These subtypes are defined by somatic genomic aberrations and alterations, such as chromosomal rearrangements or gene mutations, in their tumoral DNA [53, 54]. The largest group of molecular PCa subtypes manifests in gene fusions of ETS family transcription factors including *ERG*, *ETV1*, *ETV4*, and *FLI1*, with *TMPRSS2* [53, 55]. ETS-negative PCa subtypes exclusively posses recurrent mutations in *SPOP*, *FOXA1*, or *IDH1* among others, each representing an unique and discrete subclass. Alterations in tumor sup-

pressor *TP53* and *PTEN* are more frequent in tumors harboring an ETS fusion [53, 56]. In contrast to the frequency of the subtype defined by the *ERG* fusion, the remaining subtypes are rather rare among PCa patients [53, 57]. Regardless, the variety of distinct molecular PCa subtypes and the affiliated differences in etiopathology make treatment decisions challenging.

1.2.2 | *TMPRSS2-ERG* fusion

The *TMPRSS2-ERG* (T2E) fusion is the most common distinct molecular subtype in PCa [53, 58]. It appears in around 50% of PCa and corresponds to 90% of PCa fusions from the ETS family of transcription factors [56, 59]. The fusion gene is characterized by either a chromosomal rearrangement or deletion on chromosome 21, amalgamating *TMPRSS2* with *ERG*. Its oncogenic properties are associated with aggressive tumor progression and worse outcome [60–62]. Despite its high specificity, the low sensitivity of T2E as prognostic biomarker makes its combination with additional biomarkers inevitable [63, 64]. By differentiating between T2E-positive and T2E-negative PCa to constitute molecularly distinct PCa subtypes, prior studies exploit various pathways or gene-signatures promoting PCa malignancy [53, 58]. By distinguishing between T2E-positive and -negative tumors, deviations in DNA methylation profiles have been reported, which prove that these molecular PCa subtypes can be described by diverse oncogenic pathways [65].

1.3 | Prostate cancer on genomic level

1.3.1 | High throughput sequencing

The sequencing of the human genome in the early 2000s, began to revolutionize medical and biological research by empowering the analysis of the whole genome [66]. Since then, high throughput methods have facilitated comprehensive large-scale next-generation sequencing projects, such as the 1000 Genomes Project [67] and the International Hap Map Project [68] exploring the human genome, which cover millions of single nucleotide variants (SNVs) and single nucleotide polymorphisms (SNPs) of thousands of individuals.

Next-generation sequencing, such as methods provided by Illumina, has prevailed for whole genome (WGS) and whole exon sequencing (WXS) in research. Its continuously declining costs allow the generation of large genomic data for exploring genetic variation of diseases, opening new possibilities in cancer research [69]. Moreover, third generation technology has been developed recently and is becoming more popular. Contrary to the preceding sequencing generation, technologies such as PacBio instruments and Oxford Nanopore, have brought sequencing to a new level by enabling the sequencing of long reads, greater sequencing speed, or even decentralized sequencing [69]. High throughput methods, such as DNA-seq or RNA-seq, transform genetic or transcriptomic material into big data by assembling sequence reads across the genome or transcriptome. More selective alternatives are costum- designed microarrays, which obtain specific known variants or gene expressions and have become a popular approach in population studies [69, 70].

1.3.2 | Population studies and genetic variation

The accelerating technical progress in high throughput sequencing techniques and microarrays facilitated the assembly of population studies. In cancer research, population studies conducted as genome-wide association studies (GWAS) have become popular to extensively test for common genetic variants associated with disease traits. Since its beginning, the number of detected variants affecting PCa has increased in the past ten years. Benafif *et al.* reported 170 common variants from more than 40 GWAS, and the numbers are still on the rise [71].

GWAS compare the allele frequencies of SNPs in different sample groups. While common SNPs have a minor allele frequency (MAF) of more than 5%, rare SNPs are only possessed by under 1% of the population. Thus, PCa cases are compared to a healthy control group to identify genetic variants or susceptibility genes associated with increasing PCa risk or genetic predisposition for PCa onset [70]. In contrast to a case-control GWAS, a case-only GWAS is conducted on patients only and compares patient subgroups with different disease traits. This approach identifies genetic variation associated with disease progression and outcome of tumor subtypes with divergent etiopathology [72]. To avoid false positives or biases in GWAS, it is important to correct for population stratification.

In GWAS, identified SNPs significantly associated with disease traits, may not necessarily be the causal SNP but rather its tag SNPs. Tag SNPs are representatives of a haploblock, a genomic region with minor genetic recombination [73]. A causal or functional SNP is in high linkage equilibrium (LD) with its tag SNPs, which describes the non-random association of variants [70, 73]. SNPs in LD with each other are closely linked variants that are non-independently inherited together on the same haploblock [73]. With genome-wide significant tag SNPs, potential risk loci can be easily identified.

These identified risk loci can be explored with fine-mapping to find adjacent causal SNPs and elucidate their function. Fine-mapping is a complex process that prioritizes potential causal SNPs based on their LD and pairwise correlation for further studies. Moreover, potential candidates are scrutinized regarding their risk effect, genomic annotation, and function influencing the disease trait. Causal SNPs are often associated with altered gene expression levels [74]. But, mRNA may not necessarily be directly affected by a genetic variation in its coding region. In fact, most trait-associated SNPs identified in GWAS were located in intronic or intergenic regions. Therefore, examining the epigenetic background of causal SNPs specifically in non-coding regions can shed light on their influence on regulatory mechanisms [74].

From these scattered tag SNPs, the remaining unknown variants on a haploblock can be accurately predicted with imputation methods based on the genome of a reference population. Imputation is also an important aspect in meta-analysis, which accumulates the results of multiple GWAS. A single GWAS may indeed find significant results but is not robust enough to evaluate whether significant variants were found coincidentally or resulted from biased study design. A sufficient number of studies combined in a meta-analysis increases both validity and power of the results. Due to different genotyping and sequencing technologies, however, merging studies can lead to quality issues. Imputation can solve this by harmonizing the data from different platforms [70].

1.3.3 | Epidemiology

Multiple risk factors influence PCa progression and tumorigenesis. PCa is known as the tumor of older men, but it is also driven by environmental factors and life style choices. Genetic factors, such as predisposition due to family history and ethnicity affect the probability of PCa or disease outcome [75]. Allele frequencies of SNPs vary among

populations [76, 77]. While some risk loci (8q24) are associated with PCa risk independently of ethnicity, other risk loci and susceptibility alleles could be verified only in Europeans so far or manifest deviating frequencies among ethnicities [78–82]. Particularly among Africans and African Americans, PCa and resultant mortality are much more frequent than among populations of European origin [83, 84]. Despite higher incidences of aggressive tumors, only 13% of men with African descent have tumors with T2E fusions [85], which emphasizes the PCa differences between ethnicities. Likewise, in East Asian populations, the amount of T2E-positive PCa is only a fraction of that among Europeans and in contrast to them, is not associated with any clinicopathological phenotype for East Asian patients [86, 87]. Powell *et al.* even implied that biomarkers, such as the T2E fusion, are not eligible for patients of non-European descent [88]. This variability in incidences among ethnicities is referable to genetic diversity and population structure but not health care reasons [76].

1.4 | The Aim of this thesis project was to assess the contribution of germline variation and somatic mutations to PCa prognosis and prognostication

PCa is a complex, heterogeneous disease whose tumor progression is not yet fully understood. Distinguishing between benign and aggressive PCa is still challenging. Detecting further transcriptomic and genomic biomarkers to improve clinical tests and understand the underlying mechanism driving PCa is inevitable.

The aim of this thesis was to assess the contribution of germline variation and somatic mutations to PCa prognosis and prognostication, which was explored by two approaches based on transcriptomic and genomic data. To this end, gene expression profiles of PCa from different subtypes were screened to highlight their subtype specific differences regarding their pathways and involved gene signatures. Furthermore, the survival of theses subgroups was examined to evaluate potential subtype specific biomarkers. The identified biomarkers were evaluated both experimentally and computationally. Several case-only GWAS were exploited in a meta-analysis to find significant SNPs that may conduce the development of high-grade PCa. Afterward, these SNPs were fine-mapped to identify potential causal SNPs, whose potential function and

regulatory effect on PCa progression was inferred.

2

Data and methods

Detailed information on the software and datasets mentioned in the following Sections is shown in Tables 2.5, 2.6, 2.7, and 2.8.

2.1 | Cohorts

2.1.1 | PCGA^{LMU}

The Prostate Cancer Genome Atlas of the Ludwig-Maximilians-University of Munich (PCGA^{LMU}) is a German collaboration between the Max-Eder-Research Group of the Institute of Pathology of the LMU, the Urologic Clinic and Polyclinic of the University of Munich, the Institute of Genetic Epidemiology at the Helmholtz Center Munich, and the Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) of the University Medical Center of the Johannes Gutenberg University Mainz.

Blood samples of 800 PCa patients before surgery were collected at the Urologic Clinic of Munich between 2017 and 2019. Data management and pseudonymization was managed by the Helmholtz Centre under the administration of Prof. Dr. Konstatin Strauch. The overall study setting and data analysis was carried out by the Max-Eder-Research Group of the Pathological Institute of the LMU. The PCGA^{LMU} study was approved by the LMU Ethics committee (Project Nr: 810-16). Prof. Dr. med Alexander Buchner was responsible for the recruitment and elucidation of study patients. All

study participants signed a patient's agreement for the use of their genetic and bio material as well as their clinical records.

2.1.2 | TCGA-PRAD

PCa samples from The Cancer Genome Atlas (TCGA) were incorporated for meta-analysis. The TCGA-PRAD cohort comprises 440 samples of prostate adenocarcinoma with comprehensive genetic, transcriptomic, clinical, and survival data. For TCGA-PRAD, Abeshouse *et al.* also provided information on the presences of the most common gene fusions and mutations affecting PCa [53].

Both projects of this thesis, were based on the TCGA-PRAD cohort. However, due to their time frame and different project target settings, the final samples used to differ between both projects.

Small SNP assembly for detection of ethnical background

A small assembly of SNPs processed from DNA-seq data on level 2 were directly downloaded from TCGA Data Portal (controlled access), which was closed in June/July 2016 after the NCI launched the Genomic Data Commons (GDC) [89]. To the authors knowledge, those files were not included in any of the GDC Data Portal data releases, but can be found in the NCI GDC Legacy Archive (Table 2.8). These data were only used for the roughly determination of the patients ethnical background to filter European samples used for transcriptome analysis.

Broad set of exonic variants for GWAS

Genomic data and clinical records from 440 PCa samples from the TCGA-PRAD cohort were downloaded with gdc-client [90] from GDC-portal [91] of the National Cancer Institute, which hosts TCGA database (controlled access). The extensive dataset of germline variants was retrieved by whole exon sequencing (WXS). Raw reads, sequenced from normal blood, enabled to retrieve germline data. The downloaded data were already aligned against human reference genome assembly GRCh38.

The clinical patient's characteristics of both TCGA-PRAD cohorts used in both projects are shown in Table 2.1 and 2.2.

2.1.3 | ICGC-CA

The ICGC-CA cohort is part of the PCAWG study (PanCancer Analysis of Whole Genomes) [92] established by the International Cancer Genome Consortium (ICGC, [93]). Access to the controlled individualized data was granted via the Data Access Compliance Office (DACO). With the Score-client (Table 2.7) all the germline data, extracted from blood samples, could be downloaded as alignments from the ICGC data portal [94].

The ICGC-CA cohort was assembled in Canada and comprises WGS, transcriptomic, and clinical data of 116 prostate adenocarcinoma samples of patients that underwent radical prostatectonomy or image guided radiotherapy [95].

2.1.4 | 1000 Genomes Project

The 1000 Genomes Project is a comprehensive genomic dataset describing the genetic variation of humans worldwide [67]. In its final phase it comprises 2,504 individuals of 26 populations around the world. Their genomes were sequenced by combining low-coverage WGS, deep exome sequencing, and dense microarray genotyping.

Here, this dataset was used as reference set to genetically determine the ethnic background of patients. The focus was on three so-called super-populations regarding people of African (n = 658), East-Asian (n = 504), and European (n = 503) descent (Table 2.4). People with European ancestry originate from France (CEU, n = 96), Finnland (FIN, n = 99), United Kingdom (GBR, n = 91), Spain (IBS, n = 107), and Italy (TSI, n = 107) [67]. With these five populations, the samples of all three genomic cohorts could be located more precisely within Europe and classified as Central European, Scandinavian, or Mediterranean.

Genotype data of autosomal chromosomes were downloaded in VCF format phase 3 (Table 2.8).

2.1.5 | GSE46691

The PCa cohort published under the accession number GSE46691 was assembled from the Mayo Clinic Radical Prostatectomy Tumor Registry [25]. Patients were treated with radical prostatectomy between 1987 and 2001. From the total number of 639 tumors, RNA was extracted and profiled by gene expression microarray (Affymetrix Human Exon 1.0 ST Array) for 545 samples [25]. Patients' characteristic is shown in Table 2.2. Raw RNA data is available at Gene Expression Omnibus (GEO).

2.1.6 | GSE16560

A watchful waiting cohort of 1,256 men was conducted over 30 years in Sweden [96]. From those patients, 281 "extreme" cases were selected, which can be differentiated between those with lethal and those with indolent PCa (10-years survival without metastasis). Patients' characteristic is summarized in Table 2.2. Normalized RNA data is available at GEO under accession number GSE16560 [96].

2.1.7 | TMA-cohort

The Institute of Pathology of the University Hospital of Bonn (Germany) assembled a well-characterized prostatectomy cohort (TMA-cohort) of 135 PCa patients with known T2E fusion (Table 2.3). As the cohort data is not publicly available via an accession number, it was referred to this cohort as TMA-cohort, in this thesis. From this cohort an immunohistochemistry (IHC) based on tissue microarrays was conducted to serve as further validation cohort for *RRM2* and *TYMS* in T2E-positive and T2E-negative tumors.

Table 2.1: Patients' characteristic of Central European samples in the $PCGA^{LMU}$, TCGA-PRAD ans ICGC-CA cohorts.

	$PCGA^{LMU}$	TCGA-PRAD	ICGC-CA
Central European patients #	751	263	84
Age (years)	67	62	62
median (range)	(43-87)	(46-78)	(46-81)
PSA (ng/ml)	8.7	7.1	7.2
median (range)	(0.2-343)	(1.6-87)	(1.6-39.5)
Gleason grade group (Gleason score)			
I (3+3)	86	20	14
II (3+4)	275	77	43
III (4+3)	164	41	25
IV (4+4, 3+5,5+3)	92	48	1
V (> 4+5)	123	77	1
Pathological T			
pT1	14	-	9
pT2	423	95	44
pT3	308	159	31
pT4	4	9	-
Pathological N (pN0/pN1)	539/73	177/48	-
Clinical M (0/1)	728/23	237/2	-
R (X/0/1/2)	-	3/166/86/3	-
Event (no/yes)	-	185/59	55/29
Time until event (month)	-	24.8	72.1
median (range)	-	(1-119)	(2-146)
T2E fusion (-/+)	65/58	155/108	36/48

Table 2.2: Patients' characteristic of samples in the European TCGA-PRAD cohort, GSE46691 cohort and validation cohort GSE16560.

	TCGA-PRAD		**GSE46691**		**GSE16560**	
T2E	negative	positive	negative	positive	negative	positive
Patients #	190	109	242	242	226	46
Age (years)	63	61	-	-	74	74
median (range)	(44-78)	(46-75)	-	-	(51-91)	(60-91)
PSA (ng/ml)	0.1	0.1	-	-	-	-
median (range)	(0-37.36)	(0-19.8)	-	-	-	-
Gleason Grade						
Group (Gleason score)						
I (3+3)	8	4	26	26	75	2
II (3+4)	47	38	104	143	90	24
III (4+3)	41	17				
IV (4+4, 3+5, 5+3)	31	22	38	22	23	4
V (> 4+5)	63	28	72	50	38	16
Pathological T						
pT2	62	36	-	-	-	-
pT3	118	71	-	-	-	-
pT4	7	2	-	-	-	-
Pathological N						
(pN0/pN1)	154/36	89/20	147/95	152/90	-	-
R (0/1)	119/66	73/30	-	-	-	-
EFS (no/yes)	128/44	82/24	-	-	71/155	3/43
Time until Event						
(month)	24	26	-	-	110	66
median (range)	(1-114)	(2-140)	-	-	(7-259)	(6-170)

Table 2.3: Patients' characteristic of samples of validation cohort TMA-cohort stratified by its patients' T2E-status. (BCR = biochemical relapse)

	TMA-cohort	
T2E	negative	positive
Patients #	88	47
Age (years)	65	65
median (range)	(48-75)	(45-75)
PSA (ng/ml)	7.5	6.6
median (range)	(1.0-163)	(1.5-58.4)
Gleason Grade Group (Gleason Score)		
I (3+3)	37	23
II (3+4)	19	12
III (4+3)	7	4
IV (4+4, 3+5, 5+3)	17	4
V (> 4+5)	7	2
Pathological T		
pT2	49	26
pT3	37	20
pT4	2	1
Pathological N (pN0/pN1)	79/8	42/5
R (0/1)	51/37	23/23
BCR (no/yes)	64/24	38/9
Time to BCR (month)	62	60
median (range)	(1-134)	(10-136)

2.2 | General methods

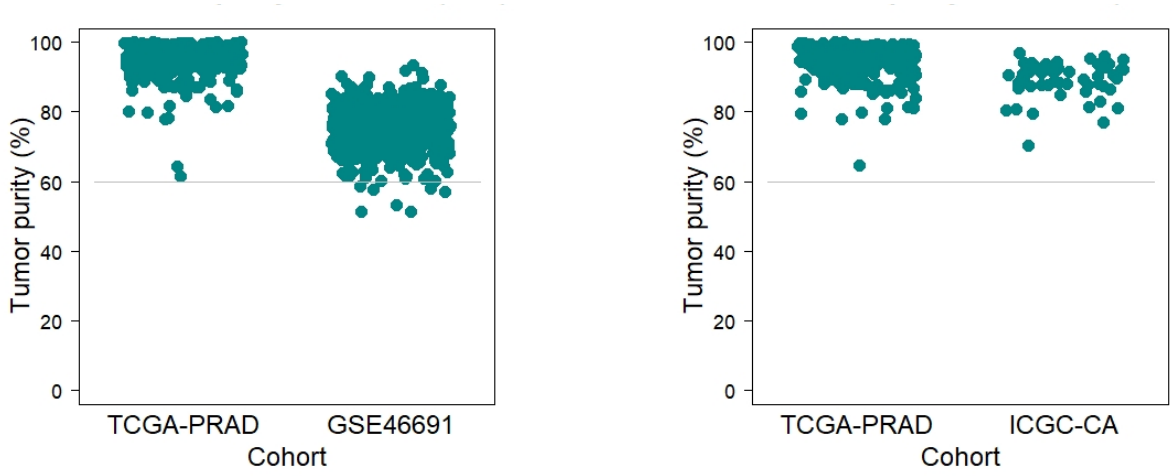
2.2.1 | Identification of ethnical origin

In GWAS, mostly homogeneous study populations are important to minimize false positive results. To keep the study population homogeneous, only people with similar descent should be included into a GWAS. Genetic clustering of people with known ancestry and those with yet-unknown reveals the descent of the latter.

To determine the genetic origin of patients, their genome was compared to a reference set of different populations. Germline variants of PCa cohorts were separately merged with the 1000 Genomes reference set comprising individuals of African, East Asian, and European descent. Genetic clustering of their genome was performed with principle component analysis (PCA) via PLINK [97, 98]. Their descent was identified using the R package mclust [99, 100] combined with precise manual inspection of the cluster visualization. The procedure was repeated with those patients identified as Europeans. They were compared to European subpopulations to enable an even closer location of their descent. Patients could be classified of Mediterranean, Scandinavian, or Central European descent. The numbers of all cohorts are described in detail in Section 2.6.2 and Table 2.4.

2.2.2 | Determination of tumor purity

Tumor purity was determined for all samples obtained with RNA-seq and gene expression microarrays (GSE46691 and GSE84042) with the ESTIMATE algorithm, which appraised the fraction of immune and stromal cells in tumor samples from gene expression pattern [99, 101]. Due to the low number of genes represented on the respective microarray, a tumor purity test on GSE16560 was not possible. All cases in the (Central) European TCGA-PRAD and ICGC-CA (GSE84042) cohorts had a higher consensus purity estimation (CPE) than 60%, which corresponds to the TCGA standard (http://cancergenome.nih.gov/cancersselected/biospeccriteria, Figure 2.1a and 2.1b), and were therefore kept for further analysis. In the GSE46691 cohort, for 7 of 545 samples, a CPE below 60% was observed and were therefore removed (Figure 2.1a).



(a) CPE of samples used for transcriptome analysis.

(b) CPE of Central European samples used for GWAS.

Figure 2.1: Tumor purity of samples: Consensus purity estimation (CPE) calculated for several cohorts.

2.2.3 | Determination of the *TMPRSS2-ERG* fusion status

Predictive approach

For the TCGA-PRAD cohort [53], Torres-García *et al.* inferred the T2E-status with PRADA based on RNA-seq split-reads [102].

Contrary, Fraser *et al.* investigated genomic rearrangements in the ICGC-CA cohort [95] using Delly [103] to discover breakpoints leading to the T2E fusion.

Estimation via gene expression level

In the Affymetrix dataset (GSE46691), the T2E-status was estimated from *ERG* expression levels, which show high concordance with the T2E-status [104]. As approximately 50% of PCa patients harbor the T2E fusion [59], all samples were classified into either T2E-positive or -negative corresponding to their individual *ERG* expression level laying above/below the median *ERG* expression. Those 10% of the microarray samples, whose *ERG* expression levels lay between the 45th and 55th percentile (Figure 2.2), were excluded to reduce the number of potentially misclassified samples.

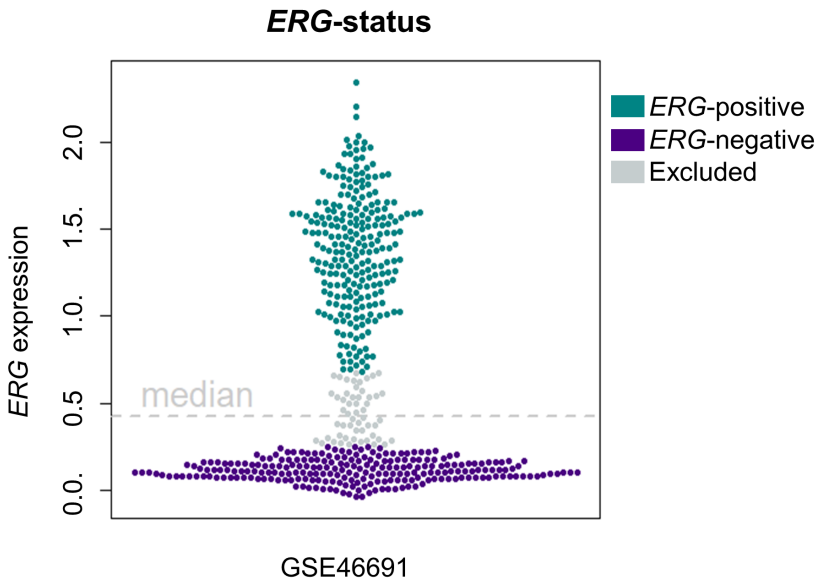


Figure 2.2: *ERG*-status of samples of GSE46691 cohort estimated from *ERG* expression levels [105]. Excluded samples (gray) lay between the 45th and 55th percentile.

Experimental detection of T2E fusion

In his Swedish cohort (GSE16560), Sboner *et al.* detected *ERG* rearrangements using fluorescence in situ hybridization (FISH) assay and quantitative polymerase chain reaction [96].

Immunohistochemical detection of *ERG*

Formalin-fixed paraffin-embedded (FFPE) prostate tumor samples were retrieved from the archive of the LMU Munich's Institute of Pathology matching 200 randomly selected samples of the PCGA^{LMU} cohort. As the gene expression level of ERG directly correlates with the presence of T2E fusion [104], an immunohistochemical (IHC) staining on ERG could reveal the presence of the T2E fusion. Therefore, 4µm sections were cut for every FFPE tumor sample. IHC staining was performed in collaboration with Jessica Kövi on Benchmark Ultra of Roche (Ventana) using the UltraViewDetection-Kit. Slides were pretreated with CC1 (pH 8.4) for 64 minutes, followed by an incubation with ERG (EP111) Rabbit Monoclonal Antibody (AC-0105, 1:80 dilution) for 28 minutes.

Afterwards, Dr. med. Fabienne Wehweck, an experienced resident pathologist on PCa, evaluated the IHC staining for ERG and classified each sample as either ERG-positive or -negative. Based on the correlation between *ERG* expression level and the T2E fusion [104], IHC staining on ERG could identify 93 T2E-positive and 107 T2E-negative PCa.

2.2.4 | Normalization of transcriptome data

Gene expression microarrays

Several publicly available gene expression datasets (GSE16569, GSE46691, and GSE84042) gained from microarrays were downloaded from GEO.

Transcriptome data of 272 PCa cases of the Swedish validation cohort, were profiled on Human 6k Transcriptionally Informative Gene Panel for DASL. With the novel DASL method (cDNA-mediated annealing, selection, ligation, and extension) Sboner *et al.* were able to determine mRNA of 6,100 genes from FFPE transurethral resection of prostate samples. Using the qspline algorithm, Sboner *et al.* normalized the data with the marginal mean of every gene as reference distribution. [96].

The dataset with accession number GSE46691 gathered 545 PCa cases profiled on Affymetrix GeneChip Human Exon 1.0 ST arrays [25]. Microarray intensities were normalized using the SCAN algorithm of SCAN.UPC [106] and the pd.huex.1.0.st.v2 annotation Bioconductor packages [99, 107] with brainarray chip description files (CDF, huex10sthsentrez, version 21), yielding one optimized probe-set per gene (gene level summarization) [108].

The GEO dataset GSE84042 comprised gene expression data from samples of the ICGC-CA cohort [95]. 53 of 73 samples could be assigned to the Central European samples. While 12 samples of those were profiled on the Affymetrix Human Gene 2.0 ST Array, the microarray signals of the remaining 41 samples were measured on Affymetrix Human Transcriptome Array 2.0. To account for this, the samples were normalized with the RMA algorithm and batch corrected using the sva package by Fraser *et al.* [95].

RNA-seq

Available transcriptome data of the TCGA-PRAD cohort consists of pre-processed RNA-seq level 2 data [53]. Therefore, Illumina HiSeq 2000 RNA Sequencing Version 2 analysis system [109] was used followed by the MapSplice algorithm for read mapping [110]. TCGA applies RSEM for transcript quantification [111]. Raw counts were normalized by division by the 75^{th} percentile of all raw counts and their multiplication by 1,000. For the transcriptome analysis project, gene expression data of 384 identified European samples were extracted from collectively 497 cases.

2.3 | Tissue microarray and immunohistochemistry

On the TMA-cohort (Section 2.1.7) an IHC-analysis based on 135 samples represented on tissue microarrays (TMAs) was conducted to validate two of the identified candidate genes, *RRM2* and *TYMS*.

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. Antigen retrieval for RRM2 and TYMS was achieved by ProTaqs IV Antigen-Enhancer (# 401602392, Quartett) and ProTaqs IX (# 401603692, Quartett). RRM2 was detected with a specific rabbit-anti-human RRM2 antibody (1:500, 60 min incubation time; HPA056994, Atlas Antibodies). 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).

For 133 of 135 patient specimens (98.5%) represented on the TMAs, evaluation of RRM2 immunoreactivity was possible. Likewise, 119 of 135 patient specimens (88.2%) could be evaluated for TYMS. RRM2 and TYMS immunoreactivities were quantified by an experienced data-blinded uropathologist (PD Dr. med. Yuri Tolkach, University of Bonn) as percentage of positive tumor cells (cytoplasmatic staining). The survMisc package for R [99] was used for optimal cut-off selection and Kaplan-Meier survival analyses. The following percentages of positive cells were selected as best cut-offs for marker positivity: $\geq 3\%$ for RRM2 and $\geq 5.5\%$ for TYMS.

These IHC analyses and their evaluation were already described and published by Gerke *et al.* [105].

2.4 | Transcriptome analysis pipeline

The integrative transcriptome analysis of T2E-positive and -negative PCa is mainly based on two cohorts, TCGA-PRAD and GSE46691, and supported by two additional validation cohorts (GSE16560 and the TMA-cohort). This analysis emphasized the transcriptomic differences between T2E-positive and T2E-negative PCa as by identifying potential subtype specific biomarkers.

This transcriptome analysis project and the following pipeline were already described and published in Gerke *et al.* [105].

2.4.1 | Sample stratification and selection

Based on the TNM-classification of tumors, datasets were stratified into cases with and without (lymph node) metastasis, which corresponds to N0M0 versus N>0 and/or M>0. As metastasis indicates aggressiveness in PCa and is associated with increased mortality [9], it was selected as prognostic factor in the transcriptome analysis.

In the transcriptome analysis pipeline based on both cohorts (TCGA-PRAD, GSE46691; Figure 2.3) only samples with information on their T2E fusion status (Section 2.2.3), sufficient tumor purity (Section 2.2.2), gene coverage of more than 90% and existing TNM-classification on metastasis were included.

Furthermore, it was possible to filter the TCGA-PRAD cohort for European descent, as incidence and aggressiveness is different in Africans and African Americans compared to Europeans [112]. This was done via principal component analysis in PLINK [97, 98] on a small pre-processed set of filtered germline variants (Section 2.2.1).

2.4.2 | Processing of transcriptomic data

For every gene, the variance across all samples was calculated with the genefilter Bioconductor package [99, 113] leading to the removal of 50% of those genes with the lowest variance in each cohort. Furthermore, transcripts with missing or ambiguous

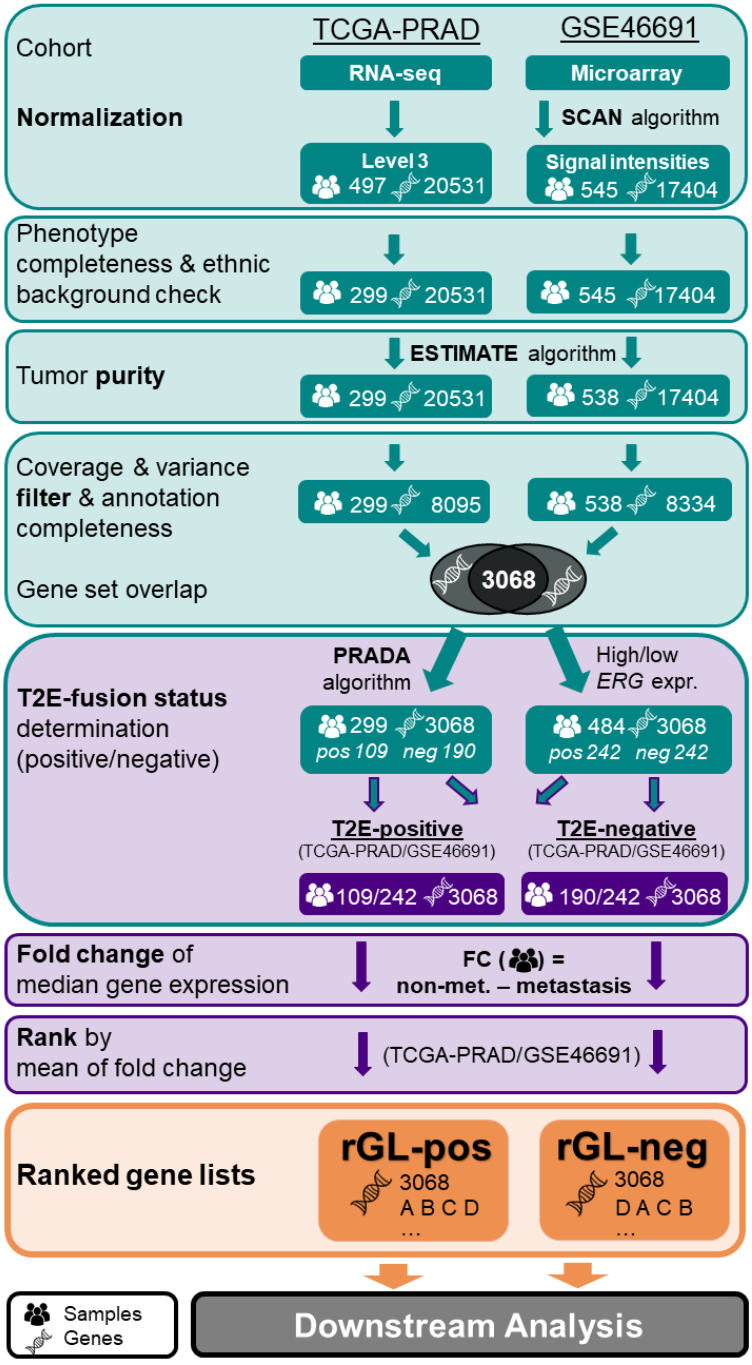


Figure 2.3: Processing pipeline of the transcriptome data from TCGA-PRAD and GSE46691 cohorts [105]. Sample selection and stratification (green) into T2E dependent subsets followed by generation of deferentially ranked gene lists (purple), rGL-pos and rGL-neg (orange).

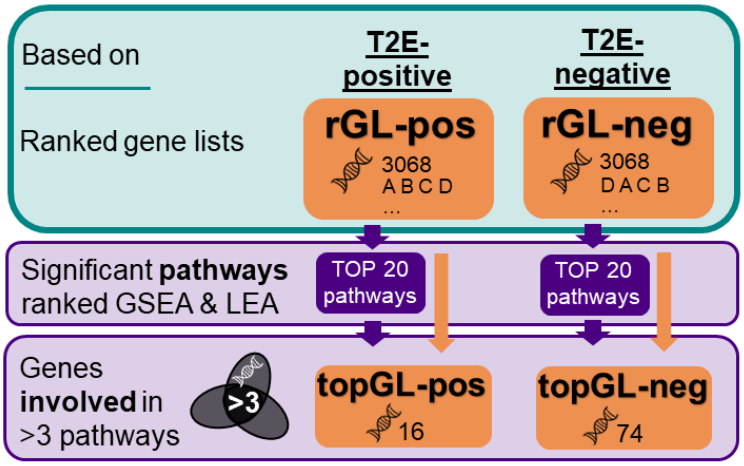


Figure 2.4: Analysis pipeline to determine the two final gene lists (topGL-pos and topGL-neg) for further downstream analysis [105].

annotation were removed. Genes represented in both cohorts assembled the final set of 3,068 variably expressed genes (Figure 2.3). After determining the T2E-status for each sample as described in Section 2.2.3, 10% of samples were removed from the GSE46691 cohort, yielding two cohorts of 299 samples (TCGA-PRAD; from 384 European samples) and 484 samples (GSE46691; from 545 samples) with 3,068 matching genes each. Next, both cohorts were split into four sub-cohorts regarding their samples' T2E-status comprising 109 T2E-positive samples and 190 T2E-negative samples (TCGA-PRAD) and 242 T2E-positive and -negative samples each (GSE46691). Afterwards, the median fold change between samples with and without metastasis at diagnosis was calculated separately for all four sub-cohorts. From this, two lists comprising the same 3,068 genes each were generated, which were ranked by their mean fold change in T2E-positive (rGL-pos) and T2E-negative (rGL-neg) cases.

2.4.3 | Downstream analysis

Based on these sample selection criteria and gene filtering (Figure 2.3), which resulted in two ranked gene lists, rGL-pos and rGL-neg, GSEA was performed followed by the selection of frequent genes (topGL-pos and topGL-neg) from its most significant gene-signatures (Figure 2.4) as described in Section 2.7.1. For every gene from both lists,

association with metastasis was tested in the two cohorts TCGA-PRAD and GSE46691. The testing procedure will be explained in more detail in Section 2.7.2. Afterwards, survival analysis as described in Section 2.7.4 was carried out for all genes of topGL-pos and topGL-neg on all European samples of TCGA-PRAD and the Swedish validation cohort (GSE16560). Only genes that were significant in association testing and survival analysis in both cohorts were considered as candidate genes.

For these significant candidate genes, GSEA was replicated based on T2E-negatives stratified by their gene expression (Section 2.7.1) for validation. Two genes (*RRM2* and *TYMS*) were additionally validated via IHC from TMAs in the TMA-cohort (Section 2.3).

2.5 | Genotyping of germline variants

2.5.1 | DNA extraction from blood samples

For the $PCGA^{LMU}$ cohort, DNA was extracted from blood with the NucleoSpin Tissue Kit (Machery-Nagel) as described in the manufacturer's protocol [114]. For genotyping, 15 μL DNA with a concentration of 60 ng/μL was used. This was executed and prepared for genotyping in collaboration with Stefanie Stein and Rebeca Alba Rubio.

2.5.2 | Genotyping

All samples assembled for the $PCGA^{LMU}$ were genotyped on the InfiniumTM Global Screening Array-24 including the multi-disease drop-in panel (GSA-MD chip) of Illumina [115]. Therefore, the prepared DNA (Section 2.5.1) was processed as described in the official Infinium HD Assay Ultra Protocol Guide provided by Illumina. Afterwards, signal intensities were measured with the Illumnia iScan System using the cluster files GSPMA24v1_0-A_4349HNR_Samples.egt and GSAMD24v2-0_20024620_A1-762Samples-LifeBrain.egt, which were provided by a consortium initiative. Evaluation was done with the Genotyping Analysis Module of the GenomeStudio Software 2.0 and revealed good data quality. This process and its evaluation was conducted by Nadine Lindemann and Dr. Jennifer Kriebel.

Genotyping was carried out in four batches of 200 samples, each. While the first

batch was genotyped on the GSA-MD chip v1.0, the subsequent batches run on GSA-MD chip v2.0. Samples genotyped on the same chip version were called together. Due to withdrawal of their patient's agreement or a later recognized diagnostic error, 12 samples were removed from the study leaving 788 PCa samples for the $PCGA^{LMU}$.

2.5.3 | Batch Effect Correction for different chip versions

As one of four batches from the $PCGA^{LMU}$ cohort ran on an earlier version of the GSA-MD chip, the samples had to be checked for batch effects in genotyping that might appear from potential differences between the two versions. The genotypes of samples from both chip versions were compared via PCA (Figure A.1). An association test between samples of both versions resulted in an inflation factor $\lambda = 1.01$. The corresponding *P*-values from all variants with MAF > 0.05 included a very small number of significant hits $P < 1.0 \times 10^{-4}$ and yielded a median *P*-value of 0.5 as suggested by Turner *et al.* [116] for quality control in GWAS. These performed tests revealed no batch effects between GSA-MD chip version 1 (n = 199) and 2 (n = 589). Therefore, the samples of all four batches from the $PCGA^{LMU}$ could be merged into one dataset and analyzed together.

2.6 | Genomic data processing pipeline

2.6.1 | Calling germline variants

The variant calling process described below is following the GATK Best Practices workflow (version 4) used at the Broad Institute [117]. While the raw alignments from WGS (ICGC-CA) run the whole pipeline, the WXS (TCGA-PRAD) could skip the cleaning step, as they were available for download already cleaned. The processing pipeline is shown in Figure 2.5. With vcftools [118] and bcftools [119], data was modified and manipulated to enable a better and faster processing of the used software in the processing pipelines.

In all following steps the reference genome and the SNP database for WGS were based on genome assembly hg19 (GRCh37), but for WXS they were based on the GRCh38 assembly.

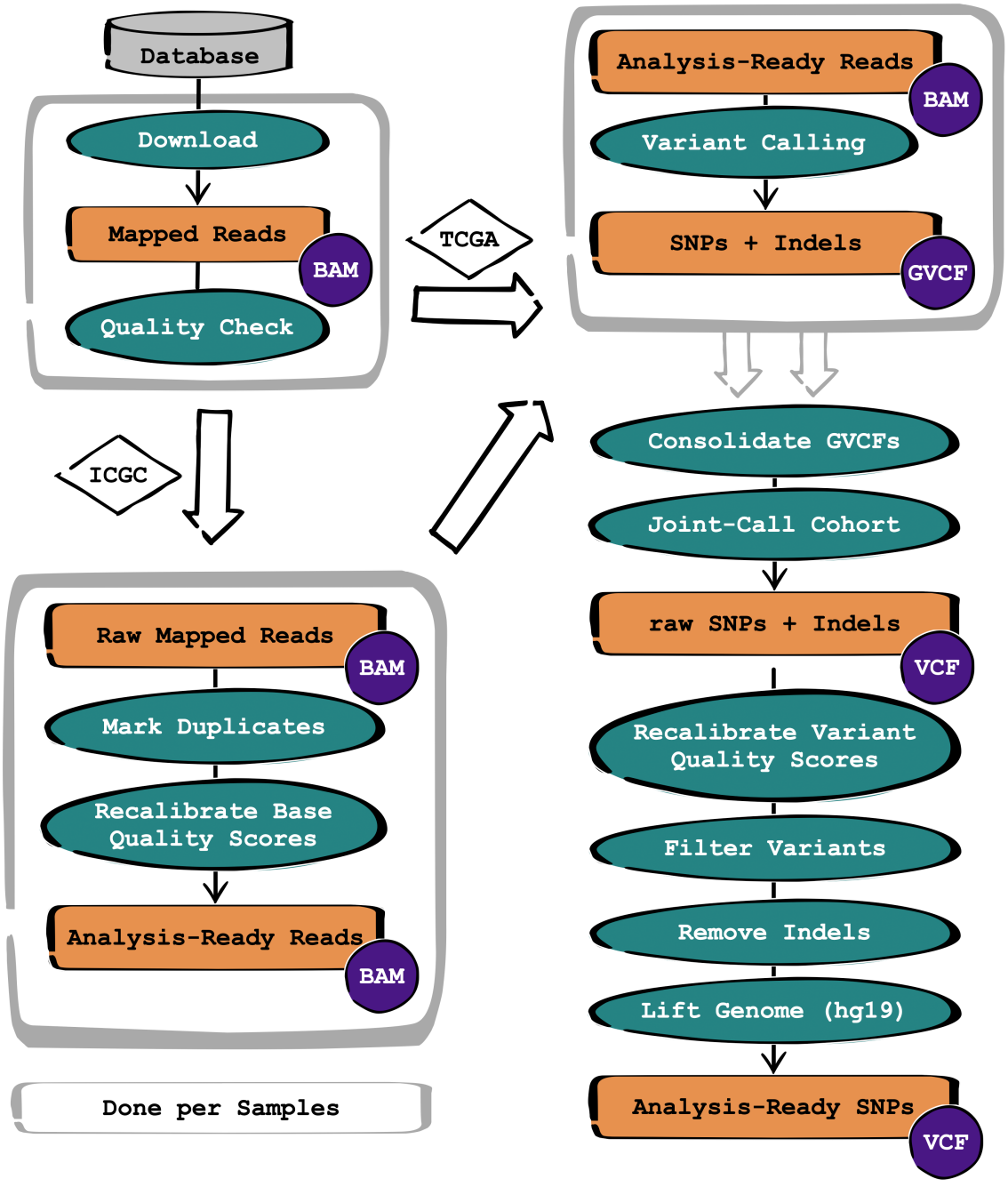


Figure 2.5: Pipeline of calling germline variants from alignments derived from normal blood and its (pre/post) processing as done for the ICGC-CA and TCGA-PRAD cohort. While the ICGC-CA cohort runs through the whole pipeline, the TCGA-PRAD cohort starts with the variant calling (right side) after its QC. Variant calling and its preprocessing is run per sample (gray box), followed by chromosome-wise post-processing of all samples combined.

Quality Check

First, all downloaded BAM files were checked for their quality with FastQC [120] followed by MultiQC [121]. Alignments failing the base quality check were removed. In the ICGC-CA cohort 3% out of 116 samples and 11% of 440 TCGA-PRAD samples did not pass the quality check or were damaged. Therefore, only 113 samples were kept from the ICGC-CA cohort and 393 from the TCGA-PRAD cohort.

Cleaning

Cleaning of alignments derived from sequencing of normal blood samples was done separately for each sample. First, duplicate reads were removed from the alignments using the MarkDuplicate method of Picard tools [122]. After indexing with samtools [119], systematic errors in base quality scores were detected with GATK BaseRecalibrator [123] using the ICGC reference genome (genome.fixed.fa; GRCh37, Table 2.8) and SNP database version 138 (hg19, Table 2.8). Afterwards, the base quality score recalibration was executed with GATK ApplyBQSR [123].

Calling Germline Variants

For every sample, germline SNPs and indels were called with GATK HaplotypeCaller [123] via local re-assembly of haplotypes with the reference confidence mode (ERC) in GVCF format. Minimum base quality score was set to 30. As the focus was on SNPs only, the maximum number of alternate alleles was reduced to 2, saving processing time. The reference genome and SNP database used were on assembly hg19 for WGS, but GRCh38 for WXS.

For WGS and WXS separately, all samples were merged with GATK GenomicsDBImport [123] and split chromosome wise. Afterwards, joint genotyping was performed with GATK GenotypeGVCFs [123] per chromosome.

Quality Filtering

With GATK VariantRecalibrator [123] a recalibration model is build for variant quality scoring, which is needed for filtering. It is used with the same corresponding ref-

erence genome as described before. It runs in SNP mode only using QD, FS, SOR MQ MQRankSum, and ReadPosRankSum parameter for calculations. Furthermore, resources regarding the hapmap, omni, 1000G, and dbsnp databases are used for training and truth sets. The exact resources for both cohorts based on GRCh38 and hg19/GRCh37 are listed in Table 2.8. Based on the resulting recalibration table, the quality of every variant was calculated with GATK ApplyVQSR [123] in SNP mode for truth sensitivity level of 99.0. These two steps for variant recalibration were only run in SNP mode but not for indels, which were culled in the next step.

With GATK SelectVariants [123] in SNP mode, those variants that did not pass the previous quality checks or were not biallelic were removed.

Unification of the genome assembly

Due to a different genome assembly of TCGA-PRAD cohort compared to the $PCGA^{LMU}$ cohort and ICGC-CA cohort, the genome of each chromosome was lifted from GRCh38 to hg19 using Picard LiftoverVCF [122] with hg38ToHg19.over.chain and the human genome reference from UCSC (Table 2.8).

Standardization

All variants were annotated with SnpSift [124] for their unique rs id using a SNP annotation file (All_20180423.vcf.gz, Table 2.8). Finally, the dataset of each cohort was converted from VCF into binary PLINK format, consisting of BED, MAP, and FAM file.

2.6.2 | Determination of ethnical origin

Clustering individuals by their genome against a reference set of multiple populations, as described in Section 2.2.1, enables to identify their genetic descent. To receive a comparative homogenous study cohort, only PCa patients with Central European descent were considered in this thesis. Therefore, the cohorts $PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA underwent this procedure to define which samples to use for GWAS analysis.

PCA on $PCGA^{LMU}$ revealed four samples with African, one with East Asian, and 781 with European descent. When focusing on European subpopulations, two of Scandinavian, 28 of Mediterranean, and 751 of Central European origin could be identi-

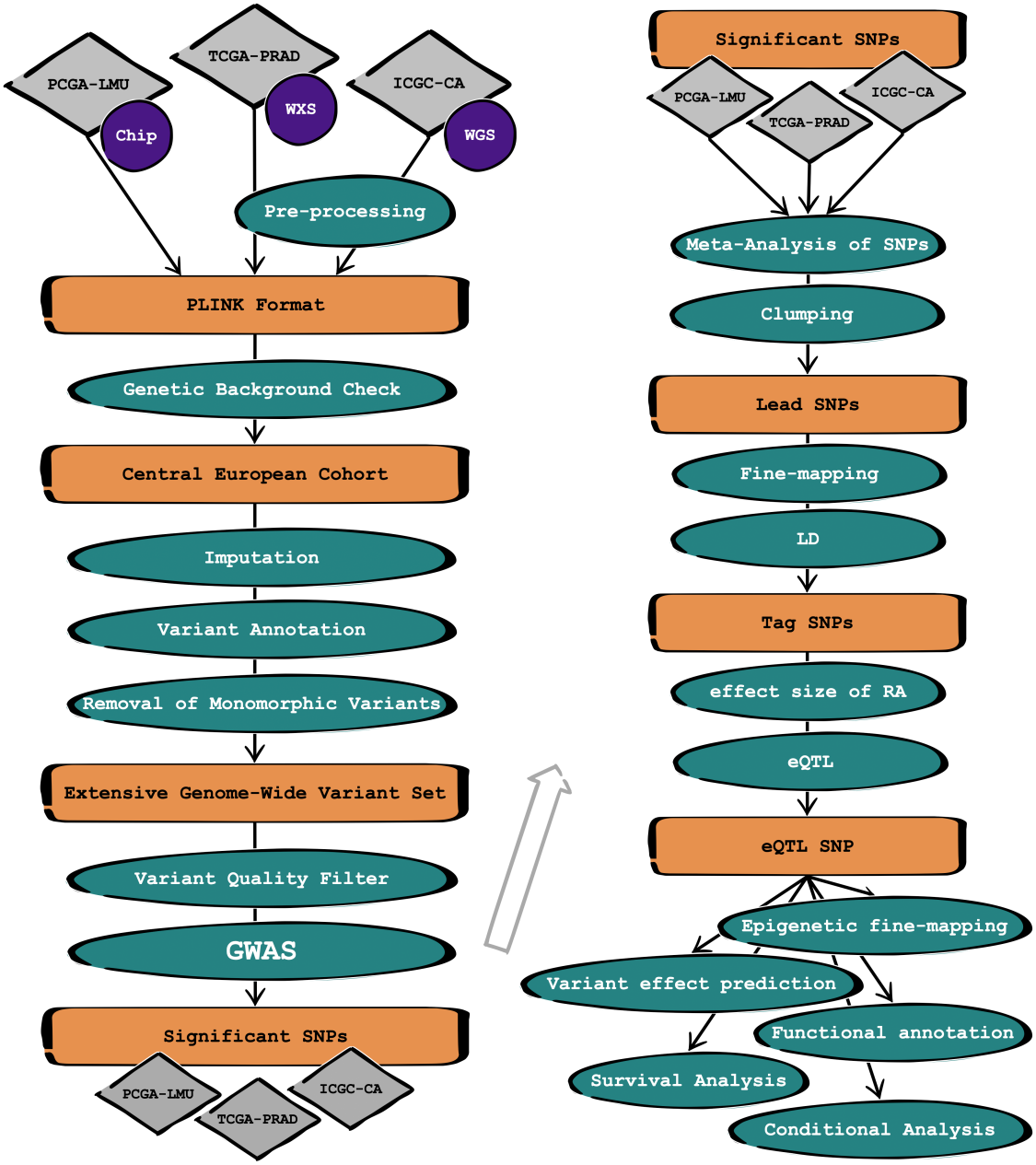


Figure 2.6: Workflow of GWAS on germline variants, lead SNP identification of its meta-analysis and the subsequent statistical analysis. Processing of the sequencing data (first green box) is described in detail in Figure 2.5. (LD = linkage disequilibrium; RA = risk allele)

Table 2.4: Ancestral population distribution in the cohorts. Final number of samples (Central European ancestry) used for further analysis are highlighted in bold font.

Population	$PCGA^{LMU}$	TCGA-PRAD	ICGC-CA	1000 Genomes
Total	788	393	113	1665
African	4	45	4	658
East Asian	1	9	3	504
Mixed ancestry	2	26	12	-
European	781	313	94	503
Central European	**751**	**263**	**84**	187
Mediterranean	28	49	10	210
Scandinavian	2	1	-	99

fied. TCGA-PRAD cohort is composed of 45 African, nine East Asian, and 313 European patients, whereas the latter can be grouped into 49 patients originating from the Mediterranean area, one from the Scandinavian area, and 263 from Central Europe. The ICGC-CA cohort contained four samples of African, three of East Asian, and 94 of European descent. Its European subpopulations split into 10 Mediteranean and 84 Central European samples. For individuals with mixed ancestry, population assignment was ambiguous (26 for TCGA-PRAD, 12 for ICGC-CA, and 2 for $PCGA^{LMU}$) and were hence excluded.

In the end, 751 Central European samples for $PCGA^{LMU}$, 263 for TCGA-PRAD, and 84 for ICGC-CA could be identified that were suitable for GWAS. All other samples were discarded. All cohorts and their corresponding populations are summarized in Table 2.4. The population clustering of the $PCGA^{LMU}$ cohort is shown in Figures A.2a, A.2b, and A.3, while those of TCGA-PRAD and ICGC-CA can be found in the Figures A.2c, A.2d, and A.4.

2.6.3 | Imputation

Genotype imputation enables the prediction of genetic variants that were not covered by previous sequencing or genotyping. Based on an individuals germline variants, missing variants can be inferred using haploblocks from a reference panel with known genome.

Before converting the variants of all autosomes into VCF, the variant file was checked with a perl script provided by the McCarthy Group (Table 2.7), which compares variants against the corresponding reference set and corrects for potential strand flips, alleles, erroneous positions, and ref/alt assignments to prepare the dataset for imputation. The imputation was executed on the Michigan Imputation Server (Table 2.7), which uses the genotype imputation algorithm Minimac3 [125] against the predominantly European population of the Haplotype Reference Consortium reference panel (HRC r1.1 2016) [126]. Phasing was performed with Eagle2 [127].

Imputed variants were filtered for good quality with GATK VariantFiltration and SelectVariants [123]. Thereby, monomorphic variants and those with a predicted $R^2 < 0.3$ were removed as proposed by Hancock *et al* [128]. Multi allelic variants were identified with bcftools norm [119] to be excluded later, as some subsequent analysis tools are not able to handle them. Remaining variants were annotated with bcftools annotate [119] followed by SnpSift [124] using the SNP database All_20180423.vcf.gz (Table 2.8). Variants located on allosomes and the mitochondrial DNA that could not be imputed on the Michigan Imputation Server were added to the variant set. Afterwards, all variants were converted into binary PLINK format for GWAS and downstream analysis.

2.7 | Statistical Analysis

2.7.1 | Gene set enrichment analysis and leading edge analysis

Two pre-ranked lists (rGL-pos and rGL-neg) of 3,068 genes that were created as described in Section 2.4 based on gene expression differences between PCa patients with and without metastasis in T2E-positive and -negative samples, were used for gene set enrichement analysis (GSEA) to identify significantly enriched gene-signatures. GSEA was performed with 1,000 permutations on gene sets from the Molecular Signatures

Database (MSigDB v6.2) that represents expression signatures of genetic and chemical perturbations (CGP) [129–131]. From each resulting list of gene-signatures sorted by normalized enrichment score (NES), the top 20 significantly enriched gene-signatures (NES > 1.6, nominal $P < 0.05$, and FDR $q < 0.3$) were selected. To identify common genes across these gene-signatures, those genes, which were involved in more than three gene-signatures, were extracted using leading edge analysis (LEA) [130] yielding two new top gene-signature based gene lists for T2E-positive and -negative samples (topGL-pos and topGL-neg) for further analysis.

Later, GSEA was repeated several times. In contrast, the lists were compiled from T2E-negative samples only and were based on genes that were similary ranked as before (Section 2.4). Moreover, these T2E-negative samples were, stratified by their median expression of five specific genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) into a high and low expression subgroup, each. Within each subgroup of T2E-negative samples, the fold change between cases with and without metastasis indicated the ranking of genes for GSEA, as described in Section 2.4.2. Once more, from GSEA on every subgroup, the top 20 gene-signatures were taken. The top 20 gene-signatures from GSEA between the corresponding subgroups (low/high) were compared for each of those five genes.

2.7.2 | Transcriptomic association testing

All candidate genes from both lists, topGL-pos and topGL-neg (Section 2.7.1), were separately tested in both PCa cohorts (TCGA-PRAD and GSE46691) for potential association with metastasis. Using Mann-Whitney-U test [99], the significance of differential expression in PCa with and without metastasis at diagnosis was determined. *P*-values were not adjusted for multiple testing. Only genes being significantly associated (significance level $\alpha < 0.05$) with metastasis in both cohorts were considered, furthermore, in transcriptome analysis.

2.7.3 | Network analysis

Based on all genes from both gene lists (topGL-pos and topGL-neg) resulting from GSEA and LEA in Section 2.7.1, three different networks were created showing genetic interaction, pathways, and physical interaction of the corresponding genes of each list with

Cytoscape [132]. Additionally, with a plug-in called GeneMania [133], the number of genes in each network was doubled by extending the list with functionally similar genes identified using available genomics and proteomics data. All networks were arranged and optimized in organic layout.

2.7.4 | Survival analysis

Even though this type of analysis is called survival analysis, the event of interest is not necessarily death. In cancer research it is also common to observe patients until a certain event or their drop out of the study. These cases are then referred to as censored. The occurring event can cover biochemical relapse (BCR) or event-free survival (EFS) additionally referring to the absence of metastasis, death, appearance of a new tumor. The survival time refers to the elapsed time from diagnosis until corresponding event. Resultant survival rates of different patient groups were compared and the corresponding *P*-values were calculated using the Mantel-Haenszel test. Survival analyses were carried out using the Kaplan-Meier method and the survival package in R [99, 134].

Survival analyses performed in this thesis were based on three different approaches of patient grouping:

In the first approach, patients with genomic data were grouped by their genotype (0/0, 0/1, 1/1). All resulting survival curves were compared to each other.

The second approach, referred to datasets with available transcriptome data. Samples were stratified into quartiles regarding their intratumoral gene expression levels. Only patient groups with the most extreme gene expression (highest versus lowest; Q1/Q4) were compared. Some survival analyses based on this approach were carried out on T2E-positive and negative sub-cohorts.

The third approach mainly refers to clinical data such as the Gleason Grade (GG). T2E-positive and -negative sub-cohorts were split by their GGG into two groups (GGG I-III and IV/V) and compared to each other. Moreover, this analysis was expanded by adding transcriptome data to analyze the potential added value of biomarker in addition to the GGG. For this specialized Kaplan-Meier survival analysis, patients of both GGG I-III and IV/V were stratified by their intratumoral gene expression levels of specific genes split into low and high (cut-off = 80^{th} percentile). The *P*-values for the difference between high and low gene expression levels were calculated separately for

GGG I-III and IV/V.

2.7.5 | GWAS - genome-wide association study

Imputed germline variants for all three cohorts ($PCGA^{LMU}$, TCGA-PRAD, ICGC-CA) were tested separately for several clinical features with PLINK [97, 98]. Beside using a homogenous population of Central European male patients, variants were filtered based on genotype call rate ≥ 95%, Hardy-Weinberg-Equilibrium ($P > 1 \times 10^{-5}$), and a minor allele frequency ≥ 1%, to ensure good sample and marker quality [116]. Additionally, the patients' age at diagnosis was used as covariate. For a GWAS on clinical features split into two groups, the logistic regression was applied. For association testing based on multiple ordinal scaled groups or ratio scaled data the linear regression was used instead. Variant associations were visualized as Manhattan and Q-Q plots (with genomic inflation factor λ) using qqman package in R [99, 135].

Clinical feature every cohort was tested for (patient characteristics of all cohorts are shown in Table 2.1):

- T2E fusion: T2E-positive vs T2E-negative PCa samples (2 groups)
- Gleason Grade Group: I+II vs III vs IV+V (3 groups, ordinal)
- Tumor growth: pT1+2 vs pT3+4 (2 groups, ordinal)
- PSA: PSA > 1 (linear)

2.7.6 | Meta-analysis

The results from the GWAS analyses with PLINK [97, 98] were formatted with python to match the input required for meta-analysis. Meta-analysis combined the P-values of all variants from all GWAS, which were tested for the same clinical feature, applying METAL [136] weighted by inverse variance using their P-value and standard error. Depending on the used regression, β (for linear) or log odds ratio (OR; for logistic) were used as effect. Additionally, heterogeneity I^2 across cohorts of every SNP was calculated.

From 7.3 million SNPs of $PCGA^{LMU}$, 9.2 million SNPs of ICGC-CA, and 2.3 million SNPs of TCGA-PRAD tested in the GWAS, only those SNPs were considered in the meta-analysis that were tested not only in the $PCGA^{LMU}$ cohort, but also in at least one of

the other two cohorts. Final SNP associations of the meta-analysis were visualized as Manhattan plot with genome-wide significance of $P < 5 \times 10^{-8}$ using qqman package in R [99, 135].

2.7.7 | Lead SNP identification via clumping

With the clumping procedure in PLINK [97, 98], lead SNPs with genome-wide significance could be identified from the results of the meta-analysis. A lead SNP refers to the most significant SNP of all SNPs in a region that are in LD with each other. SNPs with $P < 5 \times 10^{-8}$ were considered as genome-wide significant for whole genome based analyses.

2.7.8 | Fine-mapping of lead SNPs

The haploblock of the lead SNPs, which is restricted by a high recombination rate of variants, was investigated closely with LocusZoom [137]. The lead SNP was highlighted in purple, while the remaining SNPs were colored by their r^2 value against the lead SNP. With LocusZoom, different lead SNPs on the same haploblock, which are not in LD with each other, but exhibit a highly significant association signal, could be detected. Also the second identified lead SNP was visualized with LocusZoom as described above.

2.7.9 | Evaluation of lead SNPs and determination of representative tag SNPs

Lead SNPs and those in LD ($r^2 > 0.4$) with the lead SNPs, were evaluated by heterogeneity (I^2) of the meta-analysis. Additionally, the association effect (95% CI) of each SNP resulting from each single cohort was compared to the overall effect of the meta-analysis in a forest plot created with the forestplot and meta packages in R [99, 138, 139]. Furthermore, it is crucial for meta-analysis that the effect of the SNP points into the same direction for each of the tested cohorts, thus, affecting the same allele. For lead SNPs with high heterogeneity ($I^2 > 50\%$), the next most significant SNP in LD and $I^2 < 50\%$ was selected as representative tag SNP. These lead SNPs were analyzed together with their corresponding representative tag SNPs and referred to as candidate SNPs in this thesis.

2.7.10 | Effect size of risk allele

In PLINK [97, 98], the effect size was calculated against the minor allele. Due to several tested ordinal scaled groups, the effect was estimated by the regression coefficient. A positive regression coefficient indicates an increased risk for patients possessing the minor allele, while a negative value identifies the major allele as risk allele (RA) [97, 98]. With an odds ratio (OR) the effect size between the groups can be measured. Therefore, based on the number of risk and non-risk alleles the odds for the lowest ordinal group is calculated as baseline. Afterwards, the odds of the other groups are calculated and compared to the baseline as OR.

2.7.11 | Expression quantitative trait locus analysis

For detected candidate SNPs, expression quantitative trait locus (eQTL) analysis was run on datasets with both genomic and transcriptomic data available. Therefore, variant annotation regarding eQTL genes acting in cis or trans with the candidate SNPs were obtained from SNiPA [140]. With linear regression between those genes and the genotypes of the candidate SNPs, eQTLs were estimated and visualized using the beeswarm package in R [99, 141].

2.7.12 | Epigenetic fine-mapping

Candidate SNPs were also examined in their epigenetic context. ChIP-seq data from PC-3 cell lines were downloaded from ENCODE [142] (Table 2.8). Beside DNase (ENCFF504HAN), ChIP-seq data covers transcription factor CTCF (ENCFF785IOE), and histone H2AFZ (ENCFF275MKL) as well as histone modifications H3K36me (ENCFF231NLM), H3K27ac (ENCFF940SZQ), and H3K9me3 (ENCFF209KEN). ChIP-seq reads were aligned on hg19 and visualized in IGV [143].

2.7.13 | Functional annotation of eQTL variant

For functional annotation, the most significant eQTL variant was annotated with Annovar and ensembl's Variant Effect Predictor (VEP) [144].

2.7.14 | Conditional analysis

Conditional analysis was executed for the eQTL SNP. Therefore, a GWAS on variants of $PCGA^{LMU}$ against Gleason score was repeated in PLINK (Section 2.7.5) for variants on the same haploblock as the eQTL SNP, but with the genotype of eQTL SNP as additional covariate.

2.8 | STARLING - webserver organization and development

A web service called STARLING (pro**STA**te cancer **R**esearch **L**everaging **I**mportant **N**ovel **G**enomic biomarkers) was established to enable a user friendly and intuitive access to the association testing results of our GWAS comprising three cohorts ($PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA).

STARLING helps other researchers of the prostate cancer field to identify biomarkers that are associated with certain clinical features or somatic mutations.

2.8.1 | Technical aspect and structure

The web service architecture is based on the client-server model (Figure 2.7). On client-side the website content and structure was written in HTML using the Skeleton framework and designed via CSS (Cascading Style Sheet). Interactions on the website were implemented in JavaScript using jQuery library. Requests sent from client-side with Ajax (Asynchronous JavaScript and XML) are processed on server-side by PhP. Afterwards, the compiled query is passed to the MySQL database, which is stored on the server. From server-side, its JSON response is send back to the client-side, where it is prepared and formatted into the corresponding outcome (table and plot) via JavaScript.

2.8.2 | Database design and data integration

The web service is based on a MySQL database, which stores the test results for all association tests, SNPs, and cohorts. Additionally, every tested SNP is listed with its meta data. Database design is described in Figure 2.8. With a python script the local

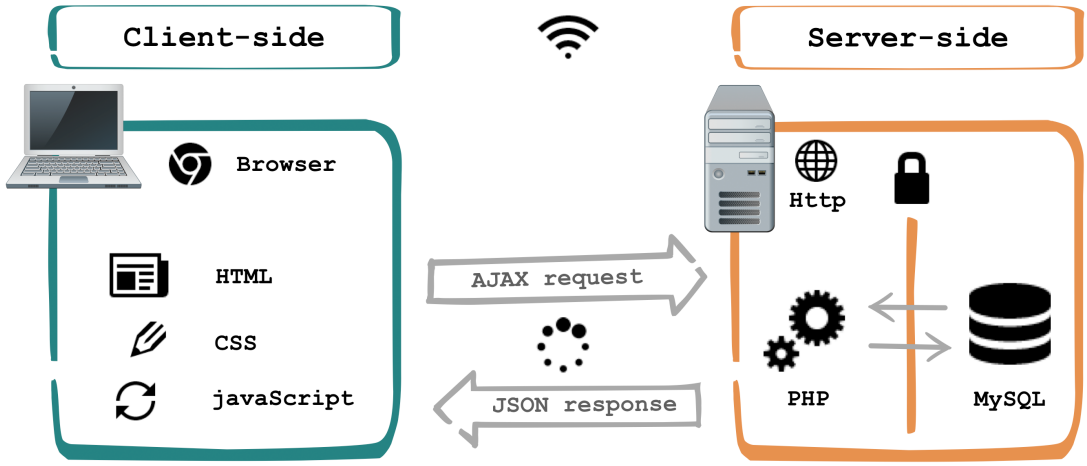


Figure 2.7: Schematic client-server model describing the architecture of STARLING.

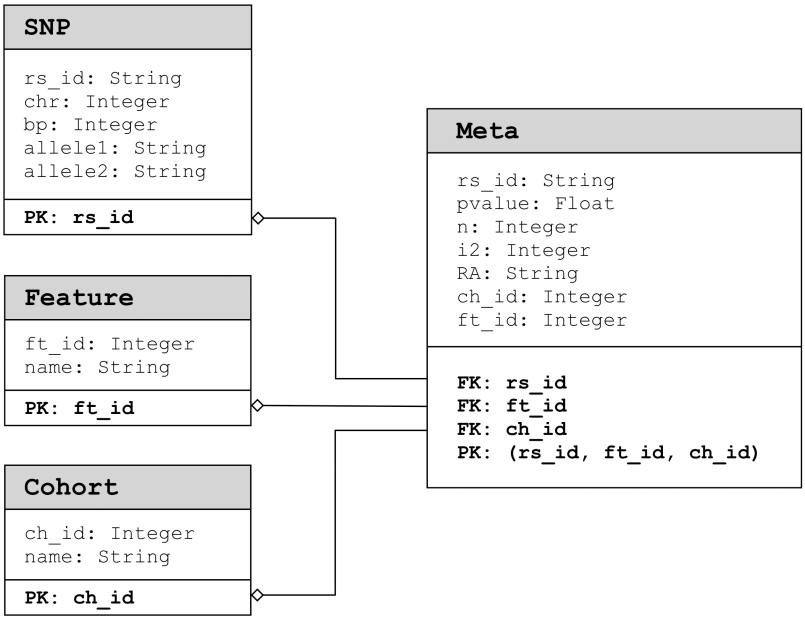


Figure 2.8: Database design used for STARLING. (PK = primary key; the remaining short cuts refer to variable names in the database)

database was created and loaded. The exported database was then uploaded onto the server via MyPhpAdmin.

2.8.3 | Data visualization of response

The requested results received from the database are shown in a table, via the DataTables plug-in for the jQuery JavaScript library, and visualized in an interactive Manhattan plot, created with JavaScript library HighCharts.

The table exclusively shows the query SNPs sorted by ascending *P*-value . Location, alleles, and risk allele of the SNPs are listed as well as the number of cohorts used for association testing and its resulting heterogeneity. Additionally, a star symbol informs about significance or genome wide significance. The table can be sorted interactively and enables full text search on SNP IDs.

Beside the table, a Manhattan plot supports the requested GWAS results. It shows the query SNPs in context to the remaining significant SNPs (P-value > 0.05) tested for association to the same phenotype. The intuitive plot highlights the query SNPs in orange. Furthermore, the interactive plot enables to identify other SNPs by hovering over them with a cursor, that could also be of the user's interest but have a lower *P*-value . By doing so, a little pop up information window to the corresponding SNP is shown.

2.8.4 | Web service security

To ensure maximum security, database access authorization and PhP scripts building a database connection were hidden in the server structure. Instead of unpopular captchas, the invisible honeypot method was applied to the request submission form to prevent bots from overloading or crashing the server with random automatically generated requests.

2.8.5 | Evaluation

Before its official launch, the website was tested for its web appearance, performance, reachability, and findability. Therefore, comprehensive search engine optimization (SEO) was done with WebSite Auditor (v4.38.1) of SEO PowerSuite Free.

2.8.6 | Hosting

The STARLING web service is hosted on the webserver (webdev-lmu.lrz.de) of the Leibniz-Rechenzentrum (LRZ) and accessible via the following domains:

- www.starling-pcga.med.uni-muenchen.de
- www.starling-pcga.med.lmu.de

2.9 | Software and file summary

Table 2.5: Most important software packages and plug-ins used in this thesis with programs and languages listed in Table 2.6 and 2.7.

Name	Software	Version	Usage	Reference
ajax/jQuery	javaScript	3.3.1	Data requests	
AnnotationDbi	R	1.46	For annotation in bioconductor	[145]
beeswarm	R	0.2.3	Beeswarm plot	[141]
BiocManager	R	1.30.4	Bioconductor package installer	[146]
DataTables	javaScript	1.10.19	Interactive tables	
ESTIMATE	R	1.0.13	Tumor purity predictor	[101]
fontawesome	CSS	5.8.1	Graphic symbols	
forestplot	R	1.9	Forest plot	[138]
genefilter	R	1.66	Filter genes	[113]
HighCharts	javaScript		Interactive plots	
meta	R	4.9	Meta-analysis	[139]
mclust	R	5.4	Population clustering	[100]
org.Hs.eg.db	R	3.8.2	human annotation database	[147]
pd.huex.1.0.st.v2	R		Gene level summarization	[107]
qqman	R	0.1.4	Manhattan and QQ plots	[135]
SCAN.UPC	R	2.28.0	Microarray normalization	[106]
skeleton	CSS		HTML framework	
survival	R	2.38	Kaplan Meier survival plots	[134]
sva	R	3.32.0	ComBat	[148,149]

Table 2.6: Programming languages and development environments used throughout this project to implement scripts and run bioinformatic software listed in Table 2.7.

Language/ Environment	Version	Description	Online
Awk	5.0.0	Text processing and data extraction	gnu.org/software/gawk/
Bash	4.4.12	Unix-shell	gnu.org/software/bash/
Bioconductor	3.9	Open source software for bioinformatics	bioconductor.org/
CSS	3	Style sheet language for website layout and presentation	
Cygwin	3.0.7	Linux Environment for Windows	cygwin.com/
HTML	5	Hypertext Markup Language for creating web pages	
javaScript	1.8.5	For interactive web pages	
Java	1.8.0 (181)	General-purpose programming language	java.com/
MySQL	8	Database management system	mysql.com/
Perl	5.26.3	General-purpose programming language	perl.org/
PhP	7.1	Scripting language for web development	php.net/
Python	2.7.14	general-purpose programming language	python.org/
R	3.5.0	Language and Environment for statistical computing	r-project.org/
XAMPP	3.2.2	PHP development environment	apachefriends.org/

Table 2.7: Bioinformatic software used in this thesis.

Software	Version	Usage	Reference	Online
Annovar	2018-04-16	Functional annotation of genetic variants	[110]	annovar.openbioinformatics.org
Bcftools	1.9	Manipulating VCF files	[119]	samtools.github.io/bcftools/
Cytoscape	3.5.1	Network Data Integration and Visualization	[132]	cytoscape.org/
FastQC	0.11.8	Quality check for bam files	[120]	bioinformatics.babraham.ac.uk/projects/fastqc/
GATK	4.1.2.0	Genome Analysis Toolkit	[123]	software.broadinstitute.org/gatk
gdc-client	0.5.4	Access to the GDC data portal (TCGA)	[90]	gdc.cancer.gov/access-data/gdc-data-transfer-tool
GSEA	3.0	Gene set enrichment analysis software	[129,130]	software.broadinstitute.org/gsea/
HRC-1000G-check-bim.pl	4.2.9	Comparison of variants to reference genome	-	well.ox.ac.uk/~wrayner/tools/index.html# Checking
IGV	2.3.97	Integrative Genome viewer	[143,150]	software.broadinstitute.org/software/igv/

Continues on next page.

Table 2.7 continued: Bioinformatic software used in this thesis.

Software	Version	Usage	Reference	Online
LocusZoom	1.3	GWAS result visualization	[137]	locuszoom.org/
METAL	2011-03-25	Meta-analysis of GWAS	[136]	csg.sph.umich.edu/abecasis/metal/
Michigan Imputation Server	1.0.4	Free Next-Generation Imputation Server	[125]	imputationserver.sph.umich.edu
MultiQC	1.7	Summarizes FastQC results	[121]	multiqc.info/
Picard	2.18	Manipulating high-throughput sequencing data	[122]	broadinstitute.github.io/picard/
PLINK	1.90b3.31	Whole genome association analysis tool	[97,98]	cog-genomics.org/plink/1.9/
RMAexpress	1.2.0	Normalization of microarray data	[151,152]	rmaexpress.bmbolstad.com/
Samtools	1.3.1	Manipulate SAM/BAM files	[119]	htslib.org/
Score-client	1.5.0	ICGC download tool	-	docs.icgc.org/download/guide/
SNiPA	3.3	Annotating genetic variants	[140]	snipa.helmholtz-muenchen.de/snipa3/

Continues on next page.

Table 2.7 continued: Bioinformatic software used in this thesis.

Software	Version	Usage	Reference	Online
SnpSift	4.3	Variant annotation	[124]	snpeff.sourceforge.net
Vcftools	0.1.15	Manipulating VCF files	[118]	vcftools.github.io/
VEP	release 98	Variant effect predictor by Ensembl	[144]	ensembl.org/Homo_sapiens/Tools/VEP
WebSite Auditor SEO Power Suite	4.38.1	Comprehensive search engine optimization	-	link-assistant.com/

Table 2.8: Data sets used in this thesis and corresponding download information.

Name	Files	Download Date	URL
1000 Genome Project (hg19)	All.chr*.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf	22.06.16	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
ChIP-seq on PC-3		27.06.19	www.encodeproject.org
ICGC reference genome (HS37D5)	genome.fixed.fa	01.03.17	http://dcc.icgc.org/releases/PCAWG/reference_data/pcawg-bwa-mem
Reference genome used for TCGA-PRAD (GRCh38)	GRCh38.d1.vd1.fa	18.09.18	https://gdc.cancer.gov./about-data/data-harmonization-and-generation/gdc-reference-files
SNP database (hg19)	dbsnp_138.hg19.vcf	07.06.19	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/
Resource files (hg19)	hapmap_3.3.hg19.sites.vcf, 1000G_omni2.5.hg19.sites.vcf, 1000G_phase1.snps.high_confidence.hg19.sites.vcf	12.07.19	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/

Continues on next page.

Table 2.8 continued: Datasets used in this thesis and corresponding download information.

Name	Files	Download Date	URL
SNP database (GRCh38)	dbsnp_146.hg38.vcf	07.06.19	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/
Resource files (GRCh38)	hapmap_3.3.hg38.vcf, 1000G_omni2.5.hg38.vcf, 1000G_phase1.snps.high_confidence.hg38.vcf.	12.07.19	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/
Lift Over reference	hg38ToHg19.over.chain	03.09.18	http://hgdownload.cse.ucsc.edu/gbdb/hg38/liftOver/
UCSC reference genome	hg19.fa	07.06.19	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/
SNP annotation (GRCh37p13_151)	All_20180423.vcf	13.05.19	ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/
HRC reference	HRC.r1-1.GRCh37.wgs.mac5.sites.vcf	23.04.19	www.haplotype-reference-consortium.org/site
Small TCGA-PRAD data assembly	TCGA-*.oxoG.snp.capture.tcga.vcf	12.01.16	https://portal.gdc.cancer.gov/legacy-archive/

3

Results

Most of the results presented in Section 3.1 have previously been published in the peer-reviewed International Journal of Cancer by Gerke *et al.* [105].

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

PCa composes several distinct molecular subtypes [53, 58] that can mainly be split into T2E-positives and T2E-negatives. In the first part of this thesis, the transcriptomes of collectively 783 PCa of two public cohorts (TCGA-PRAD and GSE46691) with matched clinicopathological data were analyzed, to emphasize the molecular differences associated with metastasis of both PCa subtypes. Contrary to other common clinical records associated with PCa, information on metastasis was available for both cohorts. Thus, metastasis was selected as surrogate for PCa aggressiveness.

Using multiple filtering steps regarding variance and regulation, a unity of 3,068 variably expressed genes was left for further analysis (Figure 2.3). After distinguishing the transcriptomes by their samples T2E-status, genes were examined for T2E-positive and -negative PCa separately, to emphasize the fusion associated differences in PCa transcriptomes. With GSEA (Figure 2.4), followed by LEA, association testing against

metastasis and survival analysis, five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) could be identified, whose high expression is associated with worse outcome in T2E-negative PCa, exclusively. Furthermore, several of them add prognostic information to clinicopathological predictors. Contrary, no similar observation was made for T2E-positive PCa.

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

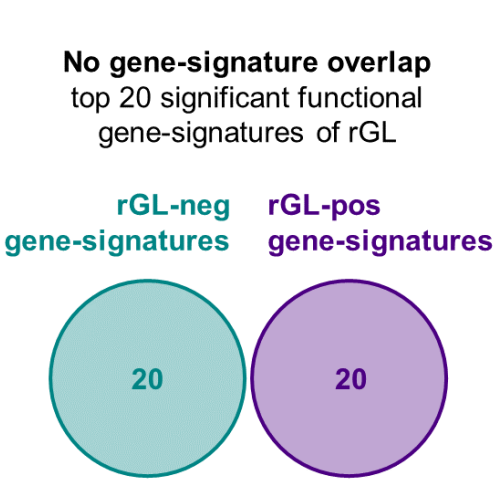
Using multiple filtering steps regarding variance and regulation, a unity of 3,068 variably expressed genes was left for GSEA. By distinguishing the transcriptomes by their samples T2E-status, those genes were ranked by their expression fold change between patients with and without metastasis (Figure 2.3). On the resulting two ranked gene lists, rGL-pos and rGL-neg, a GSEA was run selecting the resulting top 20 significant gene-signatures obtained for each gene list. There was no overlap between the top 20 significant metastasis associated gene-signatures of both T2E-positives (rGL-pos) and -negatives (rGL-neg) (Figure 3.1a, Table A.1, and A.2).

Using LEA, genes involved in more than three of the selected top 20 gene-signatures were extracted to create two new 'top gene-signature' gene lists (topGL-pos and -neg, Figure 2.4). Accordingly, 16 genes of rGL-pos were overrepresented in significant gene-signatures of T2E-positives and summarized in topGL-pos (Table A.3). TopGL-neg contained 74 genes frequently occurring in significant gene-signatures of T2E-negatives (rGL-neg, Table A.4). By comparing the genes of topGL-pos to topGL-neg, only two genes (*RRM2* and *TYMS*) were part of both lists (Figure 3.1b). However, both shared genes were involved in different gene-signatures of T2E-positive and -negative cases.

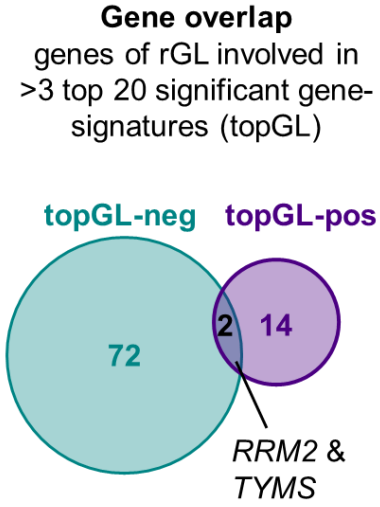
Altogether, these results indicated that T2E-positive and -negative PCa are characterized by distinct metastasis associated gene signatures [105].

3.1.2 | Frequent genes involved in metastasis associated gene-signatures are predominantly coding genes

Beside protein coding genes, non-coding genes were examined accordingly regarding miRNAs, ncRNAs, and lncRNAs. However, only 20 non-coding genes were found



(a) Overlap between top 20 significant gene-signatures of rGL-pos and -neg as identified by GSEA [105].



(b) Overlap of overrepresented genes in top metastasis associated gene-signatures in T2E-positive and -negative cases [105].

Figure 3.1: Venn diagrams showing a) T2E-positive and -negative PCa are characterized by distinct metastasis associated gene-signatures and b) the overlap between the corresponding overrepresented genes of these gene-signatures.

among the unity of genes from both discovery datasets. Only one of those (*DLEU2*) was among the top 20 significantly enriched gene-signatures, but too infrequent to be included in topGL-pos or -neg gene list as candidate. Therefore, it was not further pursued with non-coding genes [105].

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

Subsequently, all genes of both gene lists (topGL-pos and -neg) obtained from the significant gene-signatures were tested in the discovery cohorts (TCGA-PRAD and GSE46691) separately for significant differential expression depending on the formation of metastases. In T2E-positive cases (topGL-pos), three genes (*GMNN*, *TROAP*, and *WEE1*) out of 16 were significantly ($P < 0.05$) higher expressed in PCa samples with metastasis (Table 3.1). In T2E-negative cases (topGL-neg) 29 of 74 genes were significantly ($P < 0.05$)

higher expressed in PCa patients with metastasis. By comparing these significantly differentially expressed and metastasis associated genes in the two gene lists, no overlap could be found (Table A.3 and A.4).

These results suggest that, depending on the T2E-status, distinct genes are linked to metastasis in PCa [105].

3.1.4 | Metastasis associated genes are not forming hubs in networks to enable gene function prediction

For every gene list (topGL-pos and -neg) three networks, visualizing genetic interactions, pathways, or physical interactions between genes, were created, each. Beside those genes, each network was extended with the same amount of functionally similar genes. This results in three networks with 32 genes based on topGL-pos (Figure A.10) and three networks consisting of 148 genes based on topGL-neg (Figure A.11).

None of the networks included all genes of each gene list in one cluster. While both genetic interaction networks formed one main cluster and several single nodes, the other four networks (physical interaction and pathway, respectively) additionally formed several small clusters around single genes of the gene lists. These network pattern did not allow an implication on the gene function affecting the aggressiveness of PCa. The genetic interaction networks of both topGL-pos and topGL-neg (Figures A.10c and A.11c) were highly connected and included many of the genes from the corresponding gene list. But, no hub gene, which is defined by a very high node degree and a low connectivity between its neighbor nodes and whose loss would destroy the whole network, could be identified. This indicates that none of these genes played a major role in the networks, which could correspond to an important effect in the development of metastases. Also, each network in its entirety did not reflect a coherency between the genes from the gene lists, topGL-pos and topGL-neg.

Due to the observed pattern of multiple small clusters and missing hub genes in the examined networks, no biologic conclusion regarding gene function could be drawn from the genes of topGL-pos and topGL-neg, respectively. Therefore, network analysis was not pursued.

3.1.5 | Identified prognostic biomarkers are subtype specific

Based on Kaplan-Meier analysis of two independent cohorts, a potential correlation between the identified metastasis associated genes and event-free survival (EFS) was explored. Apart from the TCGA-PRAD cohort, this analysis was supported by another independent microarray-based validation cohort, GSE16560, to verify the findings. As before, only those genes significantly ($P < 0.05$) and concordantly associated with EFS in both cohorts were further pursued. None of the genes from topGL-pos was significantly associated with EFS in T2E-positive PCa and, therefore, not accepted as candidate as prognostic biomarker (Table A.3). Contrary, seven genes identified based on T2E-negative cases were consistently associated with EFS (*APOE*, *ASPN*, *BGN*, *COL1A1*, *LY96*, *RRM2*, and *TYMS*) and suit potential prognostic biomarkers. All seven identified genes accorded with their observation of higher expression levels being associated with shorter EFS (Figure 3.2, Table A.4). Strikingly, this effect of consistent association with EFS could not be seen for these genes in T2E-positives cases (Figure 3.2).

However, only five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) were associated with metastasis (Section 3.1.3) and EFS in both discovery cohorts and the first validation cohort as summarized in Table 3.1. Thus, these five identified genes from topGL-neg qualified for potential subtype specific biomarkers and were further validated.

Altogether, this confirms that the identified potential prognostic biomarkers *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS* could be employed for outcome prediction in T2E-negative PCa, exclusively [105].

3.1.6 | Common mutations associated with PCa did not bias previously conducted results

Beside T2E, multiple other additional molecular events affecting tumor suppressors can occur in PCa and lead to worse outcome [53, 55]. The TCGA-PRAD cohort was re-investigated for common mutations in PCa regarding suppressor genes *SPOP*, *TP53*, and *PTEN*, which could be inferred from exome sequencing data [53]. These mutated genes were investigated for a potential bias on previous results regarding EFS in Section 3.1.5.

The most frequent mutations are affecting the *SPOP* gene and occur in around 10%

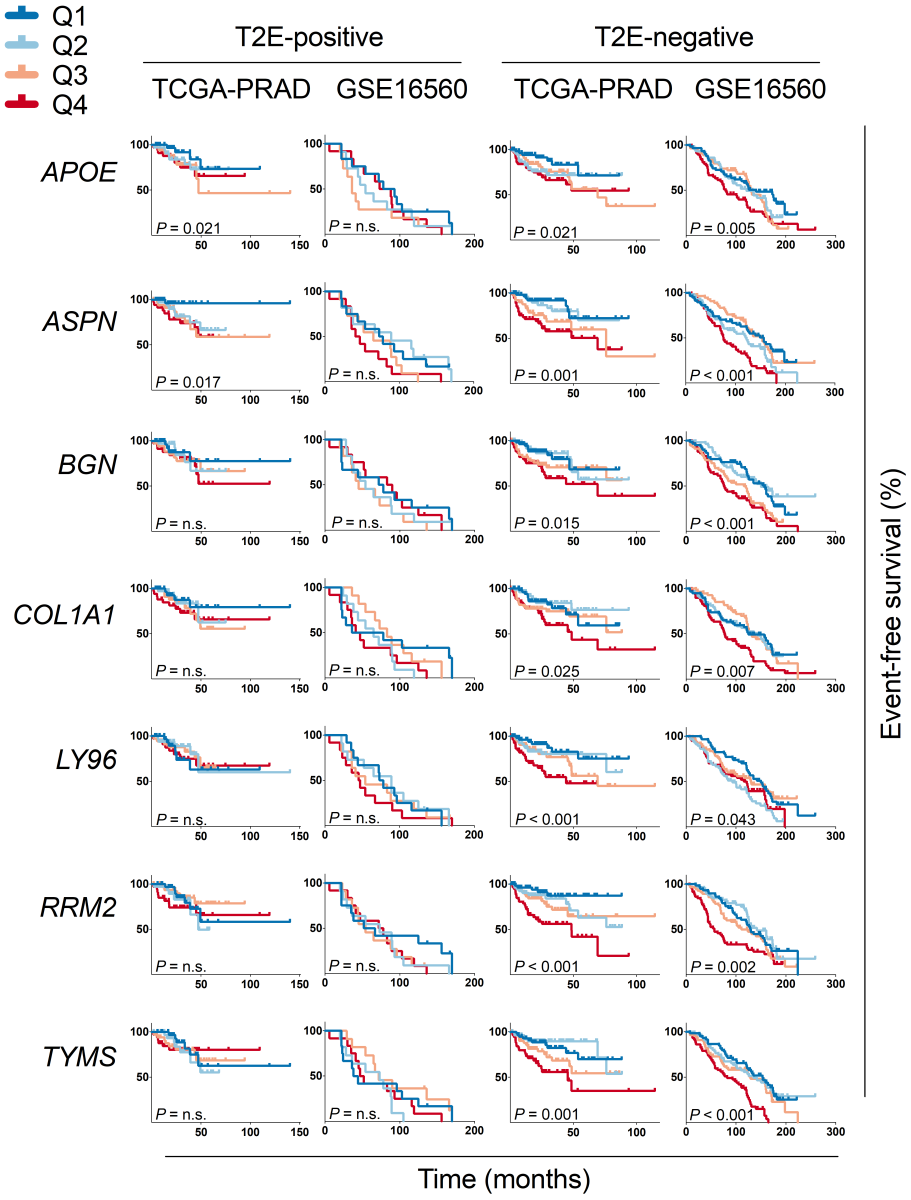


Figure 3.2: The prognostic value of identified biomarkers depends on the T2E-status [105]. Kaplan-Meier plots for event-free survival (EFS)-associated genes (*APOE*, *ASPN*, *BGN*, *COL1A1*, *LY96*, *RRM2*, and *TYMS*) of topGL-neg derived from samples of the TCGA-PRAD and GSE16560 cohorts. Patiens were split by their T2E-status (positive/negative) and stratified by their quartile intratumoral gene expression level of the corresponding gene. Using Mantel-Haenszel test, *P*-values were calculated between the lowest (Q1) an highest (Q4) gene expression quartiles.

Table 3.1: Result summary of genes in both gene lists (topGL-pos and topGL-neg) that passed at least one of the applied tests (association test against metastasis and survival analysis on EFS) for all cohorts, as well as those two genes (*RRM2* and *TYMS*), which were included in both gene lists. Data for all genes was extracted from Table A.3 and A.4. Genes being significant in all tests are highlighted in bold font. [105]

Dataset	GSE46691	TCGA			GSE16560	
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
LY96	n.s.	<0.001	0.001	low	0.043	low
RRM2	0.044	<0.001	<0.001	low	0.002	low
TYMS	0.009	0.018	0.001	low	<0.001	low

of PCa, but in T2E-negative cases exclusively [53, 153]. The removal of 20 PCa cases harbouring a mutation in the *SPOP* gene from the T2E-negative sub-cohort did not affect the significance of previous results of the five subtype specific genes being associated with clinical outcome, described in this thesis. These results are not shown as they hardly differ from those in Figure 3.2. This suggests that *SPOP* mutations do not impact the validity of *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS* for T2E-negative tumors [105]. Furthermore, two additional genes (*TP53*, *PTEN*) possessing common mutations in PCa were investigated, like *SPOP* mutations before, as they could have biased the results from the TCGA-PRAD cohort (overall mutation frequency of 7% and 2%, respectively). Beside eleven TP53-mutated cases, only two PCa samples harbouring a PTEN mutation could be identified in the T2E-negative TCGA-PRAD cohort. Again, removing these minor amount of mutated cases from the T2E-negative sub-cohort did not impact the significant associations of *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS* with clinical outcome.

This indicates, that none of the examined gene mutations in *SPOP*, *TP53*, or *PTEN* could have confounded the previous results of this thesis [105].

3.1.7 | T2E-negative PCa stratified by candidate biomarker expression deviate in their metastasis associated gene-signatures

Five identified subtype specific genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) were investigated, separately, regarding their gene expression level (low/high) being associated with different gene-signatures. GSEA was repeated for T2E-negative subgroups that were split by their corresponding gene expression into high and low for each of the five genes. By comparing the top 20 gene-signatures resulting from the high and low subgroup for each gene, only a minor amount of gene-signatures did overlap (Tables A.5, A.6, A.7, A.8, and A.9). The shared gene-signatures between the high and low subgroups ranged from 15% for *COL1A1* to 45% for *RRM2*. The median overlap for all five genes accounts for 38% of their corresponding top 20 gene-signatures.

These slight overlaps imply that T2E-negative tumors expressing high or low levels of the given candidate marker genes may be driven by mostly distinct pathways and thus may differ in their (patho)biology [105].

3.1.8 | Validation by IHC endorsed *RRM2* and *TYMS* as biomarkers for T2E-negative cases

With IHC a potential T2E dependent prognostic value of PCa biomarkers should be explored. Therefore, TMAs from 135 PCa cases (TMA-cohort) were stained by IHC for RRM2 and TYMS. These two genes were chosen as representatives of the five identified genes, as for both a specific antibody was available. The BCR-free survival (BFS) of T2E-positive and -negative cases were analyzed separately by stratifying the patients by their percentage of RRM2-positive tumor cells as well as TYMS-positive tumor cells. For RRM2 a cut-off $\geq 3\%$ and for TYMS a cut-off $\geq 5\%$ was selected according to the respective median percentage of both RRM2- and TYMS-positive tumor cells across the complete cohort.

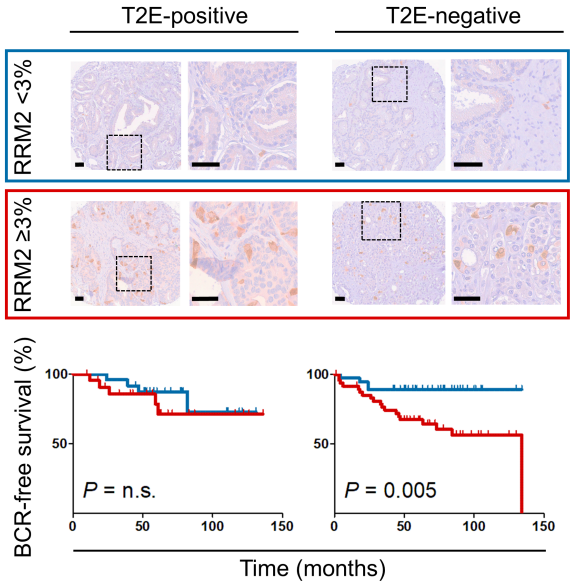
Patients with T2E-negative PCa, which possess a high percentage of RRM2-positive tumor cells, could be observed with significant ($P = 0.005$) worse BFS than those with low RRM2-positivity (Figure 3.3a). Moreover, a significant ($P = 0.004$) lower BFS rate for patients with T2E-negative PCa exhibiting a high percentage of TYMS-positive tumor cell could be seen (Figure 3.3b). Contrary, no association of neither RRM2-positivity nor TYMS-positivity with BFS was detected in T2E-positive cases.

These results verify that the prognostic value of biomarkers in PCa depends on the T2E-status and suggest that analyses not considering the T2E-status may affect outcome prediction.

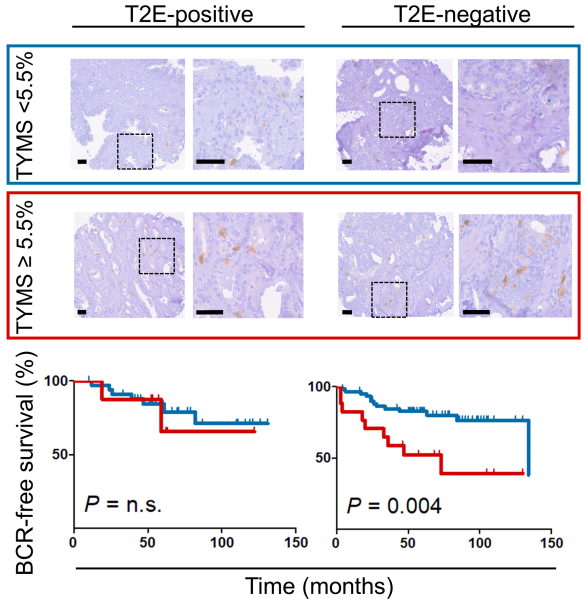
3.1.9 | Subtype specific biomarkers add prognostic information to Gleason grading

Currently, the GG is the most widely used predictor for patient outcome in PCa. Its Gleason grading groups (GGG) were recently redefined based on the Gleason score and cover five categories (I-V, Table 1.1) [50], which have been proven to be of high prognostic significance in large cohorts [43, 154]. Nevertheless, it still remains challenging and hazardous to predict the risk of developing an aggressive PCa for individuals based on GG only [155, 156].

Therefore, both cohorts (TCGA-PRAD and GSE16560) were stratified by their T2E-status followed by GGG (I-III vs. IV/V) to be compared via Kaplan-Meier analyses. As



(a) Validation of RRM2 (median percentage of positive tumor cells with cutoff ≥ 3%)



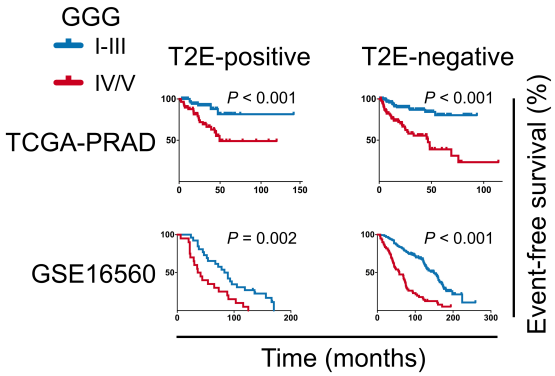
(b) Validation of TYMS (median percentage of positive tumor cells with cutoff ≥ 5.5%)

Figure 3.3: Validation of genes as prognostic biomarkers for T2E-negative samples by IHC [105]. Top a/b): Representative micrographs of T2E-positive and negative PCa stained for the corresponding genes by IHC. Scale bars = 50μM 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 median percentage of positive tumor cells using the Mantel-Haenszel test.

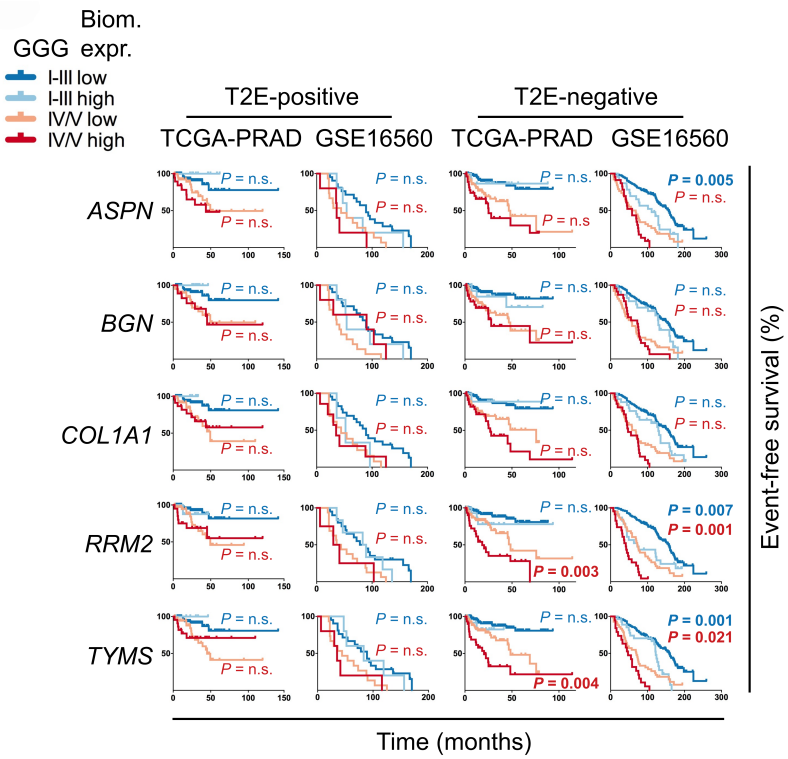
expected, a significant association ($P < 0.002$) of worse EFS with high GGG (IV/V) could be observed for both cohorts, regardless of the T2E-status (Figure 3.4a).

Even though the identified genes did not outperform the GG as biomarker, the assumption arose that they might increase the prognostic value of the GG and enhance outcome prediction together. Depending on their samples T2E-status, *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS* were examined, whether their expression level might add prognostic information to the GG. Thus, the previously described subgroups were further divided regarding their gene expression of the potential subtype specific biomarker into high and low. As shown in Figure 3.4b, high *RRM2* and *TYMS* expression was associated with signifcantly worse outcome for patients in both GGG subgroups (high-/low) with T2E-negative tumors. Interestingly, this additive prognostic effect was completely absent in T2E-positive cases of both cohorts (Figure 3.4b). For *ASPN*, *BGN*, and *COL1A1*, slight effects could be seen that either present statistical trends or reached statistical significance only in one cohort. The results are summarized in Table A.10.

Altogether, these results indicate that at least two out of five genes (*RRM2* and *TYMS*) can add prognostic information to routine GG for patients with T2E-negative PCa [105].



(a) T2E-positive and T2E-negative patients, separately, stratified by their GGG each.



(b) Samples stratified by their T2E-status, GGG (by high or low expression (cut-off = 80^{th} percentile) of the indicated biomarkers. High versus low biomarker expression were compared separately for high (IV/V, red color) and low (I-III, blue color) GGG. *P*-values are listed in Table A.10.

Figure 3.4: Kaplan-Meier analysis of EFS from the TCGA-PRAD and GSE16560 cohorts. Survival rates of patients with T2E-positive and -negative PCa were compared separately a) regarding their GG and b) additionally stratified by gene expression levels of the candidate biomarkers to emphasize the prognostic information, which subtype specific biomarkers add to GG. Patients groups were compared by calculating the *P*-value using the Mantel-Haenszel test. [105]

3.2 | GWAS on germline variants identifies potential risk loci for prostate cancer aggressivenes on 7q31.33

The following analysis results are comprised of three PCa cohorts. First, 800 PCa patients were enrolled in the $PCGA^{LMU}$ cohort that were genotyped for germline variants. The second cohort, ICGC-CA, comprises WGS data of 113 PCa samples. The last cohort, TCGA-PRAD, consists of 393 PCa subjects for which WXS data was available. An overview on processing of germline sequencing data is shown in Figure 2.5. After identifying and selecting Central European samples to achieve a homogenous study group ($PCGA^{LMU}$: n = 751, ICGC-CA: n = 84, TCGA-PRAD: n = 263), their germline variant data were imputed to enlarge their number of variants as well as to harmonize the cohorts for better combination of analysis results later (Figure 2.6). For each cohort, a GWAS was performed against several clinical features, whose results were combined by meta-analysis. From this a genome-wide significant locus 7q31.33 was detected. Its lead SNPs (rs12537032, rs191029826 and rs76326523) and corresponding representative tag SNPs (rs74999840 and rs73451279) were further fine-mapped and examined regarding their potential function and effect on PCa progression (Figure 2.6). The results of these analyses are presented in the following subsections.

3.2.1 | T2E fusion is not associated with germline variants

Germline variants of patients with a T2E-positive and -negative tumor were tested genome-wide for association in the $PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA cohort, separately. However, a meta-analysis of the results obtained from the three cohorts revealed neither a SNP with genome-wide significance nor any trend of potential association (Figure A.6a).

3.2.2 | Tumor growth and PSA show trends of germline association in PCa

Next, each cohort ($PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA) was tested for potential association of its variants with PSA and tumor growth (encapsulated vs non-encapsulated tumor). But, in both Manhattan plots from the TCGA-PRAD and ICGC-CA cohorts, some common artifacts occurred affecting the GWAS results. Therefore, these results were not pursued for further analysis. After combining the results of all three cohorts regarding tumor growth in a meta-analysis, no genome-wide significant lead SNP could be identified. Nevertheless, the resulting Manhattan plot (Figure A.6b) showed several trends for potential association on chromosomes 1, 6, 8, 9, 17, and X. These results indicate potential association of germline variants in PCa with tumor growth, which might be identified by increasing the study size or adding further cohorts to meta-analysis for clarification.

3.2.3 | GWAS identifies one genome-wide significant lead SNP on 7q31.33 locus associated with GG

Furthermore, germline variants of all three cohorts ($PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA) were tested in a GWAS against the GGG, which were categorized into three groups of lower (grade I and II), middle (grade III), and high GG (grade IV and V). The genomic inflation factor λ of all tested cohorts ranges from 1.001 to 1.078 indicating no genome-wide inflation caused by population stratification (Figure A.5). In the Manhattan plot in Figure 3.5 only one single genome-wide significant hit ($P = 2.372 \times 10^{-9}$) associated with GGG can be observed. The intronic SNP rs12537032 is located on 7q31.33 and marks a locus that is potentially associated with GGG in PCa.

3.2.4 | Fine-mapping reveals second lead SNP on 7q31.33

After identifying the genome-wide significant locus 7q31.33 via the lead SNP rs12537032, fine-mapping on the haploblock of the lead SNP, which is restricted by a higher recombination rate, revealed other significant SNPs. However, several of these SNPs were not in LD with the lead SNP. Thus, a second significant signal determined by SNP rs191029826 ($P = 1.747 \times 10^{-6}$) could be detected (blue in Figure 3.6). Another rep-

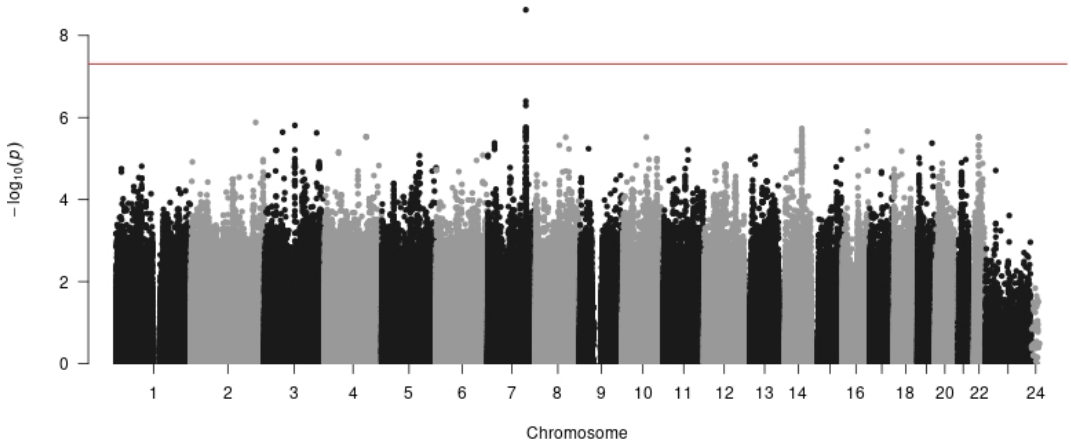


Figure 3.5: GWAS results from meta-analysis against GGG. Genome-wide significance is marked with a red line, revealing one candidate lead SNP rs12537032 on 7q31.33.

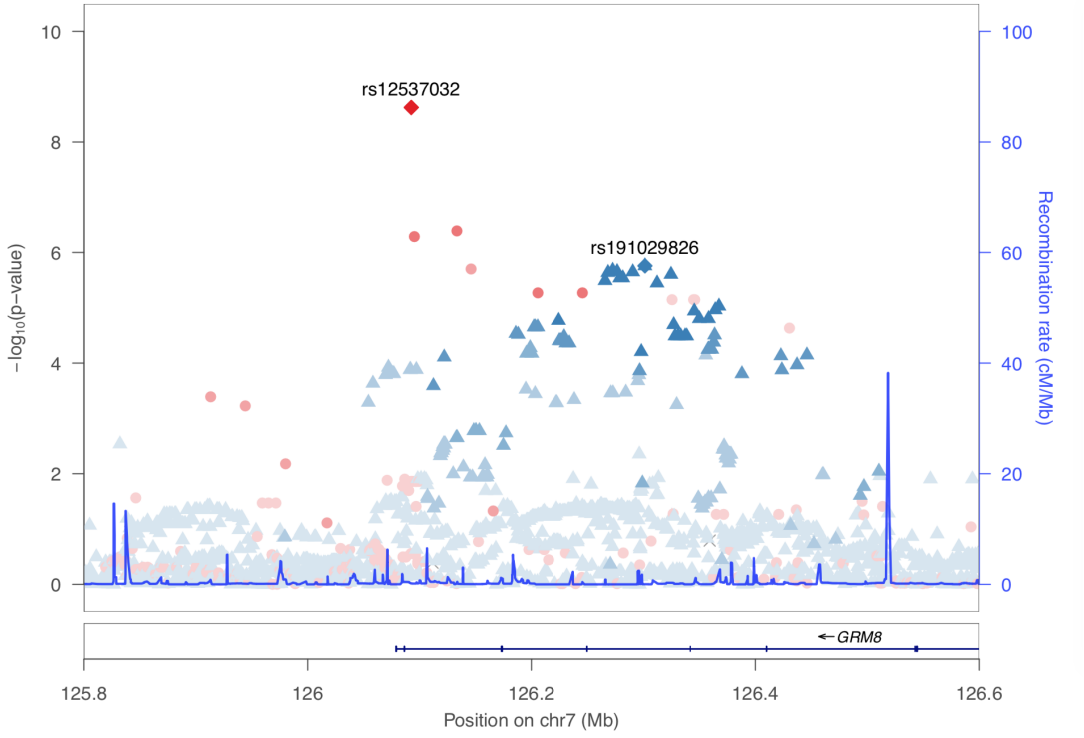
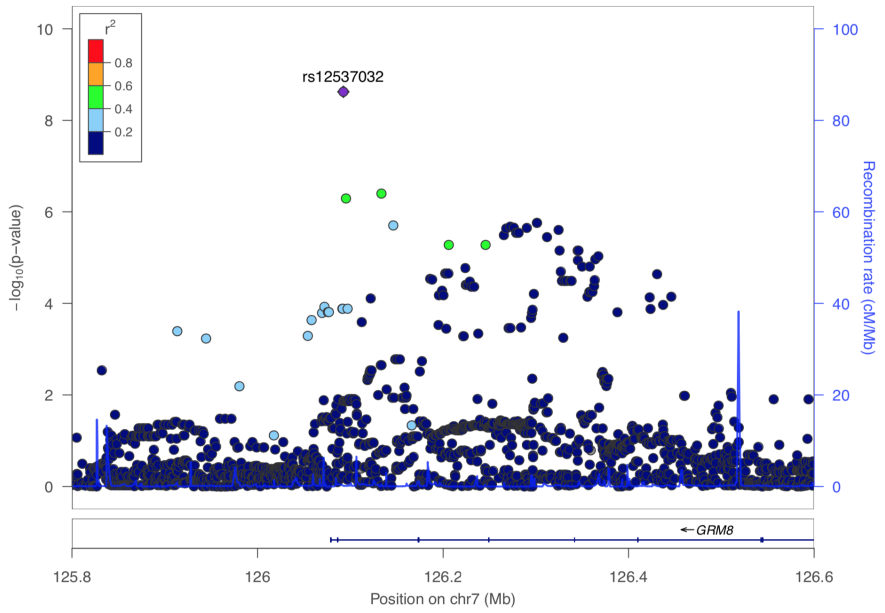


Figure 3.6: Fine-mapping of locus 7q31.33 reveals two independent signals (red and blue). Color gradient of SNPs indicate their LD to the lead SNP, whereas red refers to lead SNP rs12537032 and blue to rs191029826 and rs76326523, respectively. The latter overlaps with rs191029826 and cannot be seen in this Figure.

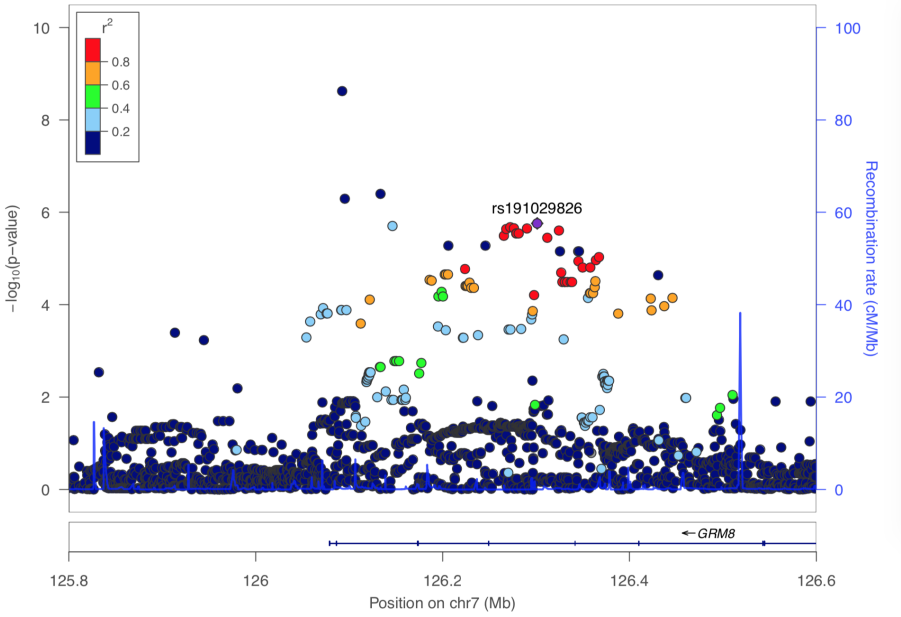
resentative SNP rs76326523, which is in total LD to rs191029826, could be observed, but is mostly not considered additionally, here, as corresponding results are identical.

3.2.5 | Evaluation of lead SNPs identifies additional tag SNPs

The lead SNPs (rs12537032 and rs191029826/rs76326523) were further fine-mapped and evaluated regarding their meta-analysis results. As the locus was not covered by TCGA-PRAD, the results relied on the $PCGA^{LMU}$ and ICGC-CA cohorts only. Despite the genome-wide significance, the heterogeneity for the lead SNPs was greater than 50% (Figure 3.8a and 3.9a). Therefore, a second representative tag SNP was chosen for each. Both representative tag SNPs were in LD ($r^2 > 0.4$, Figures 3.7a and 3.7b) with their lead SNP, respectively, had a lower heterogeneity than 50% and were the most significant SNP of those fulfilling the previous conditions. This led to the additional tag SNP rs74999840 for lead SNP rs12537032 and rs73451279 representing the second signal from rs191029826/rs76326523. Both representative tag SNPs possesed a heterogeneity $I^2 \leq 5\%$ and were significantly associated ($P = 7.1 \times 10^{-5}$) with GG. The five tag SNPs and the study results obtained for them were evaluated with forest plots shown in Figures 3.8a, 3.8b, 3.9a, and 3.9b. For all of them their estimated effect of the single GWAS (black square) was positive and consistent among the studies. Thus, their pooled results from the meta-analysis (black diamond) were distant from the no effect line. This indicates that for all five tag SNPs, the minor allele was significantly associated with higher GG. Due to the smaller number of PCa patients in the ICGC-CA cohort, its 95% confidence interval (CI; Figure 3.9a and 3.9b, horizontal grey line) crossed the no effect line implying no significant study effect. However, as the results pooled in the meta-analysis were nevertheless significant and only a few SNPs could be identified as potential tag SNPs, the corresponding tag SNPs (rs191029826, rs76326523, rs73451279) were not rejected but, as an exception, further investigated in this thesis. As causality could not be confirmed for any of them, all five SNPs were referred to as tag SNPs.

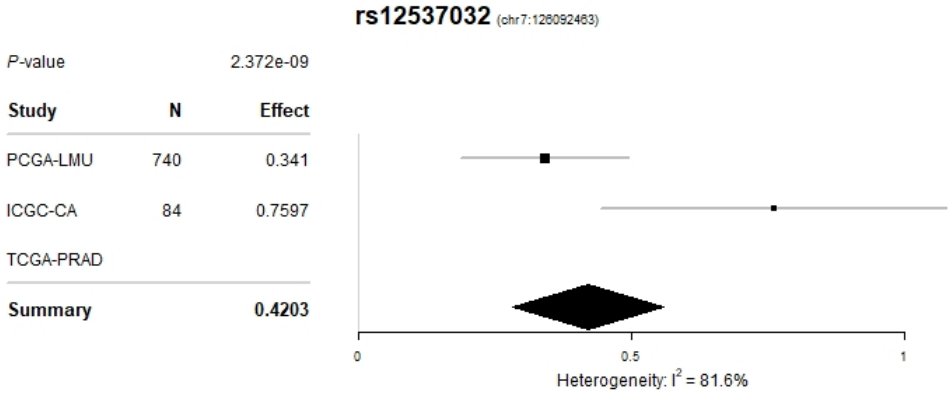


(a) Lead SNP rs12537032

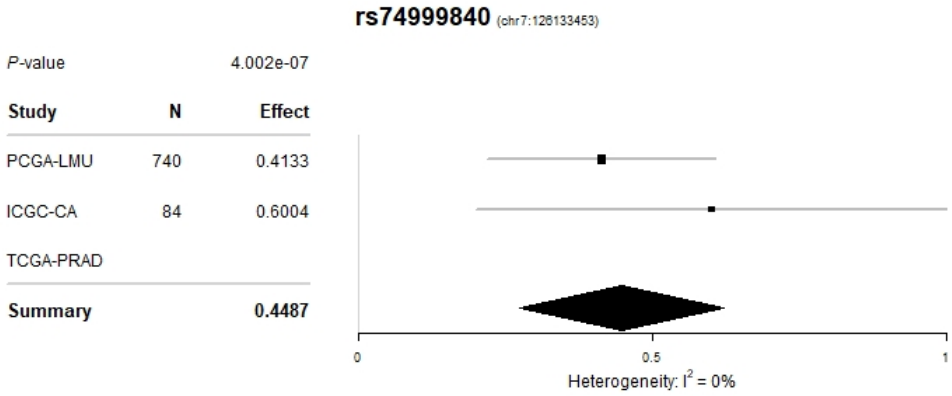


(b) Lead SNP rs191029826

Figure 3.7: Local association results from meta-analysis focusing on the haploblock of the lead SNPs. Variants were colored according to their LD (r^2) to the lead SNP a) rs12537032 and b) rs191029826, both highlighted in purple.

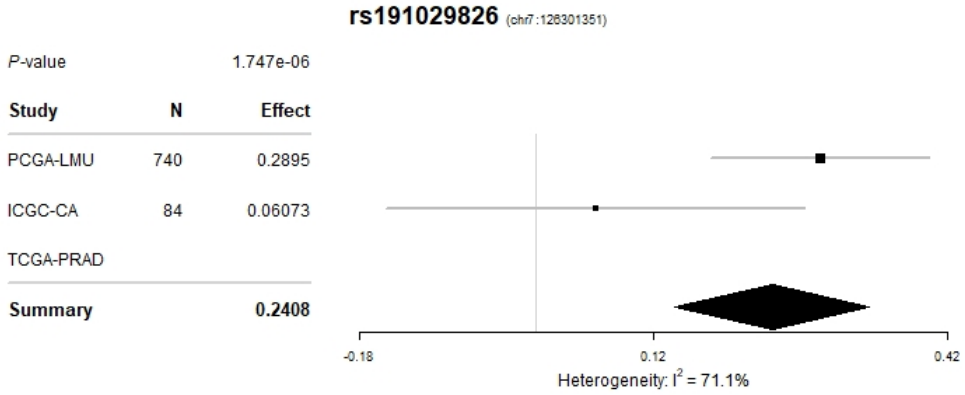


(a) Lead SNP rs12537032

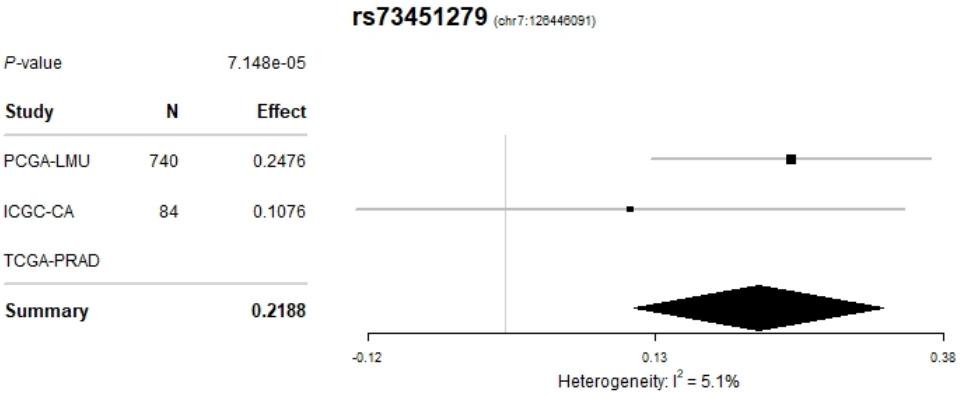


(b) Tag SNP rs74999840

Figure 3.8: Forrest plots illustrating the genome-wide significant results of the meta-analysis and the single GWAS it relied on in detail for a) lead SNP rs12537032 and its b) representative tag SNP rs749998040 on locus 7q31.33. The vertical line located at 0 represents the point of no effect. The effect size is marked by a black square framed by its corresponding 95 % confidence interval (grey). Results from TCGA-PRAD are not shown as 7q31.33 was not covered by the study.



(a) Lead SNP rs191029826/rs76326523



(b) Tag SNP rs73451279

Figure 3.9: Forrest plots illustrating the results of the meta analysis and the GWAS it relied on in detail for a) lead SNP rs191029826/rs76326523 and its b) representative tag SNP rs73451279 regarding the second identified signal on locus 7q31.33. The vertical line located at 0 represents the point of no effect. The effect size is marked by a black square framed by its corresponding 95 % confidence interval (grey). Results from TCGA-PRAD are not shown as 7q31.33 was not covered by the study.

3.2.6 | Minor allele of tag SNPs can increase risk for aggressive PCa up to three times

Next, an examination of the alleles from $PCGA^{LMU}$ samples revealed, that the risk allele (RA) referred to each tag SNP's minor allele. This became apparent by the positive linear regression coefficient resulting from the GWAS, which represents the effect β of the risk allele (Table 3.2). This effect was transformed to the OR, which was calculated from GGG I/II to both other groups (III and IV/V) separately.

For patients with the risk allele T for rs12537032 (β = 0.34, CI[0.19-0.49]), the risk of developing a tumor of grade III was increased (OR(III) = 1.746). Moreover, their risk for a tumor entering stage IV/V was further increased to an OR of 2.640 (Table 3.2). For rs74999840 (β = 0.41, CI[0.22-0.61]) the risk for PCa stage III as well as IV/V tripled (OR(III) = 3.088, OR(IV/V) = 3.466). In contrast to rs12537032, the rising risk of developing a GG of III or IV/V for patients carrying the risk allele at rs74999840 did only slightly differ (OR(III) vs. OR(IV/V), $\Delta OR \sim 0.5$, Table 3.2).

The other lead SNPs rs191029826 and rs76326523, respectively, which were not in LD with rs12537032, had both the same allele frequency and effect (β = 0.29, CI[0.18-0.40]) on PCa progression. Patients with the corresponding risk allele of both SNPs (Table 3.2) had a higher risk of PCa with GG III (OR(III) = 1.826). Moreover, their risk for a tumor entering stage IV/V was more than twofold higher (OR(IV/V) = 2.212) than for patients without the risk allele. PCa patients with minor allele G for the representative tag SNP of the former two rs73451279 (β = 0.25, CI[0.13-0.37]), had indeed an increased risk for aggressive PCa (OR(III) = 1.376). For patients harboring the risk allele, the chance to develop PCa of GG IV/V doubled (Table 3.2).

These results indicate that Central Europeans with the risk allele in their genotype of the identified tag SNPs (rs12537032, rs74999840, rs191029826, rs76326523, rs723451279) may have a two to threefold higher risk to develop an aggressive PCa of GG III, IV or V.

Table 3.2: Summary of GWAS results based on $PCGA^{LMU}$ data against GG including the risk of developing aggressive PCa. Global MAF was taken from dbSNP [77]. (RA = risk allele, nRA = non-risk allele, # = number of , RAF = risk allele frequency, GGG = Gleason grade group, OR = odds ratio)

SNP	Alleles (A/a)	RA	GGG	# nRA	# RA	RAF	OR to baseline	Effect β (CI 95%)	GWAS P-value	Global MAF
rs12537032	C/T	T	I/II	684	38	0.053	1.000			
			III	299	29	0.088	1.746			
			IV/V	375	55	0.128	2.640	0.34		
			total	1358	122	0.082		(0.19-0.49)	1.47×10^{-5}	0.02
rs74999840	C/T	T	I/II	704	18	0.025	1.000			
			III	304	24	0.073	3.088			
			IV/V	395	35	0.081	3.466	0.41		
			total	1403	77	0.052		(0.22-0.61)	2.95×10^{-5}	0.012
rs191029826/	C/A	A	I/II	633	89	0.123	1.000			
rs76326523	C/T	T	III	261	67	0.204	1.826			
			IV/V	328	102	0.237	2.212	0.29		
			total	1222	258	0.174		(0.18-0.40)	4.35×10^{-7}	0.077
rs73451279	T/G	G	I/II	642	80	0.111	1.000			
			III	280	48	0.146	1.376			
			IV/V	346	84	0.195	1.948	0.25		
			total	1268	212	0.143		(0.13-0.37)	6.83×10^{-5}	0.074

3.2.7 | Increased risk allele frequency in Central Europeans may be connected to higher number of incidences

Beside estimating the potential effect of the risk allele from RAF for Central Europeans from the $PCGA^{LMU}$ cohort, the corresponding MAF of the tag SNPs was scrutinized, which was here equivalent with RAF (Table 3.2). A difference between MAFs of Central European patients and PCa patients from an international cohort of the 1000 Genomes dataset, as provided in the dbSNP database [77], may partially explain the higher incidence rate for PCa in European men compared to men with other ethnicity [76].

For all five tag SNPs, MAF of Central Europen men was much higher than its frequency in a global cohort. The frequency for rs191029826, rs76326523, and rs73451279 with MAF > 14% was twice as high for Central Europeans compared to PCa patients worldwide (MAF > 7%, Table 3.2). Moreover, with a MAF of 8% the risk allele of lead SNP rs12537032 was increased fourfold in Central European samples of the $PCGA^{LMU}$ cohort compared to the worldwide MAF of 2%. Furthermore, its representative SNP, rs74999840 (MAF = 5%) was even five times higher than its worldwide MAF (1%).

These numbers show that risk alleles of the identified tag SNPs are highly overrepresented in Central European PCa patients compared to their frequency worldwide.

3.2.8 | Genotype of rs73451279 correlates with *GRM8* expression

All five tag SNPs (rs12537032, rs74999840, rs191029826, rs76326523, and rs73451279), which are located within different introns of *GRM8*, were inspected for cis and trans eQTL genes with SNIPA. However, beside *GRM8* no other eQTL gene was found. Thus, an eQTL analysis was performed to examine the potential effect of the tag SNPs on *GRM8* expression. For rs12537032, rs74999840, rs191029826, and rs76326523 eQTL analysis was not significant ($P > 0.39$; Figure A.7). However, for rs 73451279 a significant correlation ($P = 0.009$) between the genotype of rs73451279 and *GRM8* expression could be observed. For PCa patients harbouring the risk allele G, decreased gene expression levels could be seen (Figure 3.10).

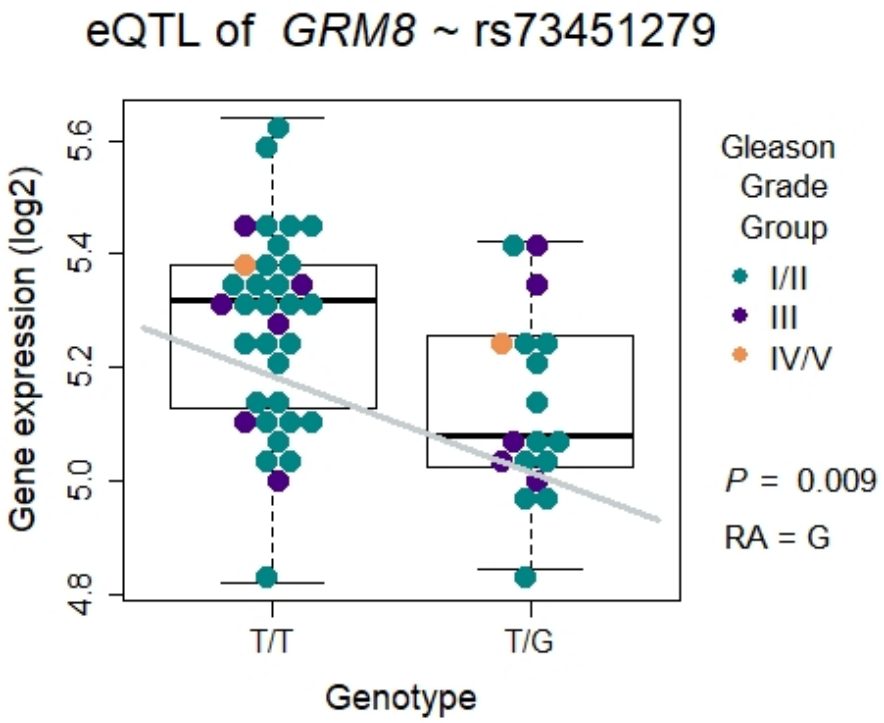


Figure 3.10: Result of eQTL analysis of rs73451279 against *GRM8* based on data from the ICGC-CA cohort. Effect size between both genotypes is represented by a gray line. (RA = risk allele)

This indicates that PCa patients with a risk allele G instead of a homogenous genotype T/T in rs723451279 have lower *GRM8* expression and are more likely to develop an aggressive tumor embodied by a high GG.

3.2.9 | *GRM8* expression is not associated with relapse in PCa

After identifying an association between rs73451279 with *GRM8* expression, its expression levels were compared regarding BCR. With Kaplan-Meier method, the BFS curves of patients split regarding their gene expression of *GRM8* were compared. But, it did not reveal any significant association ($P > 0.2$) with BCR in any of three different cohorts (ICGC-CA, GSE16560, and GSE46691), as shown in Figure A.8.

3.2.10 | Genotype of tag SNPs has no prognostic effect on BCR-free survival

To examine, whether a patients' genotype affects PCa progression such as BCR, a survival analysis was performed on samples from the ICGC-CA cohort, stratified by their genotype of each tag SNP. However, neither of the tag SNPs (rs12537032, rs74999840, rs191029826, and rs73451279) was significantly associated ($P > 0.2$) with BFS (Figure A.9). This indicates, that the genotype of the identified tag SNPs are not suitable as prognostic biomarkers for BCR of PCa.

3.2.11 | Tag SNPs are located in epigenetic inactive regions of 7q31.33

In general, *GRM8* is rather low expressed in prostate cells [157, 158]. Nevertheless, the epigentic background of *GRM8* in PCa cells was scrutinized, to verify a potential effect of the identified tag SNPs on regulatory mechanisms in *GRM8*, which might affect *GRM8* expresssion. Investigation of ChIP-seq data of histone modifications (H3K36me3 and H3K9me3) in PCa cells (PC-3, Figure 3.11), showed that this region was barely transcribed, which is in line with the observed low expression of *GRM8* [159]. This observation was also consistant with *GRM8* expression levels in PCa seen on GTEx portal [157, 158]. Also, H3K27ac did not show any enrichment for active enhancers in the region around *GRM8*. DNase activity, which displays chromatin accessibility [159], could only be observed slightly. For H2AFZ, which makes DNA more accessible to transcription factors and can mark active promoter regions [160], no enrichment could be seen in proximity of the tag SNPs. Only for CTCF one peak could be detected near rs73451279 (Figure 3.11), but was with a distance of 2.5 kb still too distant to be likely affected by the variant. Due to variant annotation and the location of the tag SNPs in introns at the end of *GRM8*, an effect on the promoter or transcription start site could be excluded.

These results, show that due to their location in epigenetic inactive regions, none of the tag SNPs is involved in regulatory mechanisms, as can be seen from the available data, so far.

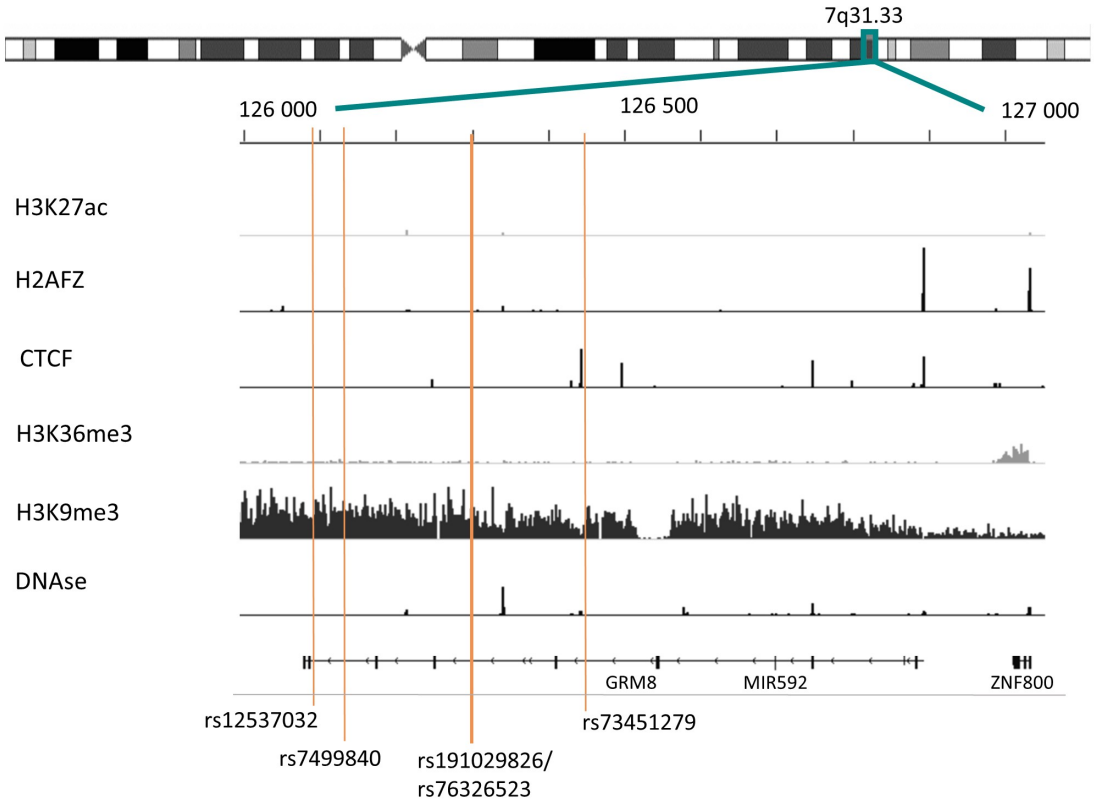


Figure 3.11: Epigenetic profile of the genomic region around *GRM8* (chr 7, 126 mb - 127 mb) on 7q31.33 locus showing the aligned reads of H3K27ac, H2AFZ, CTCF, H3K36me3, H3K9me3, and DNAse from PC-3 cell lines. Identified tag SNPs are marked by orange lines. Altogether, the epigenetic profile describes a rather inactive genomic region.

3.2.12 | Potential functional variant identified on rs73451279

With conditional association testing on the eQTL SNP rs73451279, based on data from $PCGA^{LMU}$ a significantly reduced association signal for variants in LD with rs73451279 could be observed (Figure 3.12), implying a major functional variant at the 7q31.33 locus. However, the observed association did not completely disappear, which can be explained by other linked variants having a regulatory effect on the expression level of *GRM8*.

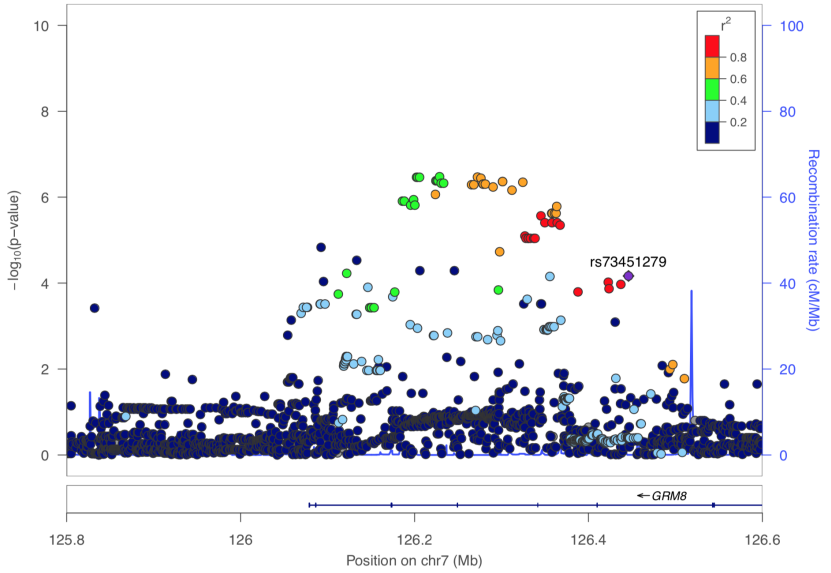
3.2.13 | rs73451279 may affect nonsense-mediated RNA decay

Using the VEP on rs73451279 to annotate its potential effect, revealed an effect of risk allele G on two *GRM8* transcripts associated with nonsense-mediated mRNA decay (NMD). The first transcript GRM8-211 (ENST00000472701.5; Figure 3.13a) is build of 12 exons, while the other transcript GRM8-202 (ENST00000341617.7; Figure 3.13b) consists of 11 exons. Even though their number of exons differes only by one, the difference was not caused by one missing exon, but a different composition and order of exons (Figure 3.13). However, the exon responsible for this effect seemed to be the same in both transcripts.

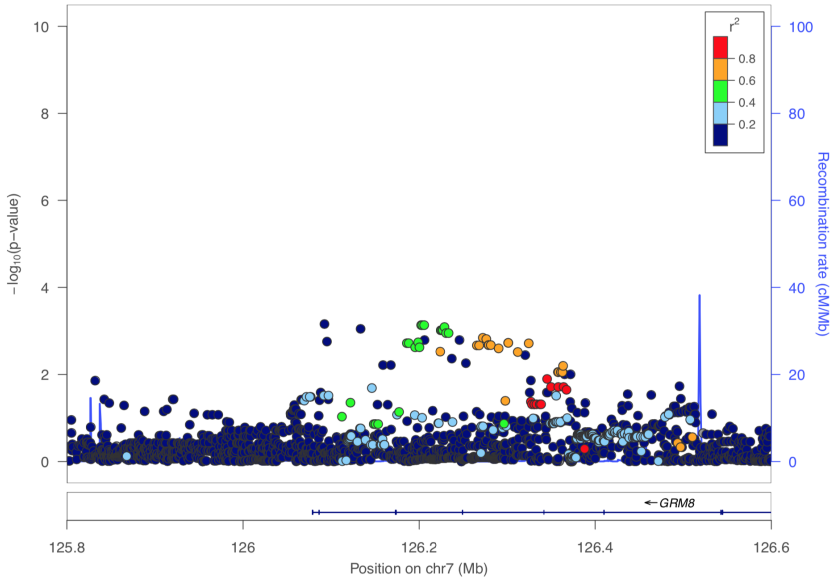
This result indicates that the risk allele of rs73452379 might activate NMD of *GRM8*.

3.2.14 | Availability and publication of GWAS results

Due to ethical reasons and patients privacy rights it was not possible to openly publish the genomic data from the $PCGA^{LMU}$. But to support PCa research of other scientists, the results of the GWAS based on $PCGA^{LMU}$ were made publicly available, as any inference to the tested individuals or their genome is not possible. Therefore, a webservice was implemented enabling the user to screen or specifically search for germline variants that might be associated to GG or tumor encapsulation in PCa patients. STARLING, which stands for 'pro**STA**te cancer **R**esearch **L**everaging **I**mportant **N**ovel **G**enomic biomarkers', can be accessed online via www.starling-pcga.med.lmu.de.

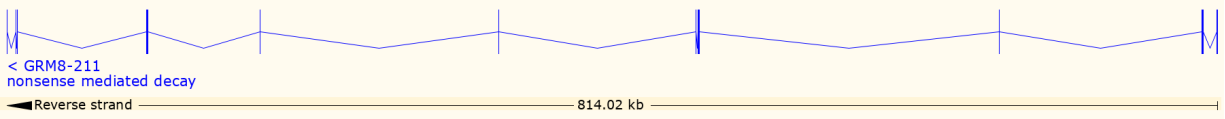


(a) Classic GWAS

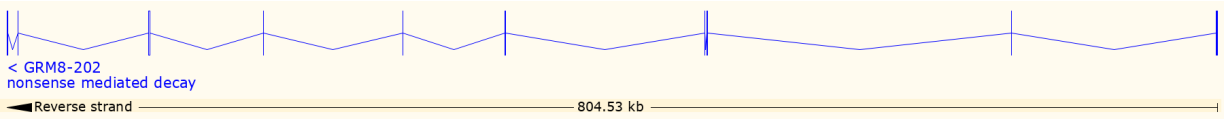


(b) Conditional GWAS regarding rs73451279

Figure 3.12: Local association results from GWAS against GGG on PCGA^{LMU} data focusing on the haploblock of tag SNP rs73451279 showing a significant difference between a) classical GWAS without condition and b) conditional analysis. Variants are colored according to their LD (r^2) to the a) tag and b) conditional SNP rs73451279 (purple), respectively.



(a) ENST00000472701.5



(b) ENST00000341617.7

Figure 3.13: *GRM8* transcripts a) ENST00000472701.5 and b) ENST00000341617.7 associated with NMD.

4

Discussion

This thesis addresses whether germline variation and somatic mutations in PCa patients can conduce to the development of aggressive tumors. With the potential biomarkers found here, predictions for improved therapy and fundamental PCa research can be supported.

This thesis reports the results of a GWAS meta-analysis from imputed PCa cohorts ($PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA), which were tested against their GGG (I/II vs. III vs. IV/V) to show a potential effect of germline variation on PCa aggressiveness. A genome-wide significant signal on locus 7q31.33 was identified that was triggered by rs12537032, an intronic variant of *GRM8*. This lead SNP revealed a second LD independent signal on the same haploblock driven by rs191029826 and rs76326523. Moreover, both were in LD with the eQTL SNP rs73451279, whose risk allele G was not only linked to higher GG but was also associated with low *GRM8* expression.

Interestingly, the discovered risk locus 7q31.33 is part of a previously discovered chromosome location 7q31-33 that was associated with aggressive PCa in German men [82]. The risk alleles of all identified tag SNPs were largely increased in Europeans compared to global allele frequencies. The minor allele for all five tag SNPs (rs12537032, rs74999840, rs191029826, rs76326523, and rs73451279) was also the risk allele multiplying the risk of an aggressive, high-grade PCa by two up to three times. Surprisingly, this applied to a European population with high minor allele frequencies of tag SNPs

but fewer incidences of aggressive tumors than in African and African American men, for whom, in contrast, the minor allele frequencies were only a small fraction. Due to different allele frequencies and the resultant incidences of aggressive tumors among ethnicities [76, 77], this risk locus may evolve to malignancy represented by or even restricted to European men.

The Glutamate Metabotropic Receptor 8 gene (*GRM8*) is located on the 7q31.33 locus [161]. Its gene product is activated by an excitatory neurotransmitter in the central nervous system. This glutamatergic neurotransmission plays a key role in brain function [161, 162]. Based on putative signal transduction mechanisms and pharmacologic properties, metabotropic glutamate receptors (mGluR) are split into three groups. Along with *GRM4*, *GRM6*, and *GRM7*, *GRM8* is part of mGluR Group III, which is connected to the inhibition of cyclic adenosine monophosphate (cAMP) cascade [161, 162]. cAMP is a well-known secondary messenger that activates and interacts with diverse proteins and kinases [163]. In PCa cells (PC-3, LNCaP), increased cAMP led to tumor growth inhibtion [164, 165]. Furthermore, the cAMP dependent protein kinase A (PKA) mediates cell proliferation and differentiation, especially in cancers. Both cAMP and PKA participate in carcinogenesis and progression of PCa [163]. cAMP and PKA signaling have been shown to regulate the activation of androgen receptor in PCa [163], which plays a significant role in PCa development [166].

All mGluRs engage in the glutamatergic system, which is the main excitatory neurotransmission system [162]. In PCa cell lines, Pissimissis *et al.* verified that mGluRs of the glutamatergic system had a potential regulatory effect on PCa. While some mGluRs (*GRM1*, *GRM2*, *GRM3*, *GRM4*, *GRM5*) were similarly expressed in both cell lines (PC-3, LNCap), *GRM8* was differentially expressed [167]. The results of this thesis suggest, that this could be explained by different GG of PCa from which the cell lines derived. Other studies [168–170] have linked *GRM1* with the Gleason score and PCa aggressiveness, which was caused by a mutation altering its splicing process. However, no studies examining a similar connection specifically between *GRM8* and PCa aggressiveness have been found.

Based on their location in the introns of *GRM8*, the identified variants might be a tag or even causal SNP affecting *GRM8* expression. Indeed, the risk allele G of tag SNP rs73451279 was significantly associated with lower *GRM8* expression, which might be

involved in developing aggressive PCa defined by high GG.

Based on this, the epigenetic profile of the variant was examined, but no regulatory effect could be observed in PCa cell line influencing the expression level of *GRM8*. In general, *GRM8* is mostly expressed in the brain and the testis, but only slightly expressed in prostate tissues [157,158]. This observation could be verified in its epigenetic profile (Figure 3.11), where an increased level of H3K9me and a decreased H3K36me3 ChIP-seq signal in PC-3 cell line indicated a restricted chromosome accessibility affecting *GRM8* expression in PCa. Also, no active regulatory mechanisms in this area influencing the GRM8 expression could be detected. However, it remains to be determined whether the candidate tag SNPs affect binding motifs of transcription factors.

Variant effect prediction revealed, that in two transcripts of *GRM8*, rs73451279 was associated with nonsense-mediated decay (NMD). NMD is a specific surveillance mechanism that recognizes defective mRNA arising from erroneous alternative splicing and prompts its degradation [171]. Germline variation of rs73451279 might affect alternative splicing, leading to a premature translation stop codon in the pre-mRNA, which provokes NMD [171]. Therefore, increased degradation of pre-mRNA could have reduced *GRM8* transcripts. This was reflected by reduced *GRM8* expression levels triggered by rs73451279 specifically for patients carrying the risk allele associated with more aggressive tumors. So far, alternative splice variants of *GRM8* have not been fully described or analyzed [161]. Whether this observation was only an association with higher GG in PCa patients or was the reason to develop an aggressive tumor remains to be determined. Further wet lab research to determine the underlying mechanism accurately is necessary. Nevertheless, the identified tag SNPs (rs12537032, rs74999840, rs191029826, rs76326523) and SNP rs73451279 might be potential biomarkers for developing PCa with higher GG.

Neither the genotype of the identified tag SNPs nor the gene expression of *GRM8* was found to be associated with worse disease outcome. This finding disputes the hypothesis that rs73451279 induces NMD and leads to decreased *GRM8* expression and, therefore, affected mechanisms that compound the development of aggressive PCa with high GG. However, the survival analysis could only be conducted on a small sample size of the cohorts, which puts its validity into question. A more representative study size might support this hypothesis, so a repetition of the survival analysis with more pa-

tients is highly recommend.

When repressing the signal from the identified eQTL SNP rs73451279 on the 7q31.33 locus by reapplying GWAS conditionally, the signal strength of the remaining variants was indeed not dissolved but significantly reduced. This implies a potential functional role of rs73451279 located on chromosome 7. However, the remaining signal also indicates that apart from rs73451279 another, yet-undetected regulatory effect caused by further SNPs may be involved affecting *GRM8* expression and thus PCa progression.

After focusing on the potential impact of germline variations on PCa progression, somatic mutations in PCa were investigated, particularly the T2E fusion, which characterizes its own molecular subtype [53,58]. It seems likely that PCa with and without the fusion gene may be driven by distinct pathways and develop via differentiated pathogenesis. Apart from this hypothesis, specific differentially expressed genes were more suitable to predict the development of metastases in PCa for one molecular subtype than the other. Under this aspect, transcriptomic data with regard to T2E fusion were explored to identify potential subtype specific biomarkers that might enable PCa outcome prediction.

Based on transcriptomic data of two large cohorts (TCGA-PRAD and GSE46691) and two validation cohorts (GSE16560 and the TMA-cohort), the molecular differences in PCa and their potential impact were explored. After screening the patients' intratumoral gene expression levels, which were stratified by their T2E-status and presence/absence of metastasis, the top 20 metastasis associated gene-signatures were identified for T2E-positive and -negative PCa, each, via enrichment analysis. Interestingly, the gene-signatures resulting from T2E-positive based data did not overlap with those based on T2E-negative data. This was in line with prior studies [53, 58] implying that T2E-positive and -negative PCa are distinct molecular subtypes whose disease progressions evolve disparately. From these subtype specific gene-signatures the most frequent genes (topGL-pos and -neg) were extracted each. Five genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) that were overrepresented exclusively in T2E-negative PCa turned out to be valuable for subtype specific outcome prediction. These findings show the importance of considering subtype specific biomarkers for risk prediction in PCa to improve their prognostic capability.

Both, Asporin (*ASPN*) and Biglycan (*BGN*) [161] are known for their association with PCa progression [172] as well as poor prognosis [173]. Indeed, these observations endorsed the results here, but apply only to T2E-negative cases. No indication of prognostic value for T2E-positive cases appeared. Moreover, a relation between BGN expression and T2E fusion was reported by Jacobsen *et al.* [173]. The results presented in this thesis could not confirm this observation in T2E-positive cases, as *BGN* was not among the genes of the top gene-signatures associated with metastasis. In contrast, it applied to T2E-negative PCa.

COL1A1 (Collagen Type I Alpha 1) encodes a component of the collagen protein (type 1), which is a supportive and major constituent of connective tissues [161]. Several studies [174–176] have declared *COL1A1* as an oncogene whose high expression is associated with progression in several tumor types and have introduced the gene as potential biomarker. Nevertheless, a potential connection of *COL1A1* and worse outcome in PCa has not yet been reported. This proposes *COL1A1* as novel potential biomarker for T2E-negative PCa.

The protein product of *RRM2* (Ribonuclease Reductase M2) is a subunit of the ribonuclease reductase that plays a crucial role in DNA synthesis [161]. Its overexpression is known to promote tumor progression [177]. Indeed, several studies [178–180] have confirmed the association of *RRM2* overexpression with PCa progression and worse outcome, but no study has distinguished between molecular PCa subtypes. The study results obtained from mRNA and protein level based on four independent PCa cohorts, here, coincided with the findings of these studies, but differentiated by the refinement that the strong prognostic power of *RRM2* in T2E-negative PCa did not apply to T2E-positive cases. One of these studies, however, conducted by Mazzu *et al.*, additionally examined transcriptomic changes of *RRM2* in the T2E-negative cell line PC-3 indicating *RRM2* overexpression to be an oncogenic trait [179, 181]. Interestingly, they partially confirmed the hypothesis reported by this thesis that transcriptomes of PCa subtype specific genes are differentially affected by regulatory mechanisms. When analyzing the epigenetic background of the candidate gene *RRM2* in PCa cells, the scientists detected a potential binding region for the transcription factor FOXM1 in the promotor region of *RRM2*, which actives transcription [179].

In the results of this thesis, similar observations could be seen for *TYMS* (thymidyl

synthetase), which is involved in DNA repair and replication [161]. Burdelski *et al.* reported a correlation between *TYMS* and worse outcome in PCa [182]. Likewise, the results disclosed here, show a significantly higher risk for short EFS for T2E-negative patients expressing high levels of *TYMS*. This observed effect was completely absent in T2E-positive PCa.

Pathway analysis was repeated, but this time focusing only on T2E-negative PCa and stratified for cases with high and low expression of the previously identified subtype specific candidate genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*). Different gene-signatures were identified for subgroups defined by high and low expression for each gene. With regard to the top 20 gene-signatures of the respective subgroups, a limited average overlap of only 38% was observed. This observed effect indicates that tumors expressing different levels of the identified marker genes differ in their (patho)biology.

So far, the GGG is a common clinicopathological marker for PCa risk-prediction. Here, the GGG outperformed the identified subtype specifc biomarkers in an comparative survival analysis in both T2E-positive and -negative cases. However, in T2E-negative PCa, the combination of both clinicopathologic and two gene markers (*RRM2* and *TYMS*) could improve outcome prediction. This advantage induced by two of the identified subtype specific genes emerged exclusively in T2E-negative PCa. For evaluating the other three candidate genes (*ASPN*, *BGN*, and *COL1A1*) regarding their additional prognostic value with GGG, larger cohorts are necessary. Yet, the availability of suitable anti-RRM2, anti-TYMS, and anti-ERG antibodies should enable a rapid translation of our findings to the clinic through the detection of the T2E-status, the *RRM2*, and *TYMS* expression levels by IHC in conjunction with GG on routine histology [105].

Beside T2E-positive PCa, other molecular subtypes can be found in PCa. Not only rare ETS translocations, but also mutations in presumable cancer driver genes like *SPOP*, *FOXA1*, and IDH1 [53] are characterisic of a few PCa patients. The second most frequent mutated gene in PCa with an observed frequency of around 10% is *SPOP* [53, 153]. Nevertheless, it had no impact on the validity of the identified candidate genes (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) for outcome prediction. Even without these *SPOP*-mutated cases, the subtype specific genes remain significant prognostic markers for T2E-negative cases. For the other, less frequent mutations occurring in *FOXA1* and *IDH1*, which were present in 1.7% and 3% of cases in the TCGA-PRAD

cohort, respectively, an impact on biomarker prediction was not comprehensible.

Furthermore, common cancer-driving mutations in tumor oncogenes, such as *TP53* and *PTEN*, which are known for their enrichment in PCa [153], were investigated for their influence on the results presented in this thesis. In the TCGA-PRAD cohort, 7% of the PCa patients possessed *TP53* mutations that were equally distributed between T2E-positive and -negative cases. Their absence in the analysis did not alter the previous results, which eliminates an alleged bias. With an overall frequency of only 2% in the TCGA-PRAD cohort, the number of PCa cases harboring a *PTEN* mutation was far too low to falsify the results.

Multiple predictive tests for PCa patients have been developed, but have not yet entirely outperformed the standard diagnostic and histological parameters, such as PSA and the Gleason score, yet.

Based on a discovery cohorts used in this thesis (GSE46691) the transcriptome based Decipher Prostate Cancer Test was recently developed [25]. With transcriptomic profiling of 22 PCa associated genes, the authors created a genomic classifier for the Decipher test, which allows risk-stratification of PCa patients after surgery [25,27]. Multiple clinical studies have certified the test [26, 28, 29]. Interestingly, the 22 Decipher genes did not encompass any of the subtype specific biomarkers identified here. Among the identified candidate genes, the absence of *RRM2* in the genomic classifier is especially surprising, as an association of *RRM2* overexpression with aggressive PCa was reported by Kosari *et al.* [180], whose co-authors from the Mayo Clinic were later involved in the development of the genomic classifier of the Decipher test [25]. However, the Decipher test does not discriminate between molecular PCa subtypes defined by T2E fusion, which could explain both the divergence of genes and the absence of *RRM2* among the Decipher genes.

Other genomic tests also do not factor the patients' T2E-status in their procedure [31, 34]. Contrary to Decipher, however, both Oncotype Dx [34] and Prolaris [31] tests had a concordance between their utilized biomarkers and the identified T2E-negative specific genes. Likewise, both Oncotype Dx and Prolaris do not differ between molceular PCa subtypes [31, 34]. Prolaris tests 31 gene transcripts, including *RRM2* [31]. Among the 17 markers used by the Oncotype Dx test [34] are both *BGN* and *COL1A1*. Checking the candidate genes against the predictive markers of both genomic tests,

however, was impossible. A fraction of their markers, which were necessary for this comparison, were missing in the unity of variably expressed genes used to discover the subtype specific candidate genes. 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 [105].

In contrast to the three genomic tests mentioned above, the EPI test respects the T2E fusion of PCa by considering *ERG* expression. Besides *ERG*, transcriptomic levels of two more genes (*PCA3* and *SPDEF*) are used to predict high-grade PCa [37, 38]. Both genes were not among the subtype specific genes identified in this thesis. However, it is important to take the fact into account that the EPI test uses gene expression from exosomes extracted from urine, but not from mRNA obtained from primary tumor. Diverse sources of gene expression can lead to divergent results. An initial study [39] validating the EPI test was promising, but it remains to be seen whether the test prevails in clinical routines. Nevertheless, this supports the findings of this thesis regarding the importance of the T2E-status in predicting high-grade PCa leading to worse outcome. Moreover, it shows that a PCa patient's T2E-status and, accordingly, the *ERG* expression level are critical factors independent of the sample tissue, which should not be discarded.

5

Conclusion, limitations, and perspective

5.1 | Conclusion and perspective

This thesis exemplified integration of comprehensive transcriptomic and genetic data with clinical records of PCa patients to emphasize both the importance of considering a patient's T2E-status in prognostic biomarker based PCa risk prediction and the effect of germline variation at potential risk loci on PCa aggressiveness.

Based on multiple transcriptomic cohorts, this study confirmed previous findings that T2E-negative and T2E-positive PCa are distinct molecular subtypes most likely induced by divers pathways. From this, five subtype specific prognostic biomarkers (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) were identified for T2E-negative PCa exclusively, whose overexpression promoted worse outcome. Even though these novel potential biomarkers did not outperform current prognostic biomarkers, they could enhance them. As affirmed by several prior studies, specifically *RRM2* stood out as promising potential transcriptomic biomarker, particularly regarding T2E-negative PCa.

After all, it remains to be determined whether factors other than the T2E-status affect the differential expression of the identified potential biomarkers *ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*. Exploring their epigenetic background may expose underlying mechanisms of progression in T2E-negative PCa. Further computational and ex-

perimental investigations focusing on T2E-negative subtypes are advisable. The results of the transcriptome analysis described in this thesis were published in an international peer-reviewed journal [105].

GWAS derived from three PCa cohorts ($PCGA^{LMU}$, TCGA-PRAD, and ICGC-CA) were combined via meta-analysis and identified five potential tag SNPs (rs12537032, rs74999840, rs191029826, rs76326523, and rs73451279) located at the PCa risk locus 7q31.33, which is associated with aggressive PCa. The intronic SNPs of *GRM8* may affect its role in the cAMP cascade and the affiliated PKA, which are known to be involved in carcinogenesis. In fact, no epigentic acitivity could be detected near *GRM8*. However, the risk allele G of rs73451279 – which was associated with an almost doubled risk of developing an aggressive PCa (high GG) – was predicted to provoke NMD resulting in decreased *GRM8* expression. Even though rs73451279 was shown to be a potential functional variant on the 7q31.33 locus, it is probably not the only variant influencing *GRM8* expression. In particular, no association of the tag SNPs genotype or *GRM8* expression with worse outcome could be verified.

In the end, of all five tag SNPs, rs73451279 was the most promising variant to qualify as potential biomarker for PCa aggressiveness, but has to be contemplated cautiously, as the validity of some results was low. Nevertheless, these results indicate a potential genomic biomarker that should be further investigated with increased number of cohorts and more comprehensive transcriptomic, genomic, and survival data. Additional research may reveal its outright function as well as the role of *GRM8* in the mechanisms provoking tumor malignancy.

5.2 | Limitations and approaches

This thesis assessed whether somatic mutations and germline variation drove PCa progression. Several potential genomic and subtype specific transcriptomic biomarkers associated with aggressive PCa were detected. But resources and time were limited so some aspects could not be examined. Limitations, complications, and open questions that arose during these projects should be addressed in continuing studies in the future, were reflected here.

With 800 participants, the PCGA^{LMU} had a sufficient number of participants for valid results. In the results of the GWAS, compelling trends of significant effects in germline variants could be observed. However, to detect common SNPs with genome-wide significance, more patient samples are needed. To increase study power, the cohort will soon be extended up to 2,000 participants. Even more important aspects are the transcriptomic data and genomic information about gene fusions, which were only extracted from a fraction of participants so far. Enlarging the cohort regarding gene expression or T2E fusion status for the entire cohort immensely enhances the predictive power and scientific value of the cohort. Amplifying the PCGA^{LMU} based GWAS with eQTL analysis or subtype specific information would be a great benefit and add new potential to the current results.

Continuing the survey of the PCGA^{LMU} participants for a certain time period would enable a follow up study on survival data, which would add a valuable component to this study. Due to the pseudonymisation and the participants reliability, however, follow up studies are challenging.

Especially for the eQTL and survival analysis regarding the candidate tag SNPs, the number of available samples played an important role as the required clinical and transcriptomic data were only available for a fraction of the samples in the ICGC-CA cohort. Therefore, the results of both analyses should be considered cautiously and are recommended to be repeated with more samples for higher validity.

The meta-analysis was performed on three different PCa cohorts, of which one only comprised WXS data. Therefore, intronic variants could only be obtained from two cohorts. Moreover, the heterogeneity was too high for both identified lead SNPs and many other variants. Increasing the number of specifically WGS PCa cohorts would improve the validity of the results. However, the number of accessible WGS PCa cohorts with comprehensive clinical, survival, and transcriptomic data as well as an adequate number of participants is limited. Surveying the academic field for proper PCa cohorts may enable a similarly constructed follow up study.

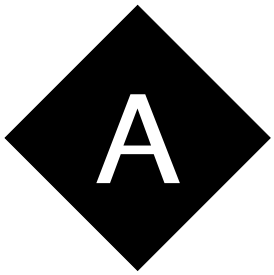
Five potential transcriptomic biomarkers (*ASPN*, *BGN*, *COL1A1*, *RRM2*, and *TYMS*) were identified specifically for T2E-negative PCa. An intensive study of the epigenetic background of these genes may illuminate the underlying regulatory mechanisms of PCa progression. Comparing epigenetic profiles in the areas near the transcriptomic

biomarkers from T2E-negative against T2E-positive PCa could also shed light on potential regulatory mechanisms influencing both PCa subtypes differentially regarding their PCa aggressiveness. Potential *ERG*-binding sites or enhancers in proximity of these five genes should also be examined. However, Unfortunately, the ChIP-seq data required for these suggested investigations were not available for this study.

For the *GRM8* gene harboring significant tag SNPs in its intronic regions, the epigenetic activities could also be examined more closely. The activity of transcription factors other than CTCF should be considered and observed. Especially the tag SNPs should be surveyed for possibly influencing binding motifs or alternative splicing, which provokes NMD. These epigenetic examinations were not performed due to unavailable ChIP-seq data for transcription factors and lack of time.

Both the transcriptome analysis and GWAS were conducted on European populations. Repeating these analyses with East Asian or African American populations would reveal varying etiopathology in different ethnicities. Validating the identified transcriptomic and genomic biomarkers in other populations would show whether they would be suitable for PCa outcome prediction in general or were only representative of specific subpopulations.

The results described in this thesis were mainly based on bioinformatic analysis. However, for potential prognostic and predictive biomarkers, extensive wet lab experiments are necessary for their validations before their establishment in PCa tests or clinical routines.



Appendix

A.1 | Abbreviations

AFR	African
AJCC	American Joint Committee on Cancer
APOE	Apolipoprotein E
ASPN	Asporin
BAM	Compressed binary file of aligned sequences
BCR	Biochemical relapse
BFS	BCR-free survival
BGN	Biglycan
cAMP	Cyclic Adenosine Monophosphate
CCP	Cell cycle progression
CDF	Chip description file
CEU	Utah residents (CEPH) with Northern and Western European ancestry
ChIP-seq	Chromain immunoprecipitation DNA-sequencing
CI	Confidence interval
CLIA	Clinical Laboratory Improvement Amendments, US
COL1A1	Collagen Type I Alpha 1
CPE	Consensus purity estimation

CTCF	CCCTC-binding factor
DACO	Data Access Compliance Office
DKFZ	Deutsches Krebsforschungszentrum
DLEU2	Deleted In Lymphocytic Leukemia 2
DRE	Digital rectal examination
EAS	East Asian
EFS	Event-free survival
EPI	Exo Prostate Intellli Score
ERG	ETS Transcription Factor ERG
ESTIMATE	Estimation of stromal and immune cells in malignant tumor using expression data
ETS family	E-twenty-six family
ETV1	ETS Variant Transcription Factor 1
geneETV4	ETS Variant Transcription Factor 4
EUR	European
eQTL	Expression quantitative trait loci
FISH	Fluorescence in situ hybridization
FDA	Food and Drug Administration, US
FDR	False discovery rate
FFPE sample	Formalin-fixed paraffin-embedded sample
FIN	Finnish population in Finland
FLI1	Fli-1 Proto-Oncogene
FOXA1	Forkhead Box A1
FOXM1	Forkhead Box M1
GATK	Genome Analysis Toolkit
GBS	British population in England and Scotland
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GG	Gleason grade
GGG	Gleason grade group
GMNN	Geminin DNA Replication Inhibitor
GRM8	Glutamate Metabotropic Receptor 8
GRM1	Glutamate Metabotropic Receptor 1

GSEA	Gene set enrichement analysis
GWAS	Genome-wide association study
I^2	Percentage of variation across studies because of heterogeneity [183]
IBS	Iberian population from Spain
ICGC	International Cancer Genome Consortium
ICGC-CA	Prostate adenocarcinoma study of ICGC
IDH1	Isocitrate Dehydrogenase 1
IGV	Integrative Genomics Viewer
IHC	Immunohistchemistry
kb	Kilo base
KLK3	Kallikrein 3
λ	Genomic inflation factor
LEA	Leading edge analysis
LD	Linkage disequilibrium
LNCaP	Prostate cancer cell line from lymph node carcinoma of the prostate
LRZ	Leibniz-Rechenzentrum
LY96	Lymphocyte Antigen 96
M0	Tumor stage, indicating no distant metastasis
MAF	Minor allele frequency
mGluR	Metabotropic glutamate receptors
N0	Tumor stage, indicating no involvement of regional lymph nodes
NES	Normalized enrichement score
NCI	National Cancer Institute
NMD	Nonsense-mediated RNA decay
OR	Odds ratio
ΔOR	Difference between ORs
PCa	Prostate cancer
PCA	Principal component analysis
PCA3	Prostate Cancer Associated 3
PC-3	Prostate cancer cell line
PCAWG	PanCancer Analysis of Whole Genomes
$PCGA^{LMU}$	Prostate cancer genome atlas of the Ludwig-Maximilian-University of Munich

PHI	Prostate Health Index
PKA	Proteine Kinase A
PRADA	Pipeline for RNA-sequencing data analysis
PSA	Prostate Specific Antigen
PTEN	Phosphatase And Tensin Homolog
QC	Quality check
r^2	Measure of LD [184]
RA	Risk allele
RAF	Risk allele frequency
rGL-pos/neg	Ranked gene list based on T2E-positive/-negative PCa samples
RNA-Seq	RNA sequencing
RRM2	Ribonuclease Reductase Regulatory Subunit M2
RT-PCR	Reverse transcription-polymerase chain reaction
SCAN	Single channel array normalization
SEO	Search engine optimization
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPDEF	SAM Pointed Domain Containing ETS Transcription Factor
SPOP	Speckle Type BTB/POZ Protein
STARLING	Prostate cancer research leveraging important novel genomic biomarkers
T2E	*TMPRSS2-ERG* fusion oncogene
TCGA	The Cancer Genome Atlas
TCGA-PRAD	Prostate adenocarcinoma study of TCGA
TMA	Tissue microarray
TMN	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)
TMPRSS2	Transmembrane Serine Protease 2
topGL-pos/neg	List of most frequent genes involved in top 20 gene-signatures based on T2E-positive/-negative PCa samples
TP53	Tumor Protein 53
TROAP	Trophinin Associated Protein

TSI	Tuscani population in Italy
TYMS	Thymidylate Synthetase
UICC	Union for International Cancer Control
VCF	Variant call format
VEP	Variant effect predictor
WEE1	WEE1 G2 Checkpoint Kinase
WGS	Whole genome sequencing
WXS	Whole exon sequencing

A.2 | Tables

Table A.1: Top 20 functional gene-signatures from ranked GSEA of rGL-pos (NES, normalized enrichment score; NOM, nominal P value; FDR, false discovery rate.) [105].

Gene-signature pathways	NES	NOM P	FDR q
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	2.06	0.00	0.10
BIDUS_METASTASIS_UP	2.00	0.00	0.10
CROONQUIST_IL6_DEPRIVATION_DN	1.95	0.00	0.11
CHANG_CYCLING_GENES	1.91	0.00	0.12
ODONNELL_TFRC_TARGETS_DN	1.91	0.01	0.09
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	1.89	0.01	0.10
ROSTY_CERVICAL_CANCER_PROLIFERATION_ CLUSTER	1.84	0.00	0.13
WHITFIELD_CELL_CYCLE_G1_S	1.82	0.00	0.14
FISCHER_DREAM_TARGETS	1.81	0.00	0.13
WEST_ADRENOCORTICAL_TUMOR_UP	1.77	0.00	0.17
NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON	1.75	0.01	0.19
SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_FUSION_UP	1.74	0.00	0.19
FISCHER_G2_M_CELL_CYCLE	1.71	0.01	0.22
ROY_WOUND_BLOOD_VESSEL_UP	1.68	0.01	0.25
ZHAN_MULTIPLE_MYELOMA_CD1_AND_CD2_UP	1.67	0.03	0.25
JOHANSSON_BRAIN_CANCER_EARLY_VS_LATE_DN	1.65	0.02	0.28
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_2	1.65	0.01	0.26
NIKOLSKY_MUTATED_AND_AMPLIFIED_IN_BREAST_CANCER	1.65	0.02	0.25
JIANG_TIP30_TARGETS_UP	1.62	0.03	0.28
HORIUCHI_WTAP_TARGETS_DN	1.62	0.00	0.27

Table A.2: Top 20 functional gene-signatures from ranked GSEA of rGL-neg (NES, normalized enrichment score; NOM, nominal P value; FDR, false discovery rate.) [105].

Gene-signature pathways	NES	NOM P	FDR q
POOLA_INVASIVE_BREAST_CANCER_UP	3.11	0.00	0.00
WIELAND_UP_BY_HBV_INFECTION	2.66	0.00	0.00
SANA_RESPONSE_TO_IFNG_UP	2.47	0.00	0.00
NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP	2.43	0.00	0.00
THUM_SYSTOLIC_HEART_FAILURE_UP	2.42	0.00	0.00
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_ LPS_DN	2.41	0.00	0.00
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_ DUCTAL_NORMAL_UP	2.40	0.00	0.00
HELLER_SILENCED_BY_METHYLATION_UP	2.39	0.00	0.00
LINDSTEDT_DENDRITIC_CELL_MATURATION_A	2.38	0.00	0.00
BROWNE_INTERFERON_RESPONSIVE_GENES	2.43	0.00	0.00
LEE_DIFFERENTIATING_T_LYMPHOCYTE	2.33	0.00	0.00
RODWELL_AGING_KIDNEY_UP	2.32	0.00	0.00
GAURNIER_PSMD4_TARGETS	2.32	0.00	0.00
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP	2.32	0.00	0.00
BOSCO_TH1_CYTOTOXIC_MODULE	2.31	0.00	0.00
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	2.31	0.00	0.00
MCLACHLAN_DENTAL_CARIES_UP	2.30	0.00	0.00
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_ LOBULAR_NORMAL_DN	2.29	0.00	0.00
KIM_GLIS2_TARGETS_UP	2.29	0.00	0.00
GRAESSMANN_RESPONSE_TO_MC_AND_SERUM_ DEPRIVATION_UP	2.29	0.00	0.00

Table A.3: Result summary of all statistical tests of topGL-pos [105].

Dataset	GSE46691	TCGA-PRAD			GSE16560	
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
ANP32E	3.74E-01	2.80E-01	8.58E-01	low	8.50E-02	low
ASF1B	4.97E-01	2.64E-01	5.89E-01	low	-	-
CDC20	2.36E-01	5.62E-02	4.43E-01	low	7.20E-02	high
CKS2	6.27E-01	1.06E-01	2.51E-01	low	9.84E-01	low
DEPDC1	1.19E-01	1.41E-01	1.63E-01	low	-	-
FAM83D	4.28E-01	1.57E-02	3.06E-01	low	-	-
GMNN	4.08E-05	5.61E-03	1.39E-01	low	6.40E-01	high
KIF4A	8.53E-01	6.14E-02	3.92E-01	low	-	-
PTTG1	7.00E-01	1.50E-01	1.70E-01	low	8.37E-01	low
RRM2	5.06E-03	1.90E-01	4.19E-01	high	2.15E-01	low
SPC25	3.91E-01	4.45E-02	5.38E-01	low	-	-
TROAP	2.11E-02	3.24E-02	6.70E-02	low	1.17E-01	high
TYMS	2.86E-05	3.12E-01	6.68E-01	high	6.87E-01	low
UBE2C	4.41E-01	7.40E-02	2.37E-01	low	1.66E-01	high
UHRF1	7.87E-02	1.78E-01	3.52E-01	low	-	-
WEE1	1.07E-04	1.65E-03	1.98E-01	low	2.11E-01	high

Table A.4: Result summary of all statistical tests of topGL-neg [105].

Dataset	GSE46691	TCGA-PRAD			GSE16560	
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
AIF1	7.65E-01	3.95E-04	2.40E-02	low	1.10E-01	low
APOC1	1.28E-01	1.69E-03	5.00E-02	low	8.60E-01	low
APOE	1.10E-01	1.20E-02	2.10E-02	low	5.00E-03	low
ARHGDIB	1.79E-02	6.77E-05	1.37E-01	low	8.24E-01	low
ASPN	7.37E-05	8.92E-06	1.00E-03	low	3.05E-04	low
BGN	2.53E-03	3.52E-04	1.50E-02	low	4.84E-04	low
BST2	2.08E-01	9.49E-03	1.22E-01	low	5.69E-01	high
C1QB	1.60E-01	9.89E-05	4.40E-02	low	7.39E-01	low
CCL2	1.21E-01	1.47E-01	8.88E-01	low	8.50E-01	low
CCL8	5.42E-02	6.22E-05	1.78E-01	low	7.45E-01	low
CCR1	9.57E-01	6.10E-06	1.62E-01	low	3.85E-01	low
CD14	7.62E-02	7.30E-02	7.30E-02	low	3.17E-01	low
CD52	1.71E-02	1.55E-04	5.80E-02	low	6.18E-01	high
CD53	1.55E-01	5.29E-06	1.21E-01	low	5.59E-01	high
CD74	5.39E-02	3.81E-04	2.53E-01	low	6.29E-01	high
CDH11	9.99E-04	4.55E-04	4.34E-01	low	8.06E-01	high
CFB	1.79E-01	2.07E-01	1.88E-01	high	3.50E-02	high
COL1A1	1.57E-04	9.64E-06	2.50E-02	low	7.00E-03	low
COL1A2	2.45E-03	2.75E-03	3.50E-01	low	2.32E-01	low
COL3A1	3.19E-04	2.34E-05	1.26E-01	low	8.00E-03	low
COMP	3.25E-01	5.55E-06	3.00E-03	low	9.60E-02	low
CTSS	6.21E-02	5.31E-05	3.77E-01	low	7.97E-01	low
CXCL11	2.93E-01	3.65E-06	5.30E-02	low	4.07E-01	high
CXCL13	2.33E-03	1.21E-02	5.19E-01	low	9.00E-03	high
CXCL9	9.40E-01	2.09E-06	3.10E-02	low	2.20E-01	high
CXCR4	9.25E-03	1.00E-03	1.66E-01	low	3.50E-01	low
EVI2B	9.61E-03	1.84E-05	4.35E-01	low	1.82E-01	high
F13A1	5.15E-01	7.11E-03	4.28E-01	low	2.68E-01	low

Continues on next page.

Table A.4 continued: Result summary of all statistical tests of topGL-neg [105].

Dataset	GSE46691	TCGA-PRAD			GSE16560	
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
FCGR2A	3.71E-01	1.98E-05	5.70E-02	low	3.70E-02	low
FN1	9.92E-04	4.01E-04	6.90E-02	low	2.90E-02	low
FYB	1.32E-01	4.65E-06	2.75E-01	low	7.99E-01	high
GBP1	1.69E-01	3.35E-04	4.36E-01	low	8.44E-01	low
GPNMB	2.45E-03	6.78E-04	9.10E-02	low	9.70E-02	low
GZMB	1.24E-01	6.68E-02	9.00E-01	low	7.70E-01	high
GZMK	8.94E-01	1.36E-03	5.30E-01	low	6.00E-03	high
HCLS1	9.99E-02	6.22E-05	1.80E-02	low	8.12E-01	high
HLA-DMB	2.32E-01	4.06E-04	1.90E-02	low	4.95E-01	high
HLA-DPA1	2.90E-02	1.64E-04	7.20E-02	low	9.38E-01	high
HLA-DPB1	5.98E-02	2.06E-03	4.60E-02	low	6.42E-01	high
HLA-DRA	2.91E-03	2.20E-05	1.60E-01	low	8.65E-01	high
HLA-DRB1	4.64E-01	8.05E-04	2.60E-02	low	9.74E-01	high
HLA-E	1.65E-01	4.97E-04	9.48E-01	high	1.65E-01	high
HLA-F	3.82E-01	5.63E-04	2.90E-02	low	9.97E-01	high
IFI27	2.35E-02	4.78E-04	1.04E-01	low	1.98E-01	low
IFI30	2.45E-03	9.36E-05	1.40E-02	low	2.75E-01	low
IFI44	6.44E-01	8.36E-07	1.10E-02	low	5.00E-02	high
IFIT3	4.59E-01	1.54E-08	2.20E-02	low	5.45E-01	high
INHBA	8.47E-03	4.50E-06	9.10E-02	low	5.07E-04	low
ISG15	9.18E-01	1.55E-04	8.64E-01	high	7.36E-01	high
LAPTM5	7.16E-02	4.74E-05	3.90E-02	low	2.49E-01	low
LRRC15	1.76E-01	3.27E-03	2.05E-01	low	-	-
LST1	1.62E-01	1.46E-03	4.70E-02	low	3.26E-01	high
LTB	5.51E-01	1.94E-04	8.70E-01	low	5.12E-01	low
LUM	2.25E-02	3.80E-02	6.96E-01	high	2.66E-01	low
LY96	8.76E-01	4.34E-05	1.00E-03	low	4.30E-02	low
LYZ	4.25E-02	2.86E-04	2.01E-01	low	2.45E-01	high
MS4A4A	8.83E-01	1.22E-03	8.10E-02	low	9.59E-01	low

Continues on next page.

Table A.4 continued: Result summary of all statistical tests of topGL-neg [105].

Dataset	GSE46691	TCGA-PRAD			GSE16560	
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
MS4A6A	9.10E-03	8.37E-06	1.70E-02	low	9.24E-01	low
PLA2G7	2.68E-05	3.95E-02	5.20E-02	low	4.31E-06	low
PLEK	9.60E-01	5.82E-06	5.76E-01	low	5.31E-01	high
POSTN	2.78E-04	3.06E-04	6.60E-02	low	7.00E-03	low
PSMB9	5.86E-01	1.01E-05	1.41E-01	low	1.93E-01	high
PTPRC	2.55E-01	2.45E-05	1.11E-01	low	9.28E-01	low
RARRES3	1.60E-01	1.00E-03	3.42E-01	low	2.93E-01	high
RGS1	4.94E-02	2.48E-05	1.98E-01	low	6.01E-01	high
RRM2	4.36E-02	8.02E-05	4.35E-05	low	2.00E-03	low
SAMHD1	1.50E-01	1.25E-03	8.06E-01	high	5.45E-01	low
SPARC	1.38E-04	6.53E-04	5.08E-01	low	1.82E-01	low
STAT1	1.32E-01	2.04E-05	5.39E-01	high	1.36E-01	high
SULF1	4.44E-04	1.55E-03	6.12E-01	low	7.58E-01	high
TRIM22	2.34E-01	1.44E-03	6.88E-01	high	8.55E-01	high
TYMS	8.81E-03	1.77E-02	1.00E-03	low	3.29E-06	low
TYROBP	1.81E-02	6.78E-04	8.10E-02	low	5.20E-01	low
UBE2L6	1.60E-01	3.87E-05	9.80E-01	high	9.16E-01	high

Table A.5: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *ASPN*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *ASPN*	High gene expression of *ASPN*
POOLA_INVASIVE_BREAST_CANCER_UP	POOLA_INVASIVE_BREAST_CANCER_UP
BROWNE_INTERFERON_RESPONSIVE_GENES	WIELAND_UP_BY_HBV_INFECTION
DER_IFN_ALPHA_RESPONSE_UP	LEE_DIFFERENTIATING_T_LYMPHOCYTE
BOSCO_TH1_CYTOTOXIC_MODULE	BROWNE_INTERFERON_RESPONSIVE_GENES
KRASNOSELSKAYA_ILF3_TARGETS_UP	BOSCO_TH1_CYTOTOXIC_MODULE
RADAEVA_RESPONSE_TO_IFNA1_UP	SANA_RESPONSE_TO_IFNG_UP
WIELAND_UP_BY_HBV_INFECTION	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP
LEE_DIFFERENTIATING_T_LYMPHOCYTE	ALTEMEIER_RESPONSE_TO_LPS_WITH_ MECHANICAL_VENTILATION
SANA_RESPONSE_TO_IFNG_UP	SENGUPTA_NASOPHARYNGEAL_ CARCINOMA_UP
DER_IFN_BETA_RESPONSE_UP	RASHI_RESPONSE_TO_IONIZING_RADIATION_6
FARMER_BREAST_CANCER_CLUSTER_1	SMIRNOV_RESPONSE_TO_IR_6HR_DN
FULCHER_INFLAMMATORY_RESPONSE_ LECTIN_VS_LPS_DN	FARMER_BREAST_CANCER_CLUSTER_1
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_ MODULE	ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_ UP
DER_IFN_GAMMA_RESPONSE_UP	VILIMAS_NOTCH1_TARGETS_UP
ZHAN_MULTIPLE_MYELOMA_LB_DN	LU_IL4_SIGNALING
GAURNIER_PSMD4_TARGETS	MCLACHLAN_DENTAL_CARIES_UP
GRAESSMANN_RESPONSE_TO_MC_AND_ SERUM_ DEPRIVATION_UP	WALLACE_PROSTATE_CANCER_RACE_UP
MORI_MATURE_B_LYMPHOCYTE_UP	DEBIASI_APOPTOSIS_BY_REOVIRUS_ INFECTION_UP
WORSCHECH_TUMOR_REJECTION_UP	GAURNIER_PSMD4_TARGETS
SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_ FUSION_UP	FULCHER_INFLAMMATORY_RESPONSE_ LECTIN_VS_LPS_DN

Table A.6: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *BGN*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *BGN*	High gene expression of *BGN*
POOLA_INVASIVE_BREAST_CANCER_UP	POOLA_INVASIVE_BREAST_CANCER_UP
FISCHER_DREAM_TARGETS	WIELAND_UP_BY_HBV_INFECTION
WIELAND_UP_BY_HBV_INFECTION	LEE_DIFFERENTIATING_T_LYMPHOCYTE
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP
BROWNE_INTERFERON_RESPONSIVE_GENES	BROWNE_INTERFERON_RESPONSIVE_GENES
PUJANA_BRCA2_PCC_NETWORK	SANA_RESPONSE_TO_IFNG_UP
RASHI_RESPONSE_TO_IONIZING_RADIATION_6	DEBIASI_APOPTOSIS_BY_REOVIRUS_ INFECTION_UP
FARMER_BREAST_CANCER_CLUSTER_1	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP
LEE_DIFFERENTIATING_T_LYMPHOCYTE	TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	RASHI_RESPONSE_TO_IONIZING_RADIATION_6
CROONQUIST_IL6_DEPRIVATION_DN	THUM_SYSTOLIC_HEART_FAILURE_UP
BOSCO_TH1_CYTOTOXIC_MODULE	MORI_MATURE_B_LYMPHOCYTE_UP
FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_REJECTED_VS_OK_UP	BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE
BERTUCCI_MEDULLARY_VS_DUCTAL_BREAST_CANCER_UP	BOSCO_TH1_CYTOTOXIC_MODULE
MORI_LARGE_PRE_BII_LYMPHOCYTE_DN	DEURIG_T_CELL_PROLYMPHOCYTIC_LEUKEMIA_DN
ICHIBA_GRAFT_VERSUS_HOST_DISEASE_D7_UP	DER_IFN_ALPHA_RESPONSE_UP
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN	MORI_LARGE_PRE_BII_LYMPHOCYTE_DN
MCLACHLAN_DENTAL_CARIES_UP	LINDSTEDT_DENDRITIC_CELL_MATURATION_A
ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_UP	RODWELL_AGING_KIDNEY_UP
MORI_MATURE_B_LYMPHOCYTE_UP	TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_DN

Table A.7: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *COL1A1*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *COL1A1*	High gene expression of *COL1A1*
FISCHER_DREAM_TARGETS	POOLA_INVASIVE_BREAST_CANCER_UP
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	WIELAND_UP_BY_HBV_INFECTION
SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_FUSION_UP	BASSO_CD40_SIGNALING_UP
POOLA_INVASIVE_BREAST_CANCER_UP	LEE_DIFFERENTIATING_T_LYMPHOCYTE
PUJANA_BRCA2_PCC_NETWORK	SANA_RESPONSE_TO_IFNG_UP
CROONQUIST_IL6_DEPRIVATION_DN	FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN
FARMER_BREAST_CANCER_CLUSTER_1	BROWNE_INTERFERON_RESPONSIVE_GENES
BROWNE_INTERFERON_RESPONSIVE_GENES	GAURNIER_PSMD4_TARGETS
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	DER_IFN_BETA_RESPONSE_UP
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_DN	WALLACE_PROSTATE_CANCER_RACE_UP
WHITEFORD_PEDIATRIC_CANCER_MARKERS	MCLACHLAN_DENTAL_CARIES_UP
RADAEVA_RESPONSE_TO_IFNA1_UP	RASHI_RESPONSE_TO_IONIZING_RADIATION_6
BENPORATH_PROLIFERATION	DER_IFN_ALPHA_RESPONSE_UP
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_DN	ALTEMEIER_RESPONSE_TO_LPS_WITH_ MECHANICAL_VENTILATION
MORI_MATURE_B_LYMPHOCYTE_UP	THUM_SYSTOLIC_HEART_FAILURE_UP
FISCHER_G2_M_CELL_CYCLE	SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	BOSCO_TH1_CYTOTOXIC_MODULE
WIELAND_UP_BY_HBV_INFECTION	LINDSTEDT_DENDRITIC_CELL_MATURATION_A
MORI_LARGE_PRE_BII_LYMPHOCYTE_DN	ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_UP
ZHAN_MULTIPLE_MYELOMA_LB_DN	FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_REJECTED_VS_OK_UP

Table A.8: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *RRM2*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *RRM2*	High gene expression of *RRM2*
POOLA_INVASIVE_BREAST_CANCER_UP	POOLA_INVASIVE_BREAST_CANCER_UP
RODWELL_AGING_KIDNEY_UP	WIELAND_UP_BY_HBV_INFECTION
SCHUETZ_BREAST_CANCER_DUCTAL_ INVASIVE_UP	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP
THUM_SYSTOLIC_HEART_FAILURE_UP	LEE_DIFFERENTIATING_T_LYMPHOCYTE
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP	BOSCO_TH1_CYTOTOXIC_MODULE
WIELAND_UP_BY_HBV_INFECTION	BROWNE_INTERFERON_RESPONSIVE_GENES
RASHI_RESPONSE_TO_IONIZING_RADIATION_6	HELLER_SILENCED_BY_METHYLATION_UP
NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP	RASHI_RESPONSE_TO_IONIZING_RADIATION_6
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_DN	GRAESSMANN_RESPONSE_TO_MC_AND_SERUM_DEPRIVATION_UP
LEE_DIFFERENTIATING_T_LYMPHOCYTE	SANA_RESPONSE_TO_IFNG_UP
MORI_MATURE_B_LYMPHOCYTE_UP	ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_UP
GAURNIER_PSMD4_TARGETS	DER_IFN_BETA_RESPONSE_UP
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN	WALLACE_PROSTATE_CANCER_RACE_UP
RODWELL_AGING_KIDNEY_NO_BLOOD_UP	FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN
PUJANA_ATM_PCC_NETWORK	DER_IFN_ALPHA_RESPONSE_UP
KIM_GLIS2_TARGETS_UP	ICHIBA_GRAFT_VERSUS_HOST_DISEASE_D7_UP
MCLACHLAN_DENTAL_CARIES_UP	RODWELL_AGING_KIDNEY_UP
ANASTASSIOU_MULTICANCER_INVASIVENESS_SIGNATURE	KRASNOSELSKAYA_ILF3_TARGETS_UP
BROWNE_INTERFERON_RESPONSIVE_GENES	GAURNIER_PSMD4_TARGETS
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_UP	QI_PLASMACYTOMA_UP

Table A.9: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *TYMS*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *TYMS*	High gene expression of *TYMS*
POOLA_INVASIVE_BREAST_CANCER_UP	POOLA_INVASIVE_BREAST_CANCER_UP
RASHI_RESPONSE_TO_IONIZING_RADIATION_6	WIELAND_UP_BY_HBV_INFECTION
LEE_DIFFERENTIATING_T_LYMPHOCYTE	BOSCO_TH1_CYTOTOXIC_MODULE
WIELAND_UP_BY_HBV_INFECTION	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP
FARMER_BREAST_CANCER_CLUSTER_1	SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_DN	TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_DN
THUM_SYSTOLIC_HEART_FAILURE_UP	TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN	RASHI_RESPONSE_TO_IONIZING_RADIATION_6
GAURNIER_PSMD4_TARGETS	BROWNE_INTERFERON_RESPONSIVE_GENES
FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_REJECTED_VS_OK_UP	LEE_DIFFERENTIATING_T_LYMPHOCYTE
BROWNE_INTERFERON_RESPONSIVE_GENES	GARY_CD5_TARGETS_UP
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_DN	DODD_NASOPHARYNGEAL_CARCINOMA_DN
MCLACHLAN_DENTAL_CARIES_UP	SCHUETZ_BREAST_CANCER_DUCTAL_ INVASIVE_UP
SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_FUSION_UP	THUM_SYSTOLIC_HEART_FAILURE_UP
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP	FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN
ICHIBA_GRAFT_VERSUS_HOST_DISEASE_35D_UP	WALLACE_PROSTATE_CANCER_RACE_UP
VILIMAS_NOTCH1_TARGETS_UP	QI_PLASMACYTOMA_UP
HADDAD_B_LYMPHOCYTE_PROGENITOR	HADDAD_T_LYMPHOCYTE_AND_NK_ PROGENITOR_DN

Continues on next page.

Table A.9 continued: Top 20 functional gene-signatures from ranked GSEA of T2E-negative samples stratified by low and high gene expression of *TYMS*. For each gene, the 20 Pathways were sorted descending by their NES [105].

Low gene expression of *TYMS*	High gene expression of *TYMS*
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN
RODWELL_AGING_KIDNEY_UP	ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER

Table A.10: Summary of *P*-values of Kaplan-Meier analyses shown in Figure 3.4b. *P*-values were determined via Mantel-Haenszel test by comparing samples with high versus low expression of the indicated subtype specific biomarkers separately in cases stratified by low (I-III) and high (IV/V) GGG. Significant genes were highlighted in bold font [105].

		TCGA-PRAD		GSE16560	
PCa subtype	Gene	GGG low (III)	GGG high (IV/V)	GGG low (I-III)	GGG high (IV/V)
T2E-positive	ASPN	0.249	0.357	0.275	0.471
	BGN	0.301	0.828	0.430	0.132
	COL1A1	0.520	0.857	0.211	0.978
	RRM2	0.622	0.800	0.467	0.704
	TYMS	0.506	0.722	0.685	0.758
T2E-negative	**ASPN**	0.919	0.151	**0.005**	0.104
	BGN	0.391	0.374	0.111	0.674
	COL1A1	0.958	0.090	0.064	0.077
	RRM2	0.262	**0.003**	**0.007**	**0.001**
	TYMS	0.391	**0.004**	**0.001**	**0.021**

A.3 | Figures

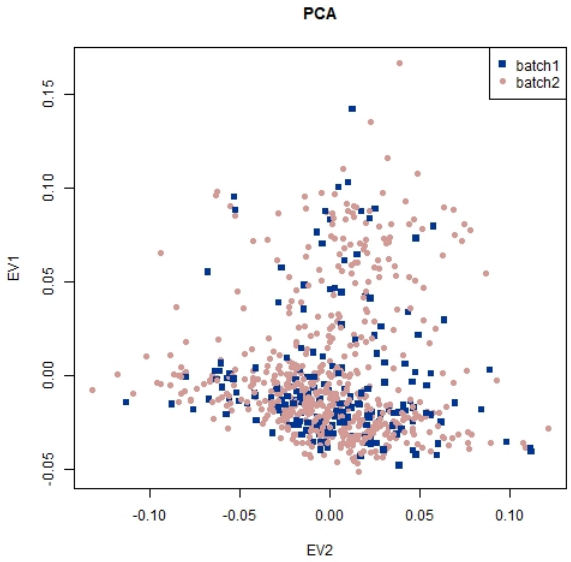
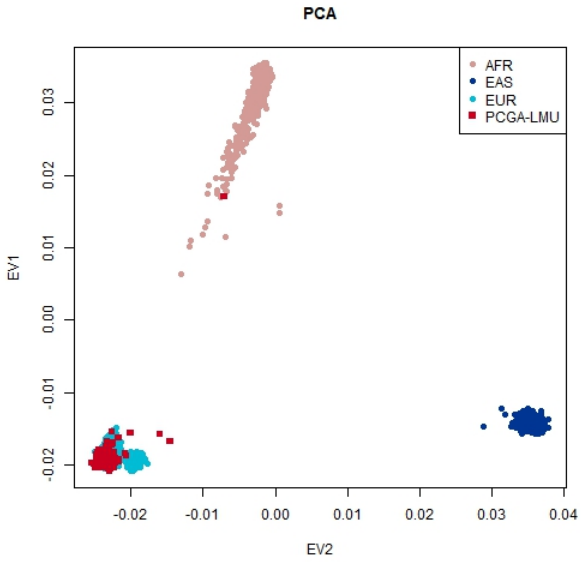
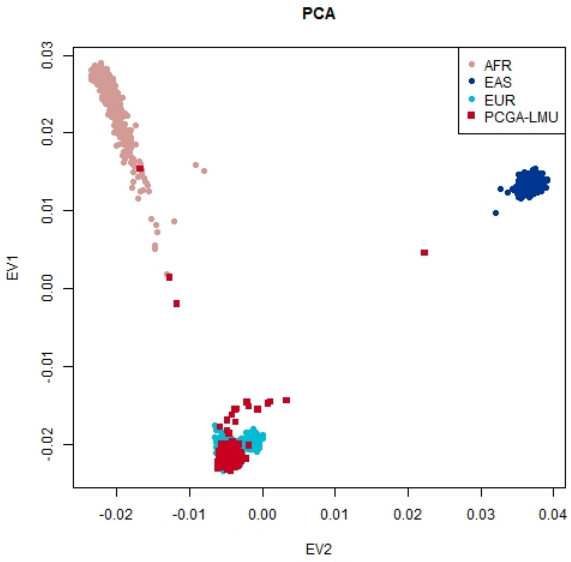


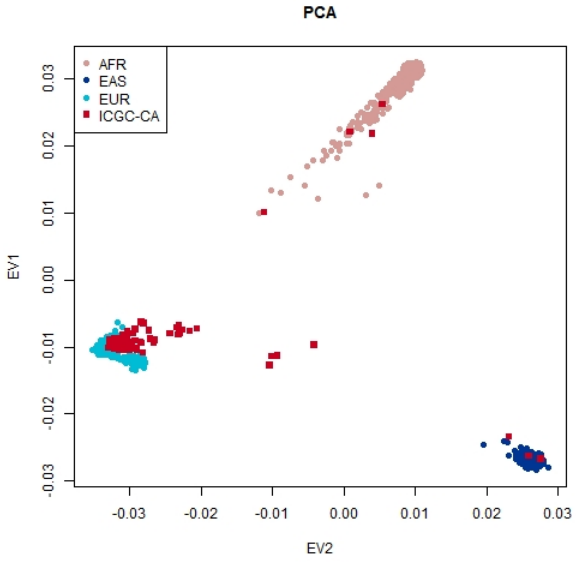
Figure A.1: Genomic differences between $PCGA^{LMU}$ samples genotyped by GSA-MD chip v1.0 (batch 1, n = 199) and GSA-MD chip v2.0 (batch 2, n = 589).



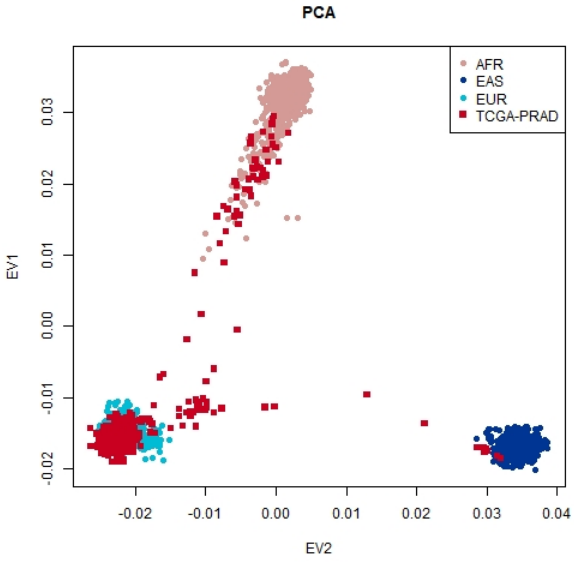
(a) PCGA^{LMU} (batch 1), n = 199



(b) PCGA^{LMU} (batch 2), n = 589

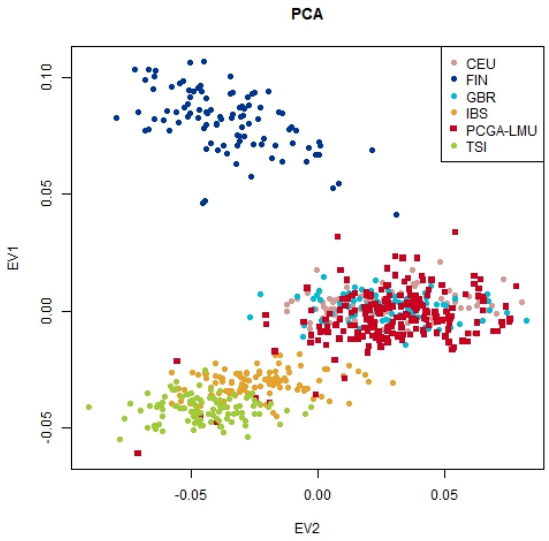


(c) ICGC-CA, n = 113

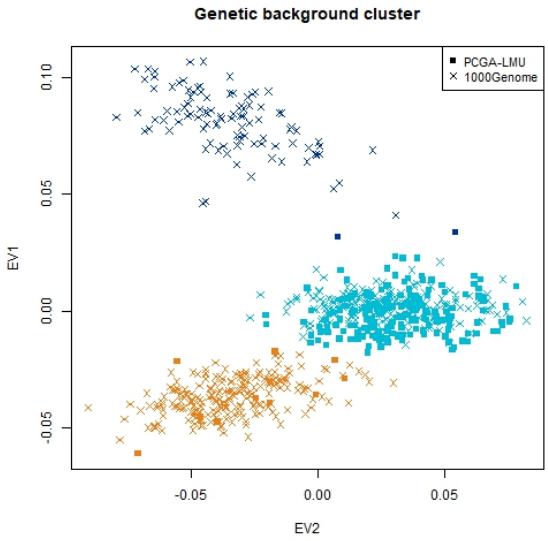


(d) TCGA-PRAD, n = 393

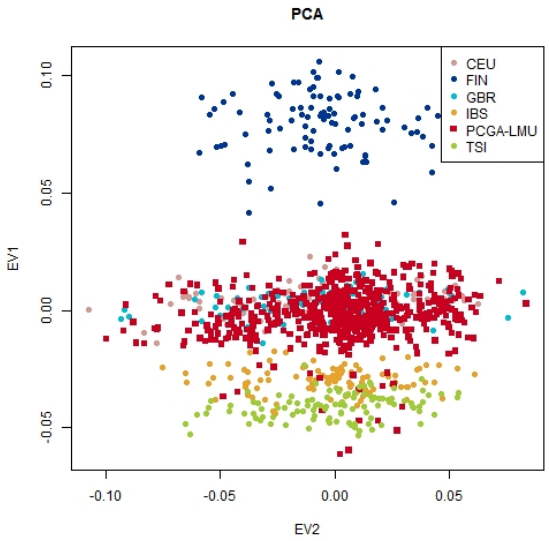
Figure A.2: Principal component analysis (PCA) of a+b) PCGA^{LMU}, c) ICGC-CA and d) TCGA-PRAD against the 1000 Genomes reference set (worldwide: AFR = African, EAS = East Asian, EUR = European) to determine each sample's ethnical background.



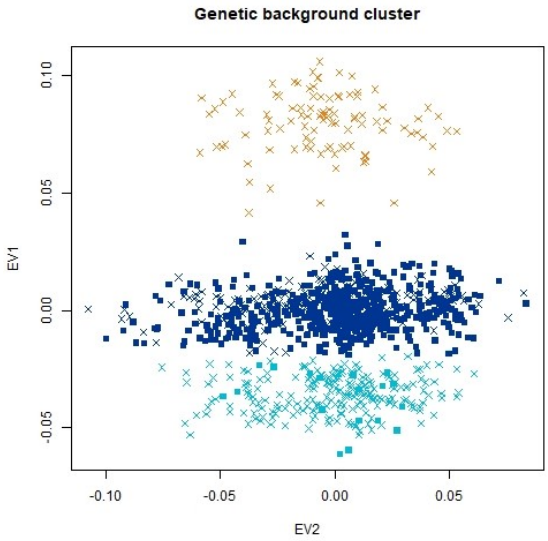
(a) European PCGA^{LMU} (batch 1)



(b) European PCGA^{LMU} (batch 1), clustered by Central European (turquoise), Mediterranean (yellow) and Scandinavian (blue).

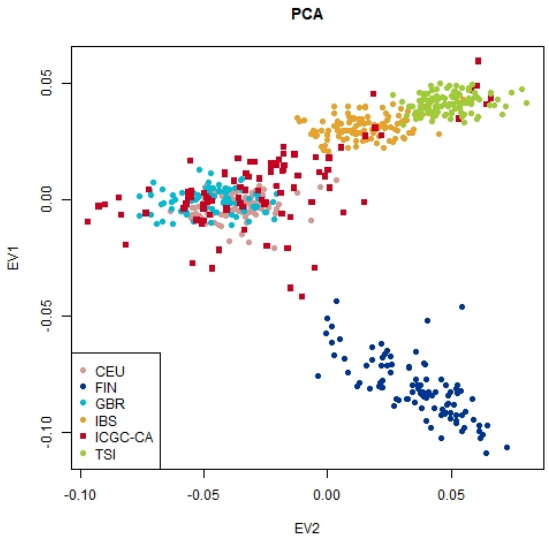


(c) European PCGA^{LMU} (batch 2)

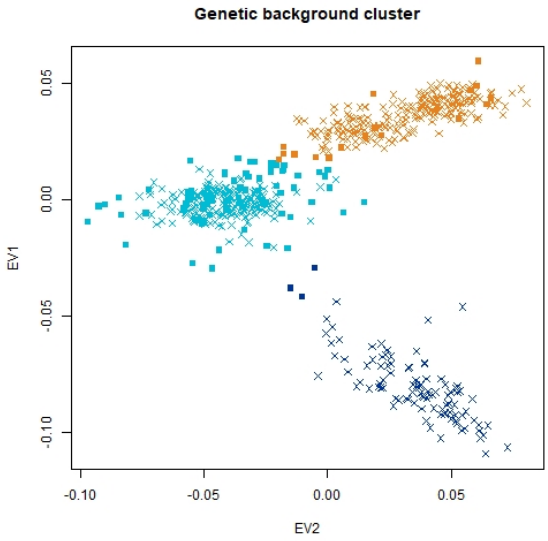


(d) European PCGA^{LMU} (batch 2), clustered by Central European (blue), Mediterranean (turquoise) and Scandinavian (yellow).

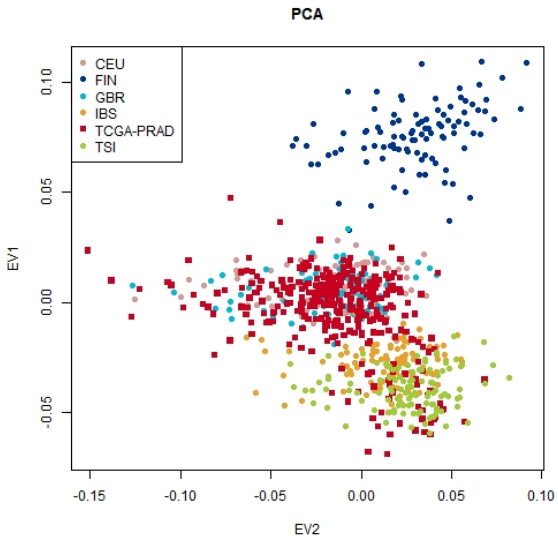
Figure A.3: Prinicipal component analysis (PCA) of European PCGA^{LMU} samples (batch 1 and 2) against European samples from the 1000 Genomes reference set colored a+c) by population and b+d) by cluster (Central European, Mediterranean, Scandinavian).



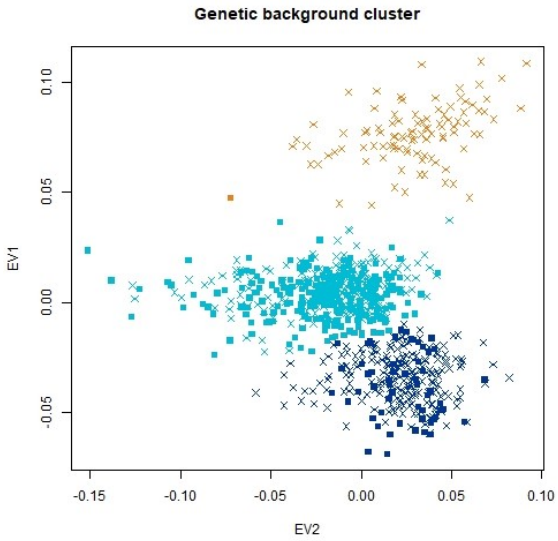
(a) European ICGC-CA



(b) European ICGC-CA, clustered by Central European (turquoise), Mediterranean (yellow) and Scandinavian (blue).

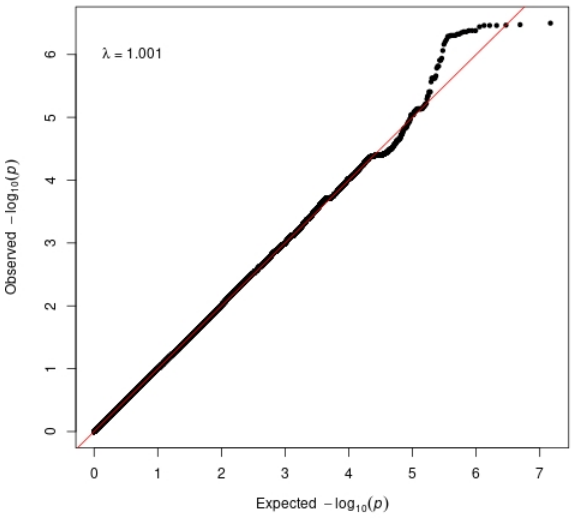


(c) European TCGA-PRAD

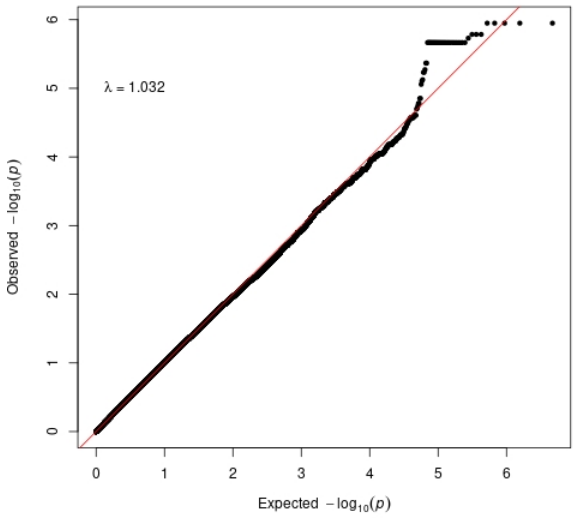


(d) European TCGA-PRAD, clustered by Central European (turquoise), Mediterranean (blue) and Scandinavian (yellow).

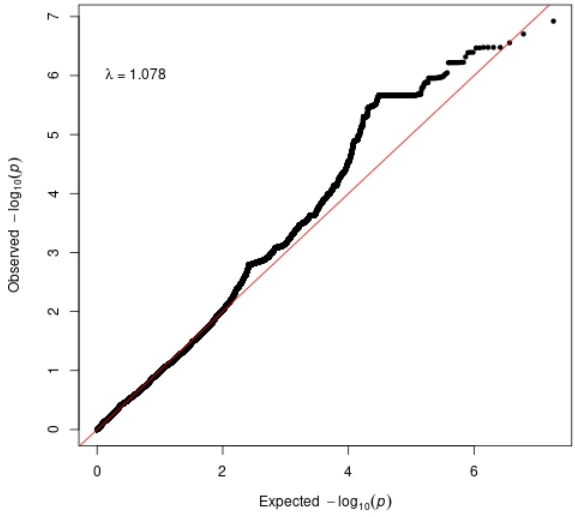
Figure A.4: Prinicipal component analysis (PCA) of European ICGC-CA and European TCGA-PRAD samples against European samples from the 1000 Genomes reference set colored a+c) by population and by b+d) by cluster of Central European, Mediterranean, Scandinavian).



(a) PCGA$^{\text{LMU}}$

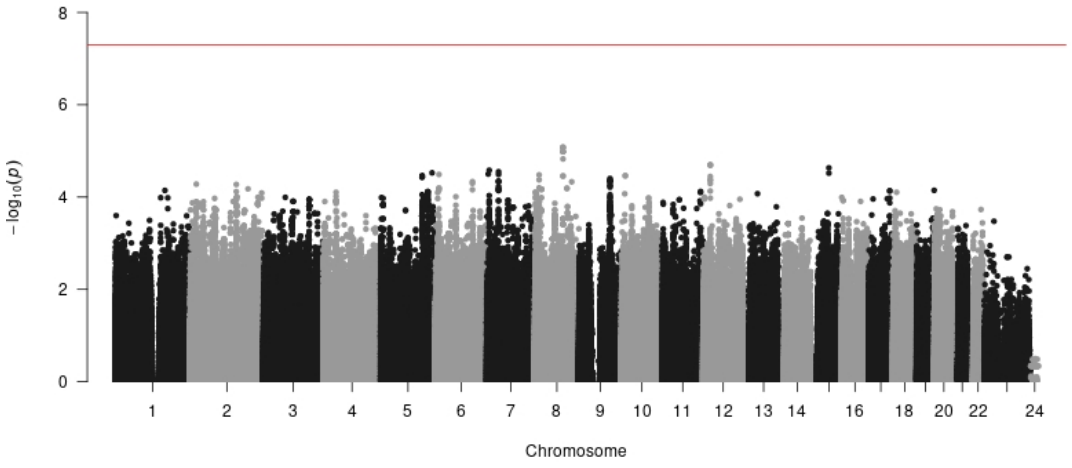


(b) TCGA-PRAD

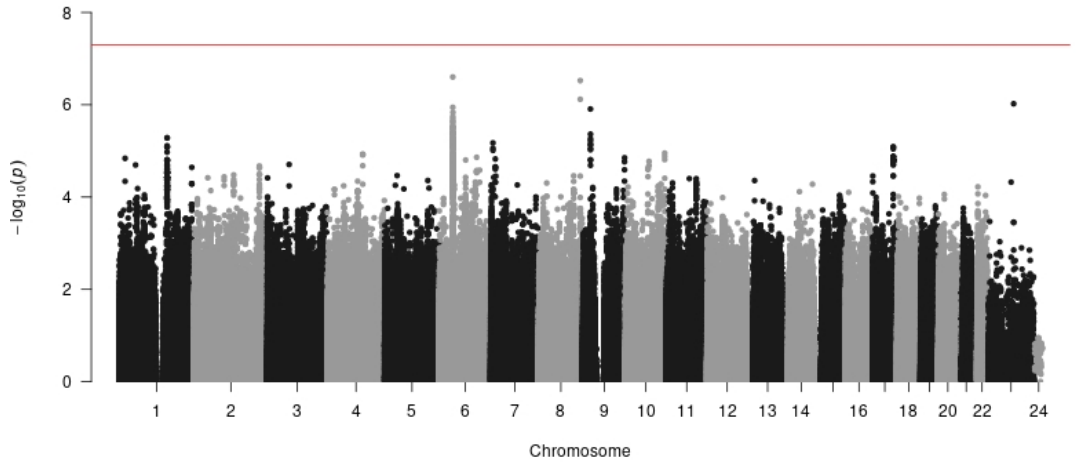


(c) ICGC-CA

Figure A.5: QQ-plots evaluating GWAS testing variants against GGG (I/II, III, IV/V) for a) PCGA$^{\text{LMU}}$, b) TCGA-PRAD and c) ICGC-CA.



(a) GWAS on T2E fusion (yes/no))



(b) GWAS on tumor encapsulation (T1+2/T3+4)

Figure A.6: Manhattan plots representing the meta-analysis results of GWAS without genome-wide significant hits. Genome-wide significance was marked with a red line.

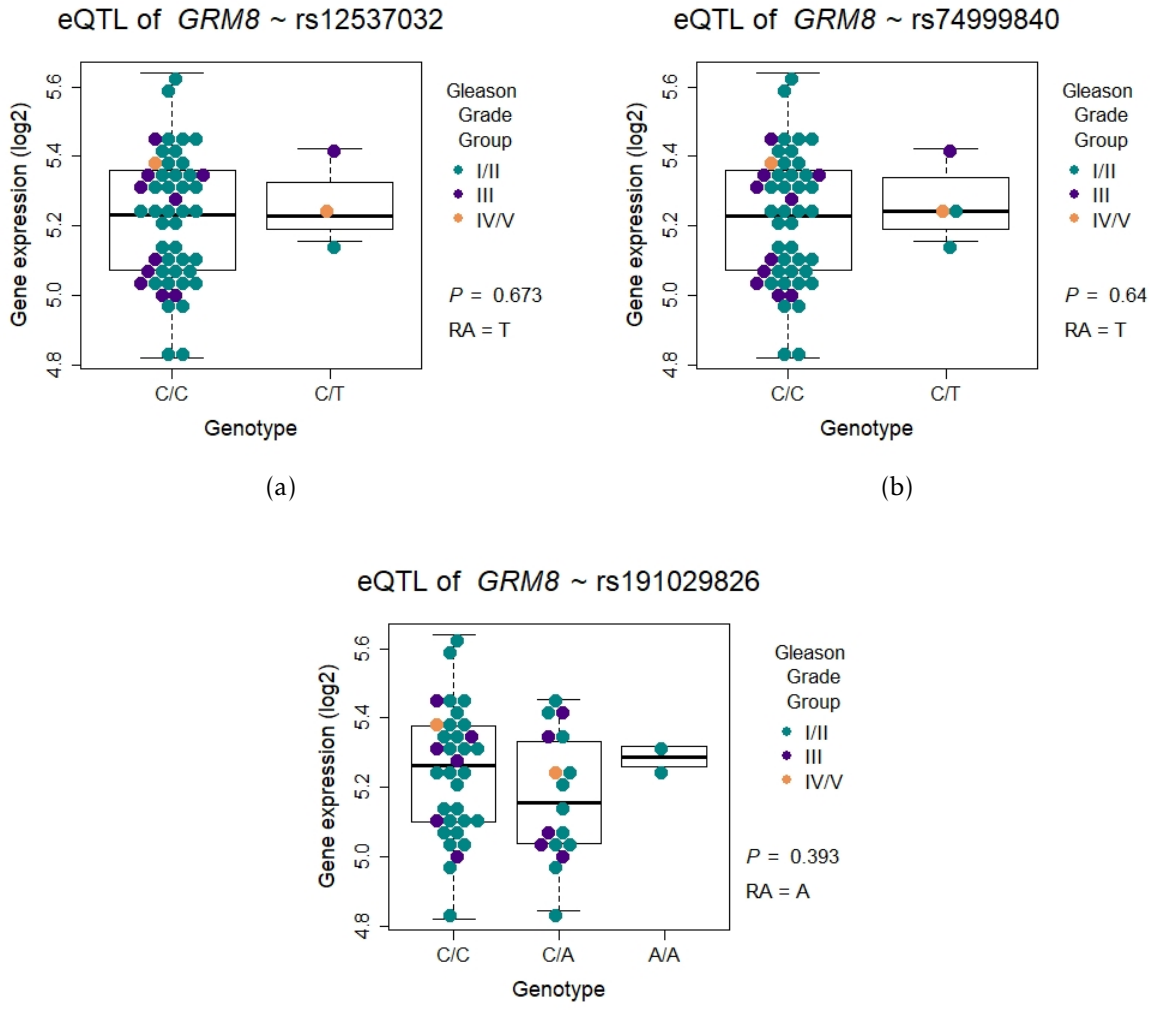
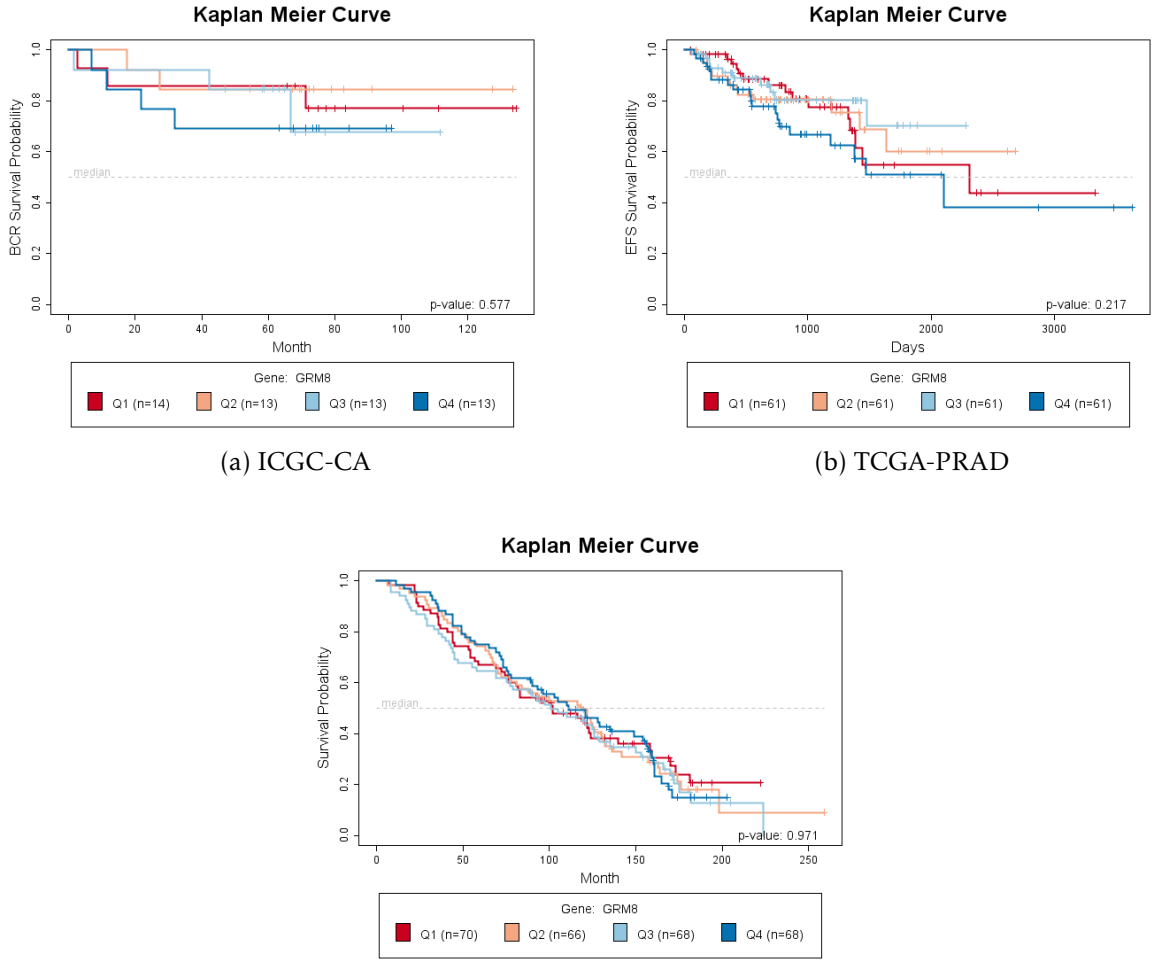


Figure A.7: Non-significant results of eQTL analysis of tag SNPs a) rs12537032 b) rs74999840 and c) rs191029826 against *GRM8* from the ICGC-CA cohort. (RA = risk allele)

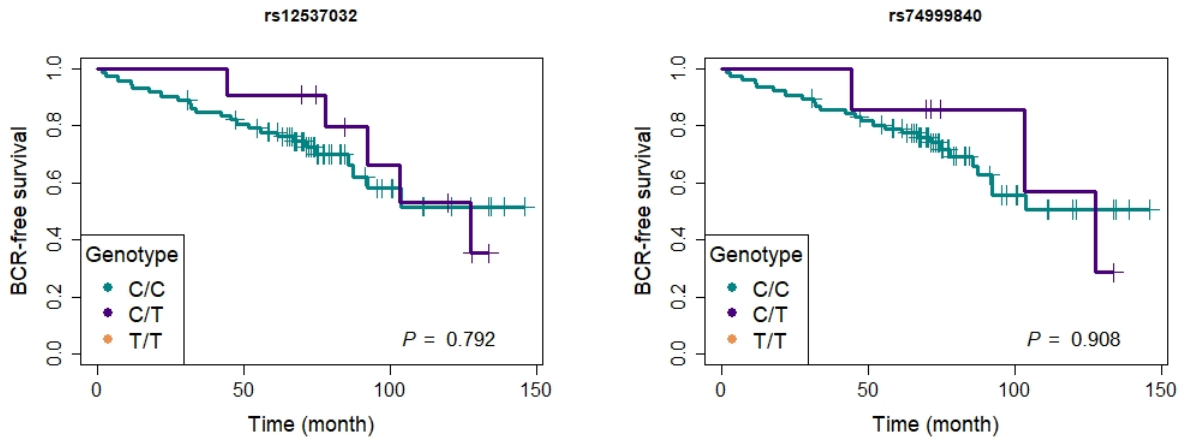


(a) ICGC-CA

(b) TCGA-PRAD

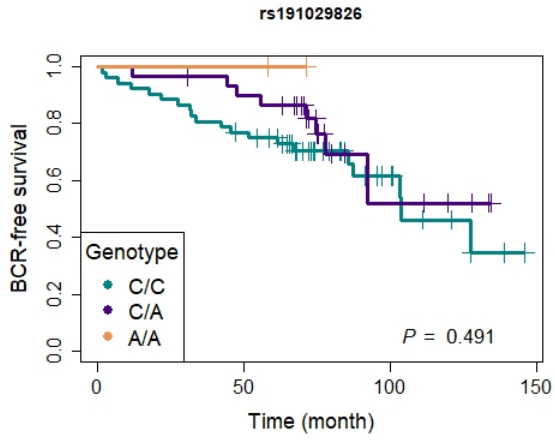
(c) GSE16560

Figure A.8: Kaplan-Meier survival plots generated from cohorts a) ICGC-CA, b) TCGA-PRAD and c) GSE16560 cohort. Samples were stratified by their quartile intratumoral gene expression level of *GRM8*. *P*-values were calculated between the most extremes (highest vs. lowest) using a Mantel-Haenszel test.

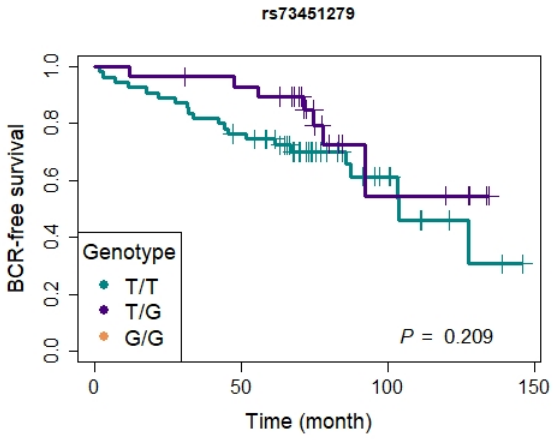


(a) rs12537032

(b) rs74999840

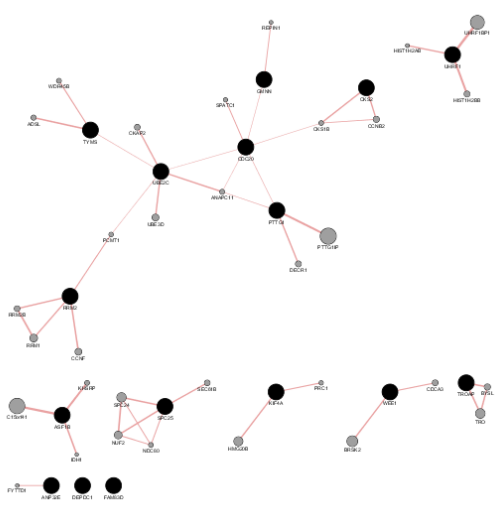


(c) rs191029826

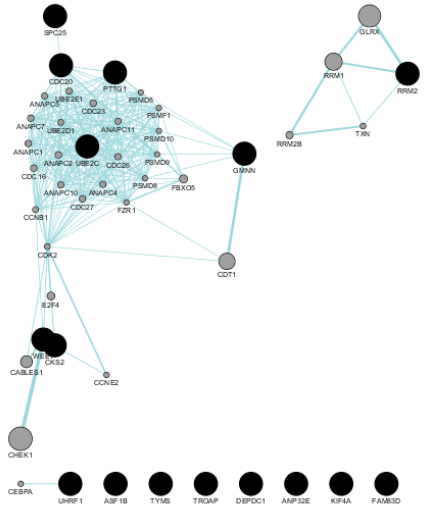


(d) rs73451279

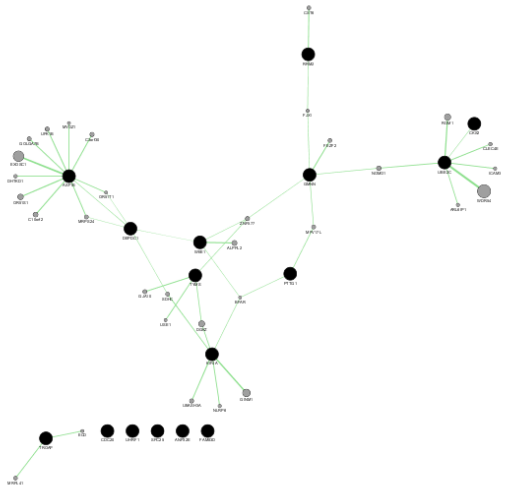
Figure A.9: Kaplan-Meier survival plots based on samples from the ICGC-CA cohort grouped by the genotype of identified tag SNPs. *P*-values were calculated between all groups using a Mantel-Haenszel test.



(a) Physical interaction network

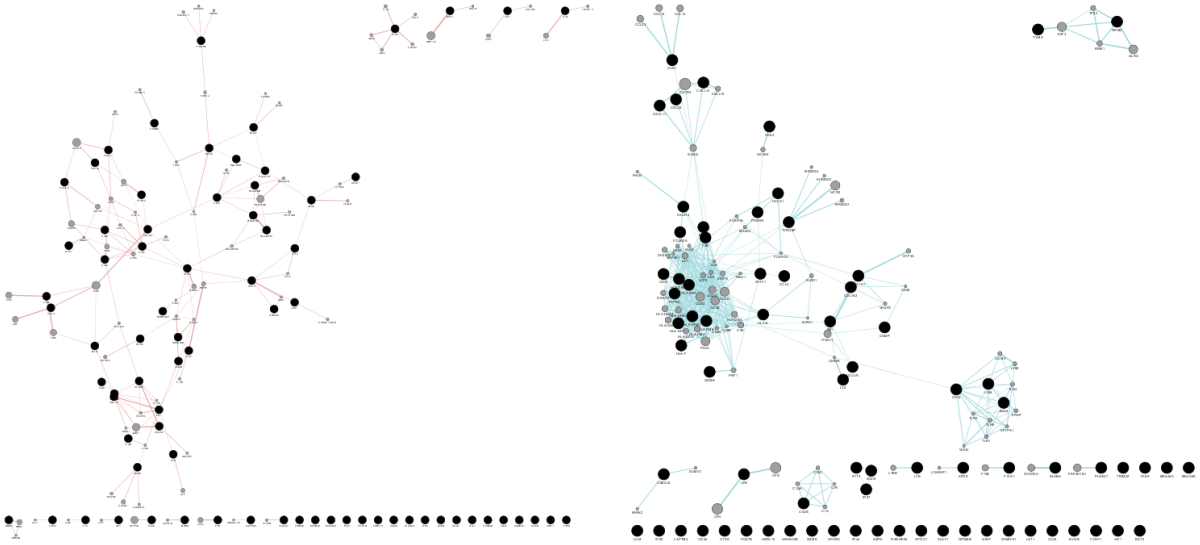


(b) Pathway network



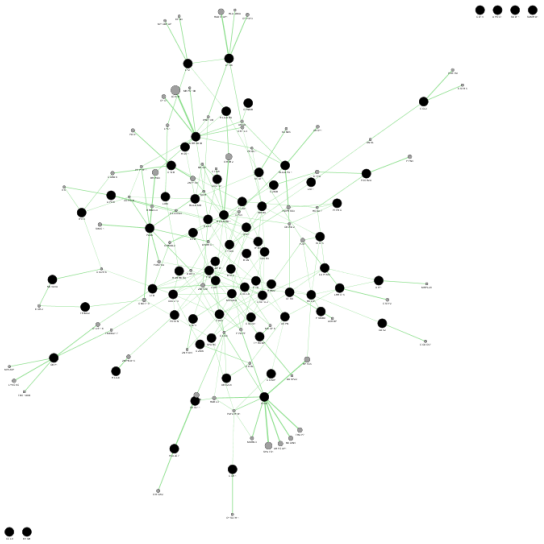
(c) Genetic interaction network

Figure A.10: Gene networks visualizing a) physical interaction, b) pathways and c) genetic interactions between genes from topGL-pos and the same amount of functionally similar genes (n = 32).



(a) Physical interaction network

(b) Pathway network



(c) Genetic interaction network

Figure A.11: Gene networks visualizing a) physical interaction, b) pathways and c) genetic interactions between genes from topGL-neg and the same amount of functionally similar genes (n = 148).

Bibliography

[1] Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. European journal of cancer (Oxford, England : 1990). 2018 Nov;103:356–387.

[2] 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 Jan;25:16–27.

[3] Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. European journal of cancer (Oxford, England : 1990). 2018 Nov;103:356–387.

[4] Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Zappa M, Nelen V, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet (London, England). 2014 Dec;384:2027–2035.

[5] Cancer Research UK. Prostate cancer survival statistics; 2018. Accessed: 2018-01-14. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/survival#heading-Four.

[6] 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 Mar;39:243–253.

[7] Stanford JL, Feng Z, Hamilton AS, Gilliland FD, Stephenson RA, Eley JW, et al. Urinary and sexual function after radical prostatectomy for clinically localized prostate cancer: the Prostate Cancer Outcomes Study. JAMA. 2000 Jan;283:354–360.

[8] Bul M, Zhu X, Valdagni R, Pickles T, Kakehi Y, Rannikko A, et al. Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. European urology. 2013 Apr;63:597–603.

[9] Wiklund F. Prostate cancer genomics: can we distinguish between indolent and fatal disease using genetic markers? Genome medicine. 2010 Jul;2:45.

[10] Daskivich TJ, Chamie K, Kwan L, Labo J, Palvolgyi R, Dash A, et al. Overtreatment of men with low-risk prostate cancer and significant comorbidity. Cancer. 2011 May;117:2058–2066.

[11] Pezaro C, Woo HH, Davis ID. Prostate cancer: measuring PSA. Internal medicine journal. 2014 May;44:433–440.

[12] Heijnsdijk EAM, Wever EM, Auvinen A, Hugosson J, Ciatto S, Nelen V, et al. Quality-of-life effects of prostate-specific antigen screening. The New England journal of medicine. 2012 Aug;367:595–605.

[13] 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 Dec;120:3642–3650.

[14] Heijnsdijk EAM, Bangma CH, Borràs JM, de Carvalho TM, Castells X, Eklund M, et al. Summary statement on screening for prostate cancer in Europe. International journal of cancer. 2018 Feb;142:741–746.

[15] Loeb S, Bjurlin MA, Nicholson J, Tammela TL, Penson DF, Carter HB, et al. Overdiagnosis and overtreatment of prostate cancer. European urology. 2014 Jun;65:1046–1055.

[16] European Prostate Cacner Awareness Day. Improving Prostate Cancer Care - Costs;. Accessed: 2019-06-17. http://epad.uroweb.org/.

[17] Thomsen FB, Berg KD, Røder MA, Iversen P, Brasso K. Active surveillance for localized prostate cancer: an analysis of patient contacts and utilization of healthcare resources. Scandinavian journal of urology. 2015 Feb;49:43–50.

[18] DKFZ. Prostatakrebs: Früherkennung und PSA-Test; 2020. Accessed: 06-03-2020. https://www.krebsinformationsdienst.de/tumorarten/prostatakrebs/psa-test-frueherkennung.php.

[19] Walsh AL, Considine SW, Thomas AZ, Lynch TH, Manecksha RP. Digital rectal examination in primary care is important for early detection of prostate cancer: a retrospective cohort analysis study. The British journal of general practice : the journal of the Royal College of General Practitioners. 2014 Dec;64:e783–e787.

[20] Naji L, Randhawa H, Sohani Z, Dennis B, Lautenbach D, Kavanagh O, et al. Digital Rectal Examination for Prostate Cancer Screening in Primary Care: A Systematic Review and Meta-Analysis. Annals of family medicine. 2018 Mar;16:149–154.

[21] Kohaar I, Petrovics G, Srivastava S. A Rich Array of Prostate Cancer Molecular Biomarkers: Opportunities and Challenges. International journal of molecular sciences. 2019 Apr;20.

[22] Alam S, Tortora J, Staff I, McLaughlin T, Wagner J. Prostate cancer genomics: comparing results from three molecular assays. The Canadian journal of urology. 2019 Jun;26:9758–9762.

[23] Khan MA, Sokoll LJ, Chan DW, Mangold LA, Mohr P, Mikolajczyk SD, et al. Clinical utility of proPSA and "benign" PSA when percent free PSA is less than 15 Urology. 2004 Dec;64:1160–1164.

[24] Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ (Clinical research ed). 2018 Sep;362:k3519.

[25] 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(6).

[26] 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 Feb;73(2):168–175.

[27] 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 Dec;190(6):2047–2053.

[28] 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 Jun;35(18):1982–1990.

[29] 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 Apr;90:148–152.

[30] Van den Broeck T, Moris L, Gevaert T, Tosco L, Smeets E, Fishbane N, et al. Validation of the Decipher Test for Predicting Distant Metastatic Recurrence in Men with High-risk Nonmetastatic Prostate Cancer 10 Years After Surgery. European urology oncology. 2019 Sep;2:589–596.

[31] 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 Mar;12(3):245–255.

[32] Cuzick J, Berney DM, Fisher G, Mesher D, Møller H, Reid JE, et al. Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. British journal of cancer. 2012 Mar;106:1095–1099.

[33] Cooperberg MR, Simko JP, Cowan JE, Reid JE, Djalilvand A, Bhatnagar S, et al. Validation of a cell-cycle progression gene panel to improve risk stratification in a

contemporary prostatectomy cohort. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013 Apr;31:1428–1434.

[34] 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(1):690.

[35] Cullen J, Rosner IL, Brand TC, Zhang N, Tsiatis AC, Moncur J, et al. A Biopsy-based 17-gene Genomic Prostate Score Predicts Recurrence After Radical Prostatectomy and Adverse Surgical Pathology in a Racially Diverse Population of Men with Clinically Low- and Intermediate-risk Prostate Cancer. European urology. 2015 Jul;68:123–131.

[36] Eggener S, Karsh LI, Richardson T, Shindel AW, Lu R, Rosenberg S, et al. A 17-gene Panel for Prediction of Adverse Prostate Cancer Pathologic Features: Prospective Clinical Validation and Utility. Urology. 2019 Apr;126:76–82.

[37] Donovan MJ, Noerholm M, Bentink S, Belzer S, Skog J, O'Neill V, et al. A molecular signature of PCA3 and ERG exosomal RNA from non-DRE urine is predictive of initial prostate biopsy result. Prostate cancer and prostatic diseases. 2015 Dec;18:370–375.

[38] McKiernan J, Donovan MJ, O'Neill V, Bentink S, Noerholm M, Belzer S, et al. A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy. JAMA oncology. 2016 Jul;2:882–889.

[39] McKiernan J, Donovan MJ, Margolis E, Partin A, Carter B, Brown G, et al. A Prospective Adaptive Utility Trial to Validate Performance of a Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer in Patients with Prostate-specific Antigen 2-10ng/ml at Initial Biopsy. European urology. 2018 Dec;74:731–738.

[40] BioTechne. Bio-Techne Receives Approval To Offer The ExoDx Prostate Intelliscore (EPI) Test In New York State; 2019. Accessed: 2020-01-12. https://investors.bio-techne.com/press-releases/detail/143/bio-techne-receives-approval-to-offer-the-exodx-prostate.

[41] Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. 2004 Mar;17:292–306.

[42] Cheng L, Montironi R, Bostwick DG, Lopez-Beltran A, Berney DM. Staging of prostate cancer. Histopathology. 2012 Jan;60:87–117.

[43] Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P. Prognostic value of the Gleason score in prostate cancer. BJU international. 2002 Apr;89(6):538–542.

[44] American Joint Committee on Cancer. Cancer Staging System; 2019. Accessed: 2019-04-10. https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx.

[45] Gleason DF. Classification of prostatic carcinomas. Cancer chemotherapy reports. 1966 Mar;50:125–128.

[46] Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. European urology. 2016 Mar;69:428–435.

[47] Epstein JI, Allsbrook WC, Amin MB, Egevad LL, Committee IG. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. The American journal of surgical pathology. 2005 Sep;29:1228–1242.

[48] Chan TY, Partin AW, Walsh PC, Epstein JI. Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy. Urology. 2000 Nov;56:823–827.

[49] Lau WK, Blute ML, Bostwick DG, Weaver AL, Sebo TJ, Zincke H. Prognostic factors for survival of patients with pathological Gleason score 7 prostate cancer: differences in outcome between primary Gleason grades 3 and 4. The Journal of urology. 2001 Nov;166:1692–1697.

[50] Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. BJU international. 2013 May;111:753–760.

[51] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Annals of surgical oncology. 2010 Jun;17:1471–1474.

[52] Cancer Research UK. TNM Staging; 2019. Accessed: 2019-04-10. https://www.cancerresearchuk.org/about-cancer/prostate-cancer/stages/tnm-staging.

[53] Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163(4):1011–1025.

[54] Tolkach Y, Kristiansen G. The Heterogeneity of Prostate Cancer: A Practical Approach. Pathobiology : journal of immunopathology, molecular and cellular biology. 2018;85:108–116.

[55] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, NY). 2005 Oct;310:644–648.

[56] Kaffenberger SD, Barbieri CE. Molecular subtyping of prostate cancer. Current opinion in urology. 2016 May;26:213–218.

[57] Arora K, Barbieri CE. Molecular Subtypes of Prostate Cancer. Current oncology reports. 2018 Jun;20:58.

[58] 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 May;25(5):745–9.

[59] 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 Aug;56:275–286.

[60] Perner S, Mosquera JM, Demichelis F, Hofer MD, Paris PL, Simko J, et al. TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion. The American journal of surgical pathology. 2007 Jun;31:882–888.

[61] Mehra R, Tomlins SA, Yu J, Cao X, Wang L, Menon A, et al. Characterization of TMPRSS2-ETS gene aberrations in androgen-independent metastatic prostate cancer. Cancer research. 2008 May;68:3584–3590.

[62] Demichelis F, Fall K, Perner S, Andrén O, Schmidt F, Setlur SR, et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. Oncogene. 2007 Jul;26:4596–4599.

[63] Sanguedolce F, Cormio A, Brunelli M, D'Amuri A, Carrieri G, Bufo P, et al. Urine TMPRSS2: ERG Fusion Transcript as a Biomarker for Prostate Cancer: Literature Review. Clinical genitourinary cancer. 2016 Apr;14:117–121.

[64] Tomlins SA, Aubin SMJ, Siddiqui J, Lonigro RJ, Sefton-Miller L, Miick S, et al. Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. Science translational medicine. 2011 Aug;3:94ra72.

[65] 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;.

[66] Gyles C. The DNA revolution. The Canadian veterinary journal = La revue veterinaire canadienne. 2008 Aug;49:745–746.

[67] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015 Oct;526:68–74.

[68] International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct;449:851–861.

[69] Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107:1–8.

[70] Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nature reviews Cancer. 2017 Nov;17:692–704.

[71] Benafif S, Kote-Jarai Z, Eeles RA, Consortium P. A Review of Prostate Cancer Genome-Wide Association Studies (GWAS). Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2018 Aug;27:845–857.

[72] Pierce BL, Ahsan H. Case-only genome-wide interaction study of disease risk, prognosis and treatment. Genetic epidemiology. 2010 Jan;34:7–15.

[73] Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nature reviews Genetics. 2003 Aug;4:587–597.

[74] Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature reviews Genetics. 2018 Aug;19:491–504.

[75] Gann PH. Risk factors for prostate cancer. Reviews in urology. 2002;4 Suppl 5:S3–S10.

[76] Rebbeck TR. Prostate Cancer Genetics: Variation by Race, Ethnicity, and Geography. Seminars in radiation oncology. 2017 Jan;27:3–10.

[77] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001 Jan;29:308–311.

[78] Matejcic M, Saunders EJ, Dadaev T, Brook MN, Wang K, Sheng X, et al. Germline variation at 8q24 and prostate cancer risk in men of European ancestry. Nature communications. 2018 Nov;9:4616.

[79] Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, et al. Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. PLoS genetics. 2011 May;7:e1001387.

[80] Robbins C, Torres JB, Hooker S, Bonilla C, Hernandez W, Candreva A, et al. Confirmation study of prostate cancer risk variants at 8q24 in African Americans identifies a novel risk locus. Genome research. 2007 Dec;17:1717–1722.

[81] Zeigler-Johnson CM, Walker AH, Mancke B, Spangler E, Jalloh M, McBride S, et al. Ethnic differences in the frequency of prostate cancer susceptibility alleles at SRD5A2 and CYP3A4. Human heredity. 2002;54:13–21.

[82] Paiss T, Wörner S, Kurtz F, Haeussler J, Hautmann RE, Gschwend JE, et al. Linkage of aggressive prostate cancer to chromosome 7q31-33 in German prostate cancer families. European journal of human genetics : EJHG. 2003 Jan;11:17–22.

[83] Eeles R, Goh C, Castro E, Bancroft E, Guy M, Al Olama AA, et al. The genetic epidemiology of prostate cancer and its clinical implications. Nature reviews Urology. 2014 Jan;11:18–31.

[84] Lloyd T, Hounsome L, Mehay A, Mee S, Verne J, Cooper A. Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008-2010. BMC medicine. 2015 Jul;13:171.

[85] Blackburn J, Vecchiarelli S, Heyer EE, Patrick SM, Lyons RJ, Jaratlerdsiri W, et al. TMPRSS2-ERG fusions linked to prostate cancer racial health disparities: A focus on Africa. The Prostate. 2019 Jul;79:1191–1196.

[86] Dong J, Xiao L, Sheng L, Xu J, Sun ZQ. TMPRSS2:ETS fusions and clinicopathologic characteristics of prostate cancer patients from Eastern China. Asian Pacific journal of cancer prevention : APJCP. 2014;15:3099–3103.

[87] Kong D, Chen R, Zhang C, Zhang W, Xiao G, Wang F, et al. Prevalence and clinical application of TMPRSS2-ERG fusion in Asian prostate cancer patients: a large-sample study in Chinese people and a systematic review. Asian Journal of Andrology. 2020;22:200–207.

[88] Powell IJ, Dyson G, Chinni SR, Bollig-Fischer A. Considering race and the potential for ERG expression as a biomarker for prostate cancer. Personalized medicine. 2014;11:409–412.

[89] TCGA consortium. TCGA history;. Accessed: 2019-10-11. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline.

[90] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. The New England journal of medicine. 2016 Sep;375:1109–1112.

[91] National Cancer Institute. Genomic Data Commons (GDC) Data Portal; 2018. Accessed: 09-2018. https://portal.gdc.cancer.gov.

[92] International Cancer Genome Consortium. PanCancer Analysis of Whole Genomes; 2019. Accessed: 11-10-2019. https://dcc.icgc.org/pcawg.

[93] International Cancer Genome Consortium. ICGC home; 2019. Accessed: 11-10-2019. http://icgc.org.

[94] International Cancer Genome Consortium. ICGC Data Portal; 2019. Accessed: 07-2019. http://dcc.icgc.org.

[95] Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. Nature. 2017 Jan;541:359–364.

[96] 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 Mar;3:8–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20233430.

[97] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007 Sep;81:559–575.

[98] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

[99] R Core Team. R: A Language and Environment for Statistical Computing. 2018;Version 3.5.0. Available from: https://www.r-project.org/.

[100] Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal. 2016;8(1):205–233.

[101] 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.

[102] 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 Aug;30(15):2224–6.

[103] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England). 2012 Sep;28:i333–i339.

[104] 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. International Journal of Cancer. 2013 Feb;132(4):807–812.

[105] Gerke JS, Orth MF, Tolkach Y, Romero-Pérez L, Wehweck FS, Stein S, et al. Integrative clinical transcriptome analysis reveals TMPRSS2-ERG dependency of prognostic biomarkers in prostate adenocarcinoma. International journal of cancer. 2020 Apr;146:2036–2046.

[106] Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012 Dec;100:337–344.

[107] Carvalho B. pd.huex.1.0.st.v2: Platform Design Info for Affymetrix HuEx-1_0-st-v2; 2015.

[108] 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 Nov;33:e175.

[109] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009 Jan;10:57–63.

[110] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research. 2010 Oct;38:e178.

[111] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011 Aug;12:323.

[112] Sathianathen NJ, Konety BR, Crook J, Saad F, Lawrentschuk N. Landmarks in prostate cancer. Nature Reviews Urology. 2018 Oct;15(10):627–642.

[113] Gentleman R, Carey V, Huber W, Hahne F. genefilter: methods for filtering genes from high-throughput experiments. 2017;(R package version 1.58.1).

[114] Macherey-Nagel. Genomic DNA from Tissue - User manual; 2017. https://www.mn-net.com/Portals/8/attachments/Redakteure_Bio/Protocols/Genomic%20DNA/UM_gDNATissue_2017.pdf.

[115] Illumina. Infinium Global Screening Array; 2018. Accessed: 2018-08. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-commercial-gsa-data-sheet-370-2016-016.pdf.

[116] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Current protocols in human genetics. 2011 Jan;Chapter 1:Unit1.19.

[117] GATK Team. GATK Best Practice Workflow: Germline short variant discovery (SNPs + Indels); 2019. Accessed: 2019-07-15. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels.

[118] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011 Aug;27:2156–2158.

[119] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009 Aug;25:2078–2079.

[120] Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338.

[121] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England). 2016 Oct;32:3047–3048.

[122] Broad Institute. Picard Tools (v2.18); 2018. http://broadinstitute.github.io/picard/.

[123] Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017;.

[124] Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Frontiers in genetics. 2012;3:35.

[125] Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature genetics. 2016 Oct;48:1284–1287.

[126] McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016 Oct;48:1279–1283.

[127] Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016 Nov;48:1443–1448.

[128] Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PloS one. 2012;7:e50610.

[129] 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. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct;102:15545–15550.

[130] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics (Oxford, England). 2007 Dec;23:3251–3253.

[131] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems. 2015 Dec;1:417–425.

[132] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003 Nov;13:2498–2504.

[133] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research. 2010 Jul;38:W214–W220.

[134] Therneau T. A Package for Survival Analysis in S; 2015. Version 2.38. Available from: https://CRAN.R-project.org/package=survival.

[135] Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. The Journal of Open Source Software. 2017;.

[136] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics (Oxford, England). 2010 Sep;26:2190–2191.

[137] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics (Oxford, England). 2010 Sep;26:2336–2337.

[138] Gordon M, Lumley T. forestplot: Advanced Forest Plot Using 'grid' Graphics; 2019. Accessed: 12-12-2019. https://cran.r-project.org/web/packages/forestplot/index.html.

[139] Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evidence-based mental health. 2019 Nov;22:153–160.

[140] Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics (Oxford, England). 2015 Apr;31:1334–1336.

[141] Eklund A. The Bee Swarm Plot, an Alternative to Stripchart. 2016;(R package version 0.2.3).

[142] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489:57–74.

[143] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013 Mar;14:178–192.

[144] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome biology. 2016 Jun;17:122.

[145] Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. 2019;(R package version 1.46).

[146] Morgan M, Ramos M. BiocManager: Access the Bioconductor Project Package Repository. 2018;(R package version 1.30.4).

[147] Carlson M. org.Hs.eg.db: Genome wide annotation for Human. 2019;(R package version 3.8.2).

[148] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England). 2007 Jan;8:118–127.

[149] Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC bioinformatics. 2015 Feb;16:63.

[150] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011 Jan;29:24–26.

[151] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). 2003 Apr;4:249–264.

[152] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England). 2003 Jan;19:185–193.

[153] Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature Genetics. 2012 May;44(6):685–689.

[154] Andrén O, Fall K, Franzén L, Andersson SO, Johansson JE, Rubin MA. How well does the Gleason score predict prostate cancer death? A 20-year followup of a population based cohort in Sweden. The Journal of Urology. 2006 Apr;175(4):1337–1340.

[155] Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer. 2008 Dec;113(11):3075–3099.

[156] 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(12):1371–1378.

[157] GTEx Portal. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal and dbGaP accession number phs000424.v8.p2; 2019. Accessed: 2019-12-02. https://www.gtexportal.org/.

[158] GTEx Consortium and Laboratory DACCLWG, groups—Analysis Working Group SM, (eGTEx) groups EG, Fund NC, NIH/NCI, NIH/NHGRI, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550:204–213.

[159] Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nature reviews Genetics. 2011 Jan;12:7–18.

[160] Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Molecular cell. 2013 Mar;49:825–837.

[161] 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 Jan;44(D1):D733–D745.

[162] Niciu MJ, Kelmendi B, Sanacora G. Overview of glutamatergic neurotransmission in the nervous system. Pharmacology, biochemistry, and behavior. 2012 Feb;100:656–664.

[163] Merkle D, Hoffmann R. Roles of cAMP and cAMP-dependent protein kinase in the progression of prostate cancer: cross-talk with the androgen receptor. Cellular signalling. 2011 Mar;23:507–515.

[164] Bang YJ, Kim SJ, Danielpour D, O'Reilly MA, Kim KY, Myers CE, et al. Cyclic AMP induces transforming growth factor beta 2 gene expression and growth arrest in the human androgen-independent prostate carcinoma cell line PC-3. Proceedings of the National Academy of Sciences of the United States of America. 1992 Apr;89:3556–3560.

[165] Macchia V, Di Carlo A, De Luca C, Mariano A. Effects of cyclic adenosine-monophosphate on growth and PSA secretion of human prostate cancer cell line. International journal of oncology. 2001 May;18:1071–1076.

[166] Heinlein CA, Chang C. Androgen receptor in prostate cancer. Endocrine reviews. 2004 Apr;25:276–308.

[167] Pissimissis N, Papageorgiou E, Lembessis P, Armakolas A, Koutsilieris M. The glutamatergic system expression in human PC-3 and LNCaP prostate cancer cells. Anticancer research. 2009 Jan;29:371–377.

[168] Koochekpour S, Majumdar S, Azabdaftari G, Attwood K, Scioneaux R, Subramani D, et al. Serum glutamate levels correlate with Gleason score and glutamate

blockade decreases proliferation, migration, and invasion and induces apoptosis in prostate cancer cells. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012 Nov;18:5888–5901.

[169] Koochekpour S. Glutamate, a metabolic biomarker of aggressiveness and a potential therapeutic target for prostate cancer. Asian journal of andrology. 2013 Mar;15:212–213.

[170] Ali S, Shourideh M, Koochekpour S. Identification of novel GRM1 mutations and single nucleotide polymorphisms in prostate cancer cell lines and tissues. PloS one. 2014;9:e103204.

[171] da Costa PJ, Menezes J, Romão L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. The international journal of biochemistry & cell biology. 2017 Oct;91:168–175.

[172] 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 Mar;116(6):775–784.

[173] Jacobsen F, Kraft J, Schroeder C, Hube-Magg C, Kluth M, Lang DS, et al. Upregulation of Biglycan is Associated with Poor Prognosis and PTEN Deletion in Patients with Prostate Cancer. Neoplasia. 2017 Sep;19(9):707–715.

[174] Zhang Z, Wang Y, Zhang J, Zhong J, Yang R. COL1A1 promotes metastasis in colorectal cancer by regulating the WNT/PCP pathway. Molecular medicine reports. 2018 Apr;17:5037–5042.

[175] Brooks M, Mo Q, Krasnow R, Ho PL, Lee YC, Xiao J, et al. Positive association of collagen type I with non-muscle invasive bladder cancer progression. Oncotarget. 2016 Dec;7:82609–82619.

[176] Li J, Ding Y, Li A. Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer. World journal of surgical oncology. 2016 Nov;14:297.

[177] 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 Jan;16(1):43–9.

[178] 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(5):1444–1456.

[179] Mazzu YZ, Armenia J, Chakraborty G, Yoshikawa Y, Coggins SA, Nandakumar S, et al. A Novel Mechanism Driving Poor-Prognosis Prostate Cancer: Overexpression of the DNA Repair Gene, Ribonucleotide Reductase Small Subunit M2 (RRM2). Clinical cancer research : an official journal of the American Association for Cancer Research. 2019 Jul;25:4480–4492.

[180] Kosari F, Munz JMA, Savci-Heijink CD, Spiro C, Klee EW, Kube DM, et al. Identification of prognostic biomarkers for prostate cancer. Clinical cancer research: an official journal of the American Association for Cancer Research. 2008 Mar;14:1734–1743.

[181] Saramäki OR, Harjula AE, Martikainen PM, Vessella RL, Tammela TLJ, Visakorpi T. TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008 Jun;14:3395–3400.

[182] 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 Apr;6(10):8377–87.

[183] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ (Clinical research ed). 2003 Sep;327:557–560.

[184] VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theoretical population biology. 2008 Aug;74:130–137.

List of Tables

List of Figures

Publications

In the following publications, I participated during my work as PhD student. The peer reviewed original articles are listed by date of publication.

Marchetto A, Ohmura S, Orth MF, Knott MML, Colombo MV, Arrigoni C, Bardinet V, Saucier D, Wehweck FS, Li J, Stein S, **Gerke JS**, Baldauf MC, Musa J, Dallmayer M, Romero-Pérez L, Hölting TLB, Amatruda JF, Cossarizza A, Henssen AG, Kirchner T, Moretti M, Cidre-Aranaz F, Sannino G, Grünewald TGP
Oncogenic hijacking of a developmental transcription factor evokes vulnerability toward oxidative stress in Ewing sarcoma.
Nature communications, 2020

Orth MF, Hölting TLB, Dallmayer M, Wehweck FS, Paul T, Musa J, Baldauf MC, Surdez D, Delattre O, Knott MML, Romero-Pérez L, Kasan M, Cidre-Aranaz F, **Gerke JS**, Ohmura S, Li J, Marchetto A, Henssen AG, Özen Ö, Sugita S, Hasegawa T, Kanaseki T, Bertram S, Dirksen U, Hartmann W, Kirchner T, Grünewald TGP
High specificity of BCL11B and GLG1 for EWSR1-FLI1 and EWSR1-ERG positive Ewing sarcoma.
Cancers, 2020

Gerke JS, Orth MF, Tolkach Y, Romero-Pérez L, Wehweck FS, Stein S, Musa J, Knott MML, Hölting TLB, Li J, Sannino G, Marchetto A, Ohmura S, Cidre-Aranaz F, Müller-Nurasyid M, Strauch K, Stief C, Kristiansen G, Kirchner T, Buchner A, Grünewald TGP
Integrative clinical transcriptome analysis reveals TMPRSS2-ERG dependency of prognostic biomarkers in prostate adenocarcinoma.
International Journal of Cancer, 2020

Steinestel K, Trautmann M, Jansen EP, Dirksen U, Rehkämper J, Mikesch JH, **Gerke JS**, Orth MF, Sannino G, Arteaga MF, Rossig C, Wardelmann E, Grünewald TGP, Hartmann W

Focal adhesion kinase confers pro-migratory and antiapoptotic properties and is a potential therapeutic target in Ewing sarcoma.

Molecular Oncology, 2020

Musa J, Cidre-Aranaz F, Aynaud MM, Orth MF, Knott MML, Mirabeau O, Mazor G, Varon M, Hölting TLB, Grossetête S, Gartlgruber M, Surdez D, **Gerke JS**, Ohmura S, Marchetto A, Dallmayer M, Baldauf MC, Stein S, Sannino G, Li J, Romero-Pérez L, Westermann F, Hartmann W, Dirksen U, Gymrek M, Anderson ND, Shlien A, Rotblat B, Kirchner T, Delattre O, Grünewald TGP

Cooperation of cancer drivers with regulatory germline variants shapes clinical outcomes.

Nature Communications, 2019

Sannino G, Marchetto A, Ranft A, Jabar S, Zacherl C, Alba-Rubio R, Stein S, Wehweck FS, Kiran MM, Hölting TLB, Musa J, Romero-Pérez L, Cidre-Aranaz F, Knott MML, Li J, Jürgens H, Sastre A, Alonso J, Da Silveira W, Hardiman G, **Gerke JS**, Orth MF, Hartmann W, Kirchner T, Ohmura S, Dirksen U, Grünewald TGP

Gene expression and immunohistochemical analysis identify SOX2 as major risk factor for overall survival and relapse in Ewing sarcoma patients.

EBioMedicine, 2019

Dallmayer M, Li J, Ohmura S, Alba Rubio R, Baldauf MC, Hölting TLB, Musa J, Knott MML, Stein S, Cidre-Aranaz F, Wehweck FS, Romero-Pérez L, **Gerke JS**, Orth MF, Marchetto A, Kirchner T, Bach H, Sannino G, Grünewald TGP

Targeting the CALCB/RAMP1 axis inhibits growth of Ewing sarcoma.

Cell Death & Disease, 2019

Orth MF, **Gerke JS**, Knösel T, Altendorf-Hofmann A, Musa J, Alba-Rubio R, Stein S, Hölting TLB, Cidrea-Aranaz F, Romero-Pérez L, Dallmayer M, Baldauf MC, Marchetto

A, Sannino G, Knott MML, Wehweck F, Ohmura S, Li J, Hakozaki M, Kirchner T, Dandekar T, Butt E, Grünewald TGP
Functional genomics identifies AMPD2 as a new prognostic marker for undifferentiated pleomorphic sarcoma.
International Journal of Cancer, 2019

Baldauf MC* and **Gerke JS***, Kirschner A, Blaeschke F, Effenberger M, Schober K, Alba Rubio R, Kanaseki T, Kiran MM, Dallmayer M, Musa J, Akpolat N, Akatli AN, Rosman FC, Özen Ö, Sugita S, Hasegawa T, Sugimura H, Baumhoer D, Knott MML, Sannino G, Marchetto A, Li J, Busch DH, Feuchtinger T, Ohmura S, Orth MF, Thiel U, Kirchner T, Grünewald TGP
Systematic identification of cancer-specific MHC-binding peptides with RAVEN.
Oncoimmunology, 2018

Baldauf MC, **Gerke JS**, Orth MF, Dallmayer M, Baumhoer D, de Alava E, Hartmann W, Kirchner T, Grünewald TGP
Are EWSR1-NFATc2-positive sarcomas really Ewing sarcomas?
Modern Pathology, 2018

Baldauf MC, Orth MF, Dallmayer M, Marchetto A, **Gerke JS**, Alba Rubio R, Kiran MM, Musa J, Knott MML, Ohmura S, Li J, Akpolat N, Akatli AN, Özen Ö, Dirksen U, Hartmann W, de Alava E, Baumhoer D, Sannino G, Kirchner T, Grünewald TGP
Robust diagnosis of Ewing sarcoma by immunohistochemical detection of super-enhancer-driven EWSR1-ETS targets.
Oncotarget, 2018

Kirschner A, Thiede M, Blaeschke F, Richter GHS, **Gerke JS**, Baldauf MC, Grünewald TGP, Busch DH, Burdach S, Thiel U
Lysosome-associated membrane glycoprotein 1 predicts fratricide amongst T cell receptor transgenic CD8+ T cells directed against tumor-associated antigens.
Oncotarget, 2016

Acknowledgements

First, I would like to express my deepest appreciation to my supervisor PD Dr. Dr. med. Thomas Grünewald, principal investigator of the Max-Eder Research Group for Pediatric Sarcoma Biology of the Pathological Institute of the Ludwig-Maximilian-University of Munich (LMU) for his encouragement, motivation, enthusiastic support, and advice. Without his guidance and adjuvant feedback this PhD thesis would not have been possible.

I greatly appreciate the support and inspiring discussions with my supervisor PD Dr. Dr. med. Thomas Grünewald, Prof. Dr. Konstantin Strauch (Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) of the University Medical Center of the Johannes Gutenberg University Mainz), Dr. Martina Müller-Nurasyid (Institute of Genetic Epidemiology, Helmholtz Center Munich), and Prof. Dr. med. Alexander Buchner (Urologic Clinic and Polyclinic of the University Munich). I would like to pay my special regards to the members of my thesis advisory committee, PD Dr. Dr. med. Thomas Grünewald, Prof Dr. Konstantin Strauch and Dr. Gabi Kastenmüller (Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich)

I would like to extend my thanks to Prof. Dr. Konstantin Strauch and both of his teams (from the Institute of Genetic Epidemiology, Helmholtz Center Munich and the Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz) for the collaboration regarding the prostate cancer study $PCGA^{LMU}$. Specifically, I am grateful to Prof. Dr. med. Alexander Buchner from the Urologic Clinic and Polyclinic of the University Munich for assembling the study participants and his tireless commitment, without whom the $PCGA^{LMU}$ would not have been possible.

I wish to show my gratitude to PD Dr. med. Yuri Tolkach for providing and evaluating the data of the TMA-cohort.

Many thanks to Dr. med. Martin Orth, Stefanie Stein, Rebeca Alba Rubio, Dr. Laura Romero Pérez, and Dr. med. Fabienne Wehweck for supporting my projects with their wet lab experiments and their pathological expertise. Furthermore, I want to thank all my colleagues of the Max-Eder Research Group for Pediatric Sarcoma Biology of the Pathological Institute of the LMU as well as my supervisor for four great years of work and research and all our adventures together.

I thank Prof. Dr. med. Thomas Kirchner for giving me the opportunity to conduct my research projects at the Pathological Institute of the LMU. Also, I want to thank Andrea Sendelhofert and Anja Heier for their technical support.

For proof-reading, inspiring discussions and their helpful comments on my manuscripts prior to submission, I want to thank Valentina Klaus and Dr. med. Martin Orth.

In the end, I wish to thank my partner for his patience and his love. I am deeply grateful to my parents, who supported me all the time. Thank you for your love all the encouragement.

Affidavit

Gerke, Julia Sophia
Thalkirchner Str. 36
80337 Munich
Germany

I hereby declare, that the submitted thesis entitled

Assessment of the contribution of germline variation and somatic mutations to prostate cancer progression and prognostication

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Munich, 03.03.2021	Julia Sophia Gerke
Place, date	Signature doctoral candidate

Confirmation of congruency between printed and electronic version of the doctoral thesis

Gerke, Julia Sophia
Thalkirchner Str. 36
80337 Munich
Germany

I hereby declare that the electronic version of the submitted thesis, entitled

Assessment of the contribution of germline variation and somatic mutations to prostate cancer progression and prognostication

is congruent with the printed version both in content and format.

Munich, 03.03.2021	Julia Sophia Gerke
Place, date	Signature doctoral candidate