Identification of Novel Prostate Cancer Biomarkers Using High-throughput Technologies

René Böttcher

Identification of Novel Prostate Cancer Biomarkers Using High-throughput Technologies

Identificatie van nieuwe prostaatkanker biomarkers met behulp van high-throughput technologieën

Proefschrift

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van de

rector magnificus

Prof.dr H.A.P. Pols

en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op

Dinsdag, 22.11.2016 om 09:30 uur.

door

René Böttcher

geboren te Berlijn, Duitsland

Erasmus University Rotterdam

Promotiecommissie

Promotor	Prof.dr.ir. G.W. Jenster
Overige leden	Prof.dr. P.J. van der Spek
	Prof.dr. R. Agami
	Dr. G. van der Pluijm
Copromotor	Prof.dr. P. Beyerlein

The work described in this thesis was conducted at the Department of Urology of Erasmus Medical Center Rotterdam, The Netherlands.

The research in this thesis was financially supported by the Center for Translational Molecular Medicine (CTMM) PCMM (grant 03O-203) and NGS ProToCol projects (03O-402).

The printing of this thesis was financially supported by the Stichting Wetenschappelijk Onderzoek Prostaatkanker, Rotterdam (SWOP).

Cover design by Anca Pora (http://ancapora.daportfolio.com/)

Printing and binding by Ridderprint BV, Ridderkerk

Erasmus University Rotterdam

Contents

Chapter 1 **9-24**
General introduction and scope of the thesis
Chapter 2 **25-50**
Long Noncoding RNA in Prostate, Bladder, and Kidney Cancer .
Published in Eur Urol. 2014;65(6):1140-51
Chapter 3 **51-76**
Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer.
Published in Oncotarget. 2015;6(6):4036-50
Chapter 4 **77-100**
Human phosphodiesterase 4D7 (PDE4D7) expression is increased in TMPRSS2-ERG positive primary prostate cancer and independently adds to a reduced risk of post- surgical disease progression.
Published in Br J Cancer. 2015;113(10):1502-11
Chapter 5 **101-128**
Human PDE4D isoform composition is deregulated in primary prostate cancer and indicative for disease progression and development of distant metastases.
Oncotarget, in press
Chapter 6 **129-150**
Using a priori knowledge to align sequencing reads to their exact genomic position.
Published in Nucleic Acids Res. 2012;40(16):e125
Chapter 7 **151-170**
General discussion
Appendices
Summary / Samenvatting **167-173**
Curriculum Vitae **179**
List of publications **180**
PhD Portfolio **182**
Acknowledgements **184**

Chapter 1

General introduction and scope of the thesis

1.1. General introduction to prostate cancer

The prostate is a gland of the male reproductive system that is highly dependent on the androgens testosterone (T) and dihydrotestosterone (DHT) for its development and homeostasis. Prostate cancer (PCa), mostly affects men above the age of fifty and has been associated with 'Western' lifestyle and diet (1, 2). PCa is the most frequently occurring gender-specific carcinoma for men, with an estimated 417,000 new cases and 70,100 cancer-related deaths in Europe in 2014 (3, 4). As these numbers indicate, a large discrepancy between diagnosed cases and fatalities exists. Many of the detected tumors are growing slowly and many men die with PCa, rather than from PCa. However, men can suffer from aggressive forms that are metastasizing and require early treatment. These highly heterogeneous outcomes highlight the necessity of well-powered risk stratification to discriminate insignificant from aggressive tumors.

If localized, PCa is treated with curative intent and clinical protocols usually involve either surgical removal of the prostate via radical prostatectomy (RP) or a radiation-based therapy (5). In case of tumor-regrowth after RP, a systemic therapy is applied to prevent the tumor from growing and spreading further (6). Since prostate cells require T or DHT for their growth, the androgen receptor (AR) pathway plays a crucial role in PCa development and progression. Androgen deprivation therapy (ADT) has been established as the standard treatment strategy for metastatic PCa (7, 8). During ADT, androgen production is actively suppressed via surgical or chemical castration, leading to a reduction in hormone levels and tumor size (7). However, ADT is criticized as provisional treatment strategy due to the inevitable occurrence of castration-resistant prostate cancer (CRPC) after treatment, a phenotype that no longer relies on normal androgen blood levels (9, 10). Once PCa has become castration-resistant, it is highly unlikely to be cured and treatment options mainly focus on chemotherapy and second line hormone therapies as palliative care. Here, it is important to note that although CRPC does not respond to ADT, AR signaling is still active and can be acted on (11). For this reason, potent anti-androgens such as enzalutamide or abiraterone, an inhibitor of androgen synthesis that blocks CYP17, are still effective in many patients, albeit only for a limited time (12–14).

The described therapy options pose a burden on both patients and healthcare systems, and as a result, it is crucial to know upfront which patients will benefit from the different treatment regiments, often referred to as 'personalized healthcare or 'precision medicine'. Despite many efforts, assigning an appropriate treatment course remains one of the major challenges in PCa therapy, as current clinical protocols lack reliable biomarkers with sufficient performance (discussed below). Therefore, novel biomarkers with better diagnostic, prognostic and predictive potential are urgently needed.

1.2. Current and emerging strategies for diagnosing and staging prostate cancer

Usually, PCa is diagnosed via a combination of different approaches to avoid misclassification based on initial test results. In Western societies, digital rectal examination (DRE), and examining blood serum levels of the prostate-specific antigen (PSA, also known as kallikrein-3 / KLK3) provide the first indication of presence of prostate cancer (15, 16). During DRE, a urologist examines the posterior side of the prostate by rectal insertion of a finger to feel for abnormalities such as bumps, indurations or increased size (17). Since DRE is limited to physiological alterations on the dorsal side of the prostate, complimentary approaches such as PSA testing have been developed, however, discussions about the currently used cutoff values for PSA serum concentration are still continuing as additional data is published (18). Moreover, despite being a highly sensitive measure of prostate tissue growth, increased PSA serum levels are not specific to PCa and can be caused by other benign conditions such as benign prostate hyperplasia (BPH) and prostatitis (19). This lack of PCa-specificity leads to the diagnosis of insignificant tumors and subsequent treatment of patients with insignificant localized disease due to PSA testing (15, 20–22). Moreover, as it is common practice to perform a prostate biopsy upon positive DRE or abnormal PSA testing to confirm presence of PCa, a substantial number of avoidable biopsies is performed every year (20, 23, 24).

Biopsies are often guided by ultrasound-based transrectal ultrasonography (TRUS) (25) and most commonly performed transrectally, though it is also possible to access the prostate tissue through the urethra or through the perineum (26). The sampled tissue is subsequently examined microscopically by a pathologist and graded according to the Gleason grading system for prognostic evaluation (27). Here, the final Gleason score consists of two grades, with the first score representing the most common tumor pattern, while the second score represents the second most common pattern found in the sample. Both grades range from 1 to 5 indicating decreasing tissue differentiation and worsening prognosis, where a combination of 4+3=7 is considered worse than 3+4=7 (27). Although guiding technologies are improving, side effects such as general discomfort during the procedure, bleeding and infections can occur due to biopsy sampling (28, 29). Moreover, biopsies may miss the area containing tumor tissue, which can lead to a false negative patient diagnosis.

In case of a negative initial tumor biopsies, additional biopsies or a urine test for the long non-coding RNA (lncRNA) PCA3 are accredited steps to validate the initial screening results (30–32). Generally, lncRNAs are similar in structure and cellular processing to messenger RNAs of protein-coding genes, however, by definition they do not harbor a functional open reading frame that is actively translated (see Chapter 2 for more information). This family of RNAs also shows highly tissue-specific expression, which in the case of PCA3 is associated with PCa development and can be used as diagnostic biomarker (30). Other recently introduced diagnostic tools for PCa include the gene fusion TMPRSS2-ERG (31, 33), the PSA-based Prostate Health Index (PHI) and 4Kscore (34–36), a combination of TMPRSS2-ERG and PCA3 called MiPS (37), the mitochondrial DNA-based Prostate Core Mitomic Test, as well as ConfirmMDx, an epigenetic test measuring DNA methylation of three marker genes to predict the results of repeat biopsies after initial negative biopsies (38, 39).

Moreover, imaging based methods such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET) and single-photon emission computed tomography (SPECT) could be used as potential diagnostic tools. However, their application for PCa diagnosis is often limited by their soft tissue resolution or a lack of appropriate tracer compounds, and hence they are mainly used for cancer staging by (bone) metastases detection at this point in time (25), with the exception of MRI (40–42). Further methodologies aiming to improve PCa prognosis are currently under development and include radiolabelled antibodies against prostate-specific membrane antigen, PSMA (ProstaScint (43)), radiolabelled bombesin analogues (44), 18F-fluoro-deoxy-glucose (FDG) (45, 46), or targeting CXCR4 (47).

Lastly, prognostic PCa classifiers based on measuring gene panels have also been under development and some have become commercially available, such as Oncotype DX (Genomic Health Inc.), Prolaris (Myriad Genetics Inc.), Decipher (GenomeDx Biosciences Inc.), ProMark (Metamark Genetics Inc), SelectMDx (MDxHealth Inc) and ExoDx Prostate IntelliScore (Exosome Diagnostics Inc) (36, 38, 39, 48, 49). It has to be noted though, that the mentioned tests represent a number of recently released products that have not yet been widely applied in a clinical context and are still under evaluation. Other promising PCa markers yet to be clinically utilized include miRNA profiles (50–52), the PTEN tumor suppressor (23, 53) and cMYC oncogene (54–56), as well as recurring copy number alterations (57, 58).

1.3. Molecular characteristics of prostate cancer

Cancer is defined as uncontrolled proliferation of abnormal cells that is termed ‘malignant’ once the involved cells are able to cross tissue borders and spread to adjacent tissues or disseminate to distant sites via the bloodstream or lymphatic vessels. Before reaching this abnormal stage, a cell has to overcome several obstacles posed by the human body and its immune system. According to Hanahan and Weinberg (2000 (59)), features required for the formation of a cancer cell are:

- Self-sufficiency in growth signals
- Insensitivity to growth-inhibitory signals
- Evasion of programmed cell death (apoptosis)
- Limitless replicative potential
- Induction of angiogenesis
- Tissue invasion and metastasis

This list has since been expanded to include other characteristics commonly observed in cancer (60):

- Evasion of immune responses
- Deregulated metabolism
- Inflammation
- Unstable DNA

Although these general properties may appear directed, they are the result of an accumulation of somatic mutations and chromosomal abnormalities that occur throughout life and mostly lead to cell death once vital functions are disrupted. However, aberrations in specific regulatory genes may also provide selective advantages by loss-of-function of tumor suppressor genes or gain-of-function of tumor promoting genes (oncogenes). Generally, the most commonly mutated genes in cancer include JAK2, BRAF, KRAS, TP53, EGFR, FLT3 and PIK3CA, with mutation frequencies varying greatly depending on the particular type of cancer (see The Cancer Gene Census available via http://cancer.sanger.ac.uk) (61). Other known genes include APC, RB1, BRCA1, BRCA2, PTEN as well as members of the RAS and MLL gene families, totaling around 150 to 200 recurrently affected genes, depending on the source (61, 62). These genes cluster in important cellular pathways such as DNA damage control (TP53, BRCA1 and BRCA2), cell survival (PIK3/AKT/mTOR and RAS/RAF/MEK/ERK) and cell fate (APC, Notch signaling) (62).

In contrast to other cancers, prostate cancer is generally characterized by a low number of somatic point mutations (single nucleotide polymorphisms and small indels) (62). Nevertheless, recurrent mutations in approximately 80 genes have been identified, including the aforementioned TP53, PTEN, EGFR and PIK3CA (COSMIC mutation frequencies when including all PCa samples: 14%, 7%, 3% and 2%, respectively) as well as AR pathway related genes (AR, SPOP, FOXA1; 4%, 9%, 5%) (61, 63–65). In addition, amplifications of the MYC oncogene on chr8q (2%-20%) as well as deletions of tumor suppressor genes located on chr8p (NKX3-1, 35%-86%) and chr16q (CDH1, 15%–27%) are common (54, 66–71). Moreover, fusions of the AR-regulated gene TMPRSS2 and members of the ETS transcription factor family are common events, with ERG being the most frequent fusion partner (approximately 50% of all PCa cases depending on study cohort) (33). However, despite its prevalence, the prognostic potential of TMPRSS2-ERG has not yet been fully clarified and conflicting results prevent a clinical application for PCa staging if one exists (39, 72).

It is also known that cross-talk of androgen signaling and other pathways such as MAPK/ERK, Wnt signaling and phosphatidylinositol 3-kinase (PI3K)/Akt pathway can play important roles in cancer development and progression (73–80). Lately, the cyclic AMP (cAMP) cAMP-dependent pathway has been gaining interest in PCa research (81–84), after numerous studies showed association of cAMP signaling and various conditions (85–89). In humans, cAMP functions as a second messenger molecule that can activate a variety of different targets after being converted from ATP by the enzyme adenylyl cyclase. As such, cAMP is part of a signaling cascade starting at G-protein coupled receptors (GPCRs), which allows various cellular responses to extracellular signals depending on the type of GPCR and its activation by an external stimulus (90). Among the many target molecules, protein kinase A (PKA) plays an important role, as it is able to phosphorylate numerous other proteins and is thereby directly involved in the regulation of processes such as glycogen conversion, muscle contraction and transcription in a cell type specific manner (91–93). Moreover, recent evidence suggests that PKA is also able to interact with AR, hinting at cross-talk between both signaling pathways (81–83, 94).

Since accumulating cAMP would lead to continuous activation of downstream targets, its concentration is controlled by phosphodiesterases (PDEs) that hydrolyze cAMP (81, 95). Individual PDE isoforms are known to contain domains that enable precise subcellular localization, allowing a tight spatial and time-dependent regulation of cAMP gradients around specific intracellular locations (95). Due to this crucial role, PDEs are being investigated (and utilized) as therapeutic targets for many conditions, such as erectile dysfunction, stroke, brain injury and Alzheimer's disease (96–99), making them potentially interesting for prostate cancer research.

1.4. Technology developments and bioinformatics

Many of the described molecular characteristics of cancers were discovered using genome-wide screenings, with the first commercially available screening platforms being based on microarray technology. Microarrays consist of a substrate (e.g. a glass slide) that provides a surface for chemical attachment of short nucleotide sequences, the so-called probes, which are complementary to the target sequences of interest. By using labeled target sequences and hybridizing them to the probes under stringent conditions, it is possible to quantify the target concentration in a sample. Target sequences are usually DNA-based and hence a reverse transcriptase step is involved to produce complementary DNA molecules from individual RNAs. This flexible approach allowed a broad range of applications such as measuring chromosomal copy number using known biallelic single-nucleotide polymorphisms (SNPs) (100) or profiling expression of thousands of genes in parallel (101).

Further technological advances in the field of nucleotide sequencing led to a massively parallelization of sequencing reactions, which is commonly referred to as next generation sequencing (NGS) (102). Initially developed for fast and cost-effective DNA sequencing, NGS found broad application in many fields, including studies of the exome (via DNA capture technologies), transcriptome (RNA-sequencing), DNA binding proteins (e.g. using chromatin immunoprecipitation (ChIP-seq)), and epigenetic factors such as DNA methylation (e.g. bisulphite sequencing). All of these NGS variations are based on the fragmentation of longer sequences such as chromosomes or mRNA transcripts into smaller pieces. These short sequences are then amplified and sequenced in a massively parallel fashion, resulting in millions of short nucleotide stretches that span between 35 and several hundred of nucleotides, depending on the sequencing technology used. After a sequencing run, the "reads" of sequence information are aligned back to a time-stamped reference genome (e.g. human genome build hg19) to allow feature quantification and identification of mutations.

Along with sequencing technology, bioinformatics solutions for data analysis have been developed at an ever increasing pace, allowing rapid hypothesis testing in various applications. For instance, while DNA-based alignment still plays an important role in modern applications and programs are continuously evolving (103–106), introduction of RNA-seq required development of splicing-aware aligners such as GSNAP, STAR, TopHat/TopHat2 and MapSplice (107–110). To quantify gene expression, early approaches relied on read counting and were not able to distinguish different RNA isoforms (111, 112), a shortcoming that subsequent methods tried to address (113, 114). Most recent developments in RNA-seq quantification use k-mer indexing and pseudo-alignments to avoid the mapping

step altogether and thereby reduce processing time for well-annotated transcriptomes (115–117).

1.5. New research insights in prostate cancer from next generation sequencing and bioinformatics

Transcriptome sequencing not only allows quantifying the expression of known genes, but can also be used to identify novel unannotated genes. Programs such as Cufflinks (118) are able to assemble the likely gene structure by utilizing information of canonical splice sites found in sequencing reads. This feature was applied to a large cohort of RNA-seq samples from prostate cancer patients, resulting in 121 previously unknown prostate cancer-associated transcripts referred to as PCATs (119). Interestingly, several of these transcripts showed a remarkable specificity for PCa. Additional validation steps showed that many of the identified transcripts were indeed long non-coding RNAs (lncRNAs), highlighting the importance of this RNA class as potential biomarkers (for more information on lncRNAs in urological cancers, see Chapter 2).

Similar to the discovery of PCa-associated lncRNAs, RNA-sequencing of small non-coding RNAs revealed that many small non-coding RNAs known as micro RNAs (miRNAs) are deregulated in PCa and can be combined in a biomarker signature (50). Furthermore, a subsequent study focusing on small nucleolar RNAs (snoRNAs) found that snoRNA-derived RNA fragments (sdRNAs) are upregulated in PCa and show association with malignancy and metastatic progression (120). These findings hint at altered RNA processing in cancerous conditions and underline the potential of NGS applications in PCa research.

DNA-sequencing was also crucial in recent studies that focused on studying tumor clonality in various cancers by identifying shared and distinct somatic mutations among the sequenced samples (121). As suggested by Peter Nowell in 1976 (122), a common observation was that several distinct subclones could be found in primary tumors and similarly, metastases could be linked to their progenitor clones due to mutual somatic events. This effectively allows the recreation of PCa progression and its spreading throughout the body, as has been performed in a recent study in PCa, which was able to trace the lineage of metastases in several PCa patients (123, 124). By identifying distinct mutational patterns between different metastatic sites, a temporal order of somatic alterations during the disease course could be created. For instance, it was confirmed that mutations in the AR were mostly present in late stage metastases and could be explained by increased selective pressure on AR signaling due to ongoing treatment strategies targeting the AR pathway.

As the presented examples illustrate, NGS technology and the associated bioinformatics data analyses not only provide many opportunities for clinical applications, but also enable research to gain further insights into the molecular foundations of prostate cancer.

1.6. Scope of this thesis

Prostate cancer is one of the most prevalent cancers in men and represents a severe burden on both patients and healthcare systems in western societies. To address the limitations of current risk stratification methods, the aim of this thesis was to identify novel biomarkers, such as RNA molecules that can discriminate PCa from non-cancerous conditions with high specificity and/or allow accurate prediction of disease outcome. Moreover, we aimed to advance next generation sequencing in clinical applications by addressing and simplifying some of the associated bioinformatics challenges.

To familiarize the reader with the concept of non-coding RNAs and the different classes found in mammalian cells, we provide an extensive review of current literature in chapter 2. The focus is on long non-coding RNAs and their potential functions and clinical applications in urological malignancies, as increasing evidence shows that expression or repression of many lncRNAs has functional consequences for the cell. Moreover, their tissue and even condition-specific transcription makes them ideal targets when searching for novel biomarkers.

Since many genomic regions are still poorly characterized and lncRNAs have often been described to reside between known genes or in so called "gene deserts", in chapter 3, we set out to investigate whether we could find any evidence of PCa-specific expression originating from such unnannotated regions. Taking into account the limitations of the PSA serum test, our aim was to discover transcripts that showed high PCa-specificity and little to no expression in normal control samples. We therefore adapted a cancer outlier profile analysis (COPA) approach first introduced by Tomlins *et al*. when describing the recurrent TMPRSS2-ETS fusions in PCa (33). Due to the fact that sequencing datasets of patient samples are still rather rare and raw data is often inaccessible, we made use of existing datasets based on a genome-wide microarray with probes targeting *in silico* predicted genes. This study design proved successful, as 334 candidate transcripts were identified and 15 uncharacterized RNAs were subsequently PCR validated, showing excellent power when combined into a gene panel for molecular diagnostic purposes.

In chapter 4, we evaluated the biomarker potential of PDE4D7, a phosphodiesterase isoform that was recently shown to be highly expressed in androgen-sensitive PCa cell lines, but not in androgen-insensitive cell lines. Since these *in vitro* experiments hinted at a possible predictive value, we made use of a large panel of existing datasets to confirm both its diagnostic as well as prognostic potential. Additionally, we investigated associations with existing clinical parameters and found that high PDE4D7 expression is associated with good clinical outcome and generally increased in samples harboring the TMPRSS2-ERG fusion gene or exhibiting ERG over-expression.

Encouraged by our PDE4D7 findings, we aimed to comprehensively study the behavior of all nine canonical human PDE4D isoforms in PCa development and progression. In chapter 5, we therefore again used a compilation of different expression datasets covering various disease stages from normal adjacent prostate to castration-resistant disease. After discovering a distinct isoform switch in localized primary disease, we set out to uncover possible

regulatory mechanisms that could explain such behavior. Utilizing copy number, DNA methylation, as well as ChIP-seq data, we found evidence that the isoform switch is actively controlled by hyper-methylation of specific regulatory sites as well as androgen signaling due to AR and ERG binding in the PDE4D locus. To show that PDE4D is also a potent biomarker, we created two signatures for diagnostic and prognostic purposes and demonstrated a direct clinical application in improving needle biopsies.

Once a malignant tumor has been identified, it is crucial to identify key driver mutations in the cancer cells that can be targeted with existing drugs, allowing a 'personalized treatment'. This approach, also commonly referred to as 'precision medicine', requires sequencing of the patient's genome using NGS. To lower the associated cost, it is possible to restrict the genomic input for sequencing to sites for which appropriate drugs exist by DNA capture technologies. However, current methods of data analysis are not adapted to this approach and rely on whole genome mapping, resulting in an increased data processing time. In chapter 6, we addressed this bioinformatics challenge and investigated whether it is possible to decrease the time needed for alignment by using only the target regions as search space. Since such an approach was not feasible with traditional DNA mappers, we implemented a novel solution based on the usage of *a priori* knowledge derived from a capture technology in development. Our results were encouraging and proved the feasibility of such endeavor by outperforming commonly used programs both in terms of speed and resource requirements.

To address the clinical need for prostate cancer biomarkers, we identified several novel RNA transcripts with good biomarker potential. Furthermore, we showed that the use of *a priori* knowledge in targeted sequencing enables a rapid data processing in an attempt to advance the utilization of next generation sequencing in precision medicine. In chapter 7, we discuss the relevance of our findings in context of basic research as well as clinical relevance.

References

1. Siegel,R., Naishadham,D. and Jemal,A. (2013) Cancer statistics, 2013. *CA Cancer J Clin*, **63**, 11–30.

2. Nelson,W.G., De Marzo,A.M. and Isaacs,W.B. (2003) Prostate cancer. *N. Engl. J. Med.*, **349**, 366–81.

3. Ferlay,J., Soerjomataram,I., Ervik,M., Dikshit,R., Eser,S., Mathers,C., Rebelo,M., Parkin,D.M., Forman,D. and Bray,F. (2013) GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase. No. 11 [Internet]. *Lyon, Fr. Int. Agency Res. Cancer.*, **11**, http://globocan.iarc.f.

4. Malvezzi,M., Bertuccio,P., Levi,F., La Vecchia,C. and Negri,E. (2014) European cancer mortality predictions for the year 2014. *Ann. Oncol.*, **25**, 1650–6.

5. Heidenreich,A., Bastian,P.J., Bellmunt,J., Bolla,M., Joniau,S., van der Kwast,T., Mason,M., Matveev,V., Wiegel,T., Zattoni,F., *et al.* (2014) EAU guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent-update 2013. *Eur. Urol.*, **65**, 124–37.

6. Heidenreich,A., Bastian,P.J., Bellmunt,J., Bolla,M., Joniau,S., van der Kwast,T., Mason,M., Matveev,V., Wiegel,T., Zattoni,F., *et al.* (2014) EAU Guidelines on Prostate Cancer. Part II: Treatment of Advanced, Relapsing, and Castration-Resistant Prostate Cancer. *Eur. Urol.*, **65**, 467–479.

7. Perlmutter,M.A. and Lepor,H. (2007) Androgen deprivation therapy in the treatment of advanced prostate cancer. *Rev. Urol.*, **9 Suppl 1**, S3–8.

8. Saad,F. and Fizazi,K. (2015) Androgen Deprivation Therapy and Secondary Hormone Therapy in the Management of Hormone-sensitive and Castration-resistant Prostate Cancer. *Urology*, **86**, 852–61.

9. Katzenwadel,A. and Wolf,P. (2015) Androgen deprivation of prostate cancer: Leading to a therapeutic dead end. *Cancer Lett.*, **367**, 12–7.

10. Ferraldeschi,R., Welti,J., Luo,J., Attard,G. and de Bono,J.S. (2015) Targeting the androgen receptor pathway in castration-resistant prostate cancer: progresses and prospects. *Oncogene*, **34**, 1745–57.

11. Karantanos,T., Corn,P.G. and Thompson,T.C. (2013) Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene*, **32**, 5501–11.

12. Bambury,R.M. and Rathkopf,D.E. (2015) Novel and next-generation androgen receptor-directed therapies for prostate cancer: Beyond abiraterone and enzalutamide. *Urol. Oncol.*, 10.1016/j.urolonc.2015.05.025.

13. Schalken,J. and Fitzpatrick,J.M. (2016) Enzalutamide: targeting the androgen signalling pathway in metastatic castration-resistant prostate cancer. *BJU Int.*, **117**, 215–25.

14. Graham,L. and Schweizer,M.T. (2016) Targeting persistent androgen receptor signaling in

castration-resistant prostate cancer. *Med. Oncol.*, **33**, 44.

15. Schröder,F.H., Hugosson,J., Roobol,M.J., Tammela,T.L.J., Ciatto,S., Nelen,V., Kwiatkowski,M., Lujan,M., Lilja,H., Zappa,M., *et al.* (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.

16. Pezaro,C., Woo,H.H. and Davis,I.D. (2014) Prostate cancer: measuring PSA. *Intern. Med. J.*, **44**, 433–40.

17. Schroder,F.H., Kruger,A.B., Rietbergen,J., Kranse,R., Maas,P. v. d., Beemsterboer,P. and Hoedemaeker,R. (1998) Evaluation of the Digital Rectal Examination as a Screening Test for Prostate Cancer. *JNCI J. Natl. Cancer Inst.*, **90**, 1817–1823.

18. Obort,A.S., Ajadi,M.B. and Akinloye,O. (2013) Prostate-specific antigen: any successor in sight? *Rev. Urol.*, **15**, 97–107.

19. Sindhwani,P. and Wilson,C.M. (2005) Prostatitis and serum prostate-specific antigen. *Curr. Urol. Rep.*, **6**, 307–12.

20. Roobol,M.J. and Carlsson,S. V (2013) Risk stratification in prostate cancer screening. *Nat. Rev. Urol.*, **10**, 38–48.

21. Andriole,G.L., Crawford,E.D., Grubb,R.L., Buys,S.S., Chia,D., Church,T.R., Fouad,M.N., Gelmann,E.P., Kvale,P.A., Reding,D.J., *et al.* (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.

22. Van der Kwast,T.H. and Roobol,M.J. (2013) Defining the threshold for significant versus insignificant prostate cancer. *Nat. Rev. Urol.*, **10**, 473–82.

23. Mohammed,A.A. (2014) Biomarkers in prostate cancer: new era and prospective. *Med. Oncol.*, **31**, 140.

24. Pal,R.P., Maitra,N.U., Mellon,J.K. and Khan,M.A. (2013) Defining prostate cancer risk before prostate biopsy. *Urol. Oncol.*, **31**, 1408–18.

25. O' Donoghue,P.M., McSweeney,S.E. and Jhaveri,K. (2010) Genitourinary imaging: current and emerging applications. *J. Postgrad. Med.*, **56**, 131–9.

26. Ghei,M., Pericleous,S., Kumar,A., Miller,R., Nathan,S. and Maraj,B.H. (2005) Finger-guided transrectal biopsy of the prostate: a modified, safer technique. *Ann. R. Coll. Surg. Engl.*, **87**, 386–7.

27. Humphrey,P.A. (2004) Gleason grading and prognostic factors in carcinoma of the prostate. *Mod. Pathol.*, **17**, 292–306.

28. Essink-Bot,M.L., de Koning,H.J., Nijs,H.G., Kirkels,W.J., van der Maas,P.J. and Schröder,F.H. (1998) Short-term effects of population-based screening for prostate cancer on health-related quality of life. *J. Natl. Cancer Inst.*, **90**, 925–31.

29. Yaghi,M.D. and Kehinde,E.O. (2015) Oral antibiotics in trans-rectal prostate biopsy and its efficacy to reduce infectious complications: Systematic review. *Urol. Ann.*, **7**, 417–27.

30. Bussemakers,M.J., van Bokhoven,A., Verhaegh,G.W., Smit,F.P., Karthaus,H.F., Schalken,J.A.,

Debruyne,F.M., Ru,N. and Isaacs,W.B. (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.*, **59**, 5975–9.

31. Dijkstra,S., Mulders,P.F.A. and Schalken,J.A. (2014) Clinical use of novel urine and blood based prostate cancer biomarkers: A review. *Clin. Biochem.*, **47**, 889–896.

32. Graif,T., Loeb,S., Roehl,K.A., Gashti,S.N., Griffin,C., Yu,X. and Catalona,W.J. (2007) Under diagnosis and over diagnosis of prostate cancer. *J. Urol.*, **178**, 88–92.

33. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.-W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R., *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–8.

34. Lepor,A., Catalona,W.J. and Loeb,S. (2016) The Prostate Health Index: Its Utility in Prostate Cancer Detection. *Urol. Clin. North Am.*, **43**, 1–6.

35. Punnen,S., Pavan,N. and Parekh,D.J. (2015) Finding the Wolf in Sheep's Clothing: The 4Kscore Is a Novel Blood Test That Can Accurately Identify the Risk of Aggressive Prostate Cancer. *Rev. Urol.*, **17**, 3–13.

36. Sartori,D.A. and Chan,D.W. (2014) Biomarkers in prostate cancer: what's new? *Curr. Opin. Oncol.*, **26**, 259–64.

37. Tomlins,S.A., Day,J.R., Lonigro,R.J., Hovelson,D.H., Siddiqui,J., Kunju,L.P., Dunn,R.L., Meyer,S., Hodge,P., Groskopf,J., *et al.* (2015) Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. *Eur. Urol.*, 10.1016/j.eururo.2015.04.039.

38. Falzarano,S.M., Ferro,M., Bollito,E., Klein,E.A., Carrieri,G. and Magi-Galluzzi,C. (2015) Novel biomarkers and genomic tests in prostate cancer: a critical analysis. *Minerva Urol. Nefrol.*, **67**, 211–31.

39. Boström,P.J., Bjartell,A.S., Catto,J.W.F., Eggener,S.E., Lilja,H., Loeb,S., Schalken,J., Schlomm,T. and Cooperberg,M.R. (2015) Genomic Predictors of Outcome in Prostate Cancer. *Eur. Urol.*, **68**, 1033–44.

40. Hamoen,E.H.J., de Rooij,M., Witjes,J.A., Barentsz,J.O. and Rovers,M.M. (2015) Use of the Prostate Imaging Reporting and Data System (PI-RADS) for Prostate Cancer Detection with Multiparametric Magnetic Resonance Imaging: A Diagnostic Meta-analysis. *Eur. Urol.*, **67**, 1112–21.

41. Schoots,I.G., Roobol,M.J., Nieboer,D., Bangma,C.H., Steyerberg,E.W. and Hunink,M.G.M. (2015) Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. *Eur. Urol.*, **68**, 438–50.

42. Schoots,I.G., Petrides,N., Giganti,F., Bokhorst,L.P., Rannikko,A., Klotz,L., Villers,A., Hugosson,J. and Moore,C.M. (2015) Magnetic resonance imaging in active surveillance of prostate cancer: a systematic review. *Eur. Urol.*, **67**, 627–36.

43. Mohammed,A.A., Shergill,I.S., Vandal,M.T. and Gujral,S.S. (2007) ProstaScint and its role in the diagnosis of prostate cancer. *Expert Rev. Mol. Diagn.*, **7**, 345–9.

44. Schroeder,R.P.J., van Weerden,W.M., Bangma,C., Krenning,E.P. and de Jong,M. (2009) Peptide

receptor imaging of prostate cancer with radiolabelled bombesin analogues. *Methods*, **48**, 200–4.

45. Høilund-Carlsen,P.F., Poulsen,M.H., Petersen,H., Hess,S. and Lund,L. (2014) FDG in Urologic Malignancies. *PET Clin.*, **9**, 457–68, vi.

46. Nanni,C., Zanoni,L. and Fanti,S. (2014) Nuclear medicine in urological cancers: what is new? *Future Oncol.*, **10**, 2061–72.

47. Chen,Q. and Zhong,T. (2015) The association of CXCR4 expression with clinicopathological significance and potential drug target in prostate cancer: a meta-analysis and literature review. *Drug Des. Devel. Ther.*, **9**, 5115–22.

48. Leyten,G.H.J.M., Hessels,D., Smit,F.P., Jannink,S.A., de Jong,H., Melchers,W.J.G., Cornel,E.B., de Reijke,T.M., Vergunst,H., Kil,P., *et al.* (2015) Identification of a Candidate Gene Panel for the Early Diagnosis of Prostate Cancer. *Clin. Cancer Res.*, **21**, 3061–70.

49. McKiernan,J., Donovan,M.J., O'Neill,V., Bentink,S., Noerholm,M., Belzer,S., Skog,J., Kattan,M.W., Partin,A., Andriole,G., *et al.* (2016) A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy. *JAMA Oncol.*, 10.1001/jamaoncol.2016.0097.

50. Martens-Uzunova,E.S., Jalava,S.E., Dits,N.F., van Leenders,G.J., Moller,S., Trapman,J., Bangma,C.H., Litman,T., Visakorpi,T. and Jenster,G. (2012) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*, **31**, 978–991.

51. Fabris,L., Ceder,Y., Chinnaiyan,A.M., Jenster,G.W., Sorensen,K.D., Tomlins,S., Visakorpi,T. and Calin,G.A. (2016) The Potential of MicroRNAs as Prostate Cancer Biomarkers. *Eur. Urol.*, 10.1016/j.eururo.2015.12.054.

52. Mlcochova,H., Hezova,R., Stanik,M. and Slaby,O. (2014) Urine microRNAs as potential noninvasive biomarkers in urologic cancers. *Urol. Oncol.*, **32**, 41.e1–9.

53. Yoshimoto,M., Ludkovski,O., DeGrace,D., Williams,J.L., Evans,A., Sircar,K., Bismar,T.A., Nuin,P. and Squire,J.A. (2012) PTEN genomic deletions that characterize aggressive prostate cancer originate close to segmental duplications. *Genes. Chromosomes Cancer*, **51**, 149–60.

54. Jenkins,R.B., Qian,J., Lieber,M.M. and Bostwick,D.G. (1997) Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization. *Cancer Res.*, **57**, 524–31.

55. Ribeiro,F.R., Henrique,R., Martins,A.T., Jerónimo,C. and Teixeira,M.R. (2007) Relative copy number gain of MYC in diagnostic needle biopsies is an independent prognostic factor for prostate cancer patients. *Eur. Urol.*, **52**, 116–25.

56. Sato,K., Qian,J., Slezak,J.M., Lieber,M.M., Bostwick,D.G., Bergstralh,E.J. and Jenkins,R.B. (1999) Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma. *J. Natl. Cancer Inst.*, **91**, 1574–80.

57. Lalonde,E., Ishkanian,A.S., Sykes,J., Fraser,M., Ross-Adams,H., Erho,N., Dunning,M.J., Halim,S., Lamb,A.D., Moon,N.C., *et al.* (2014) Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a

retrospective cohort study. *Lancet. Oncol.*, **15**, 1521–32.

58. Williams,J.L., Greer,P.A. and Squire,J.A. (2014) Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. *Cancer Genet.*, **207**, 474–88.

59. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

60. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–74.

61. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S., *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–11.

62. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–58.

63. Attard,G., Parker,C., Eeles,R.A., Schröder,F., Tomlins,S.A., Tannock,I., Drake,C.G. and de Bono,J.S. (2015) Prostate cancer. *Lancet*, **387**, 70–82.

64. Barbieri,C.E., Baca,S.C., Lawrence,M.S., Demichelis,F., Blattner,M., Theurillat,J.-P., White,T.A., Stojanov,P., Van Allen,E., Stransky,N., *et al.* (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.*, **44**, 685–9.

65. Cancer Genome Atlas Research Network. (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**, 1011–1025.

66. Gurel,B., Ali,T.Z., Montgomery,E.A., Begum,S., Hicks,J., Goggins,M., Eberhart,C.G., Clark,D.P., Bieberich,C.J., Epstein,J.I., *et al.* (2010) NKX3.1 as a marker of prostatic origin in metastatic tumors. *Am. J. Surg. Pathol.*, **34**, 1097–105.

67. Nordgard,S.H., Johansen,F.E., Alnaes,G.I.G., Bucher,E., Syvänen,A.-C., Naume,B., Børresen-Dale,A.-L. and Kristensen,V.N. (2008) Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mRNA expression, and germline haplotypes in breast cancer patients. *Genes. Chromosomes Cancer*, **47**, 680–96.

68. Matsuyama,H., Pan,Y., Yoshihiro,S., Kudren,D., Naito,K., Bergerheim,U.S.R. and Ekman,P. (2003) Clinical significance of chromosome 8p, 10q, and 16q deletions in prostate cancer. *Prostate*, **54**, 103–11.

69. Huang,S., Gulzar,Z.G., Salari,K., Lapointe,J., Brooks,J.D. and Pollack,J.R. (2012) Recurrent deletion of CHD1 in prostate cancer with relevance to cell invasiveness. *Oncogene*, **31**, 4164–70.

70. Khemlina,G., Ikeda,S. and Kurzrock,R. (2015) Molecular landscape of prostate cancer: Implications for current clinical trials. *Cancer Treat. Rev.*, **41**, 761–766.

71. Rodrigues,L.U., Rider,L., Nieto,C., Romero,L., Karimpour-Fard,A., Loda,M., Lucia,M.S., Wu,M., Shi,L., Cimic,A., *et al.* (2015) Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. *Cancer Res.*, **75**, 1021–34.

72. Xu,B., Chevarie-Davis,M., Chevalier,S., Scarlata,E., Zeizafoun,N., Dragomir,A., Tanguay,S., Kassouf,W., Aprikian,A. and Brimo,F. (2014) The prognostic role of ERG immunopositivity in

prostatic acinar adenocarcinoma: a study including 454 cases and review of the literature. *Hum. Pathol.*, **45**, 488–497.

73. Verras,M. and Sun,Z. (2006) Roles and regulation of Wnt signaling and beta-catenin in prostate cancer. *Cancer Lett.*, **237**, 22–32.

74. Kypta,R.M. and Waxman,J. (2012) Wnt/β-catenin signalling in prostate cancer. *Nat. Rev. Urol.*, **9**, 418–28.

75. Wang,Y., Kreisberg,J.I. and Ghosh,P.M. (2007) Cross-talk between the androgen receptor and the phosphatidylinositol 3-kinase/Akt pathway in prostate cancer. *Curr. Cancer Drug Targets*, **7**, 591–604.

76. Culig,Z. (2004) Androgen receptor cross-talk with cell signalling pathways. *Growth Factors*, **22**, 179–84.

77. Carey,A.-M., Pramanik,R., Nicholson,L.J., Dew,T.K., Martin,F.L., Muir,G.H. and Morris,J.D.H. (2007) Ras-MEK-ERK signaling cascade regulates androgen receptor element-inducible gene transcription and DNA synthesis in prostate cancer cells. *Int. J. Cancer*, **121**, 520–7.

78. Peterziel,H., Mink,S., Schonert,A., Becker,M., Klocker,H. and Cato,A.C. (1999) Rapid signalling by androgen receptor in prostate cancer cells. *Oncogene*, **18**, 6322–9.

79. Marques,R.B., Aghai,A., de Ridder,C.M.A., Stuurman,D., Hoeben,S., Boer,A., Ellston,R.P., Barry,S.T., Davies,B.R., Trapman,J., *et al.* (2015) High Efficacy of Combination Therapy Using PI3K/AKT Inhibitors with Androgen Deprivation in Prostate Cancer Preclinical Models. *Eur. Urol.*, **67**, 1177–85.

80. Park,H., Kim,Y., Sul,J.-W., Jeong,I.G., Yi,H.-J., Ahn,J.B., Kang,J.S., Yun,J., Hwang,J.J. and Kim,C.-S. (2015) Synergistic anticancer efficacy of MEK inhibition and dual PI3K/mTOR inhibition in castration-resistant prostate cancer. *Prostate*, **75**, 1747–59.

81. Merkle,D. and Hoffmann,R. (2011) Roles of cAMP and cAMP-dependent protein kinase in the progression of prostate cancer: Cross-talk with the androgen receptor. *Cell. Signal.*, **23**, 507–515.

82. Sarwar,M., Sandberg,S., Abrahamsson,P.-A. and Persson,J.L. (2014) Protein kinase A (PKA) pathway is functionally linked to androgen receptor (AR) in the progression of prostate cancer. *Urol. Oncol.*, **32**, 25.e1–12.

83. Desiniotis,A., Schäfer,G., Klocker,H. and Eder,I.E. (2010) Enhanced antiproliferative and proapoptotic effects on prostate cancer cells by simultaneously inhibiting androgen receptor and cAMP-dependent protein kinase A. *Int. J. Cancer*, **126**, 775–89.

84. Henderson,D.J.P., Byrne,A., Dulla,K., Jenster,G., Hoffmann,R., Baillie,G.S. and Houslay,M.D. (2014) The cAMP phosphodiesterase-4D7 (PDE4D7) is downregulated in androgen-independent prostate cancer cells and mediates proliferation by compartmentalising cAMP at the plasma membrane of VCaP prostate cancer cells. *Br. J. Cancer*, **110**, 1278–87.

85. Yoon,H.-K., Hu,H.-J., Rhee,C.-K., Shin,S.-H., Oh,Y.-M., Lee,S.-D., Jung,S.-H., Yim,S.-H., Kim,T.-M. and Chung,Y.-J. (2014) Polymorphisms in PDE4D are associated with a risk of COPD in non-emphysematous Koreans. *COPD*, **11**, 652–8.

86. Michot,C., Le Goff,C., Goldenberg,A., Abhyankar,A., Klein,C., Kinning,E., Guerrot,A.M.,

Flahaut,P., Duncombe,A., Baujat,G., *et al.* (2012) Exome sequencing identifies PDE4D mutations as another cause of acrodysostosis. *Am. J. Hum. Genet.*, **90**, 740–745.

87. Kaname,T., Ki,C.-S., Niikawa,N., Baillie,G.S., Day,J.P., Yamamura,K.-I., Ohta,T., Nishimura,G., Mastuura,N., Kim,O.-H., *et al.* (2014) Heterozygous mutations in cyclic AMP phosphodiesterase-4D (PDE4D) and protein kinase A (PKA) provide new insights into the molecular pathology of acrodysostosis. *Cell. Signal.*, **26**, 2446–59.

88. Gretarsdottir,S., Thorleifsson,G., Reynisdottir,S.T., Manolescu,A., Jonsdottir,S., Jonsdottir,T., Gudmundsdottir,T., Bjarnadottir,S.M., Einarsson,O.B., Gudjonsdottir,H.M., *et al.* (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.*, **35**, 131–138.

89. Lee,H., Graham,J.M., Rimoin,D.L., Lachman,R.S., Krejci,P., Tompson,S.W., Nelson,S.F., Krakow,D. and Cohn,D.H. (2012) Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am. J. Hum. Genet.*, **90**, 746–751.

90. Wettschureck,N. and Offermanns,S. (2005) Mammalian G proteins and their cell type specific functions. *Physiol. Rev.*, **85**, 1159–204.

91. Walsh,D.A. and Van Patten,S.M. (1994) Multiple pathway signal transduction by the cAMP-dependent protein kinase. *FASEB J.*, **8**, 1227–36.

92. Dema,A., Perets,E., Schulz,M.S., Deák,V.A. and Klussmann,E. (2015) Pharmacological targeting of AKAP-directed compartmentalized cAMP signalling. *Cell. Signal.*, **27**, 2474–2487.

93. Mellon,P.L., Clegg,C.H., Correll,L.A. and McKnight,G.S. (1989) Regulation of transcription by cyclic AMP-dependent protein kinase. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 4887–91.

94. Kasbohm,E.A., Guo,R., Yowell,C.W., Bagchi,G., Kelly,P., Arora,P., Casey,P.J. and Daaka,Y. (2005) Androgen receptor activation by G(s) signaling in prostate cancer cells. *J. Biol. Chem.*, **280**, 11583–9.

95. Houslay,M.D. (2010) Underpinning compartmentalised cAMP signalling through targeted cAMP breakdown. *Trends Biochem. Sci.*, **35**, 91–100.

96. Zhang,R., Wang,Y., Zhang,L., Zhang,Z., Tsang,W., Lu,M., Zhang,L. and Chopp,M. (2002) Sildenafil (Viagra) induces neurogenesis and promotes functional recovery after stroke in rats. *Stroke.*, **33**, 2675–80.

97. Chen,L., Staubli,S.E.L., Schneider,M.P., Kessels,A.G., Ivic,S., Bachmann,L.M. and Kessler,T.M. (2015) Phosphodiesterase 5 inhibitors for the treatment of erectile dysfunction: a trade-off network meta-analysis. *Eur. Urol.*, **68**, 674–80.

98. Titus,D.J., Oliva,A.A., Wilson,N.M. and Atkins,C.M. (2015) Phosphodiesterase inhibitors as therapeutics for traumatic brain injury. *Curr. Pharm. Des.*, **21**, 332–42.

99. Heckman,P.R.A., Wouters,C. and Prickaerts,J. (2015) Phosphodiesterase inhibitors as a target for cognition enhancement in aging and Alzheimer's disease: a translational overview. *Curr. Pharm. Des.*, **21**, 317–31.

100. Mei,R., Galipeau,P.C., Prass,C., Berno,A., Ghandour,G., Patil,N., Wolff,R.K., Chee,M.S., Reid,B.J. and Lockhart,D.J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, **10**, 1126–37.

101. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–70.

102. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

103. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

104. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

105. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–359.

106. Marco-Sola,S., Sammeth,M., Guigo,R. and Ribeca,P. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, **9**, 1185–1188.

107. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

108. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 10.1093/bioinformatics/bts635.

109. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

110. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M., *et al.* (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

111. Robinson,M., Mccarthy,D., Chen,Y. and Smyth,G.K. (2011) edgeR : differential expression analysis of digital gene expression data User ' s Guide. *Most*, **23**, 1–77.

112. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

113. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–78.

114. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

115. Patro,R., Mount,S.M. and Kingsford,C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–4.

116. Patro,R., Duggal,G. and Kingsford,C. (2015) Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment Cold Spring Harbor Labs Journals.

117. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq

quantification. *Nat. Biotechnol.*, 10.1038/nbt.3519.

118. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511–515.

119. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S., Kominsky,H.D., *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*, **29**, 742–749.

120. Martens-Uzunova,E.S., Hoogstrate,Y., Kalsbeek,A., Pigmans,B., Vredenbregt-van den Berg,M., Dits,N., Nielsen,S.J., Baker,A., Visakorpi,T., Bangma,C., *et al.* (2015) C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget*, **6**, 17430–44.

121. Caldas,C. (2012) Cancer sequencing unravels clonal evolution. *Nat. Biotechnol.*, **30**, 408–10.

122. Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–8.

123. Boutros,P.C., Fraser,M., Harding,N.J., de Borja,R., Trudel,D., Lalonde,E., Meng,A., Hennings-Yeomans,P.H., McPherson,A., Sabelnykova,V.Y., *et al.* (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.*, **47**, 736–45.

124. Gundem,G., Van Loo,P., Kremeyer,B., Alexandrov,L.B., Tubio,J.M.C., Papaemmanuil,E., Brewer,D.S., Kallio,H.M.L., Högnäs,G., Annala,M., *et al.* (2015) The evolutionary history of lethal metastatic prostate cancer. *Nature*, **520**, 353–357.

Chapter 2

Long Noncoding RNA in Prostate, Bladder, and Kidney Cancer

Elena S. Martens-Uzunova[1], René Böttcher[1,2], Carlo M. Croce[3], Guido Jenster[1], Tapio Visakorpi[4], George A. Calin[5,6]

1. Department of Urology, Erasmus Medical Centre, Rotterdam, The Netherlands
2. Department of Bioinformatics, Technical University of Applied Sciences Wildau, Wildau, Germany
3. Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, Ohio State University, Columbus, OH, USA
4. Institute of Medical Technology, University of Tampere, Tampere University Hospital, Tampere, Finland
5. Department of Experimental Therapeutics and Leukemia, University of Texas MD Anderson Cancer Center, Houston, TX, USA
6. Center for RNA Interference and Non-Coding RNAs, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Published in

European Urology. 201;65(6):1140-51

Supplementary Material is available via

http://www.europeanurology.com/article/S0302-2838(13)01324-9/fulltext/long-noncoding-rna-in-prostate-bladder-and-kidney-cancer

Abstract

Context: Genomic regions without protein-coding potential give rise to millions of protein-noncoding RNA transcripts (noncoding RNA) that participate in virtually all cellular processes. Research over the last 10 yr has accumulated evidence that long noncoding RNAs (lncRNAs) are often altered in human urologic cancers.

Objective: To review current progress in the biology and implication of lncRNAs associated with prostate, bladder, and kidney cancer.

Evidence acquisition: The PubMed database was searched for articles in the English language with combinations of the Medical Subject Headings terms long non coding RNA, long noncoding RNA, long untranslated RNA, cancer, neoplasms, prostate, bladder, and kidney.

Evidence synthesis: We summarise existing knowledge on the systematics, biology, and function of lncRNAs, particularly these involved in prostate, kidney, and bladder cancer. We also discuss the possible utilisation of lncRNAs as novel biomarkers and potential therapeutic targets in urologic malignancies and portray the major challenges and future perspectives of ongoing lncRNA research.

Conclusions: LncRNAs are important regulators of gene expression interacting with the major pathways of cell growth, proliferation, differentiation, and survival. Alterations in the function of lncRNAs promote tumour formation, progression, and metastasis of prostate, bladder, and kidney cancer. LncRNAs can be used as noninvasive tumour markers in urologic malignancies. Increased knowledge of the molecular mechanisms by which lncRNAs perform their function in the normal and malignant cell will lead to a better understanding of tumour biology and could provide novel therapeutic targets for the treatment of urologic cancers.

Patient summary: In this paper we reviewed current knowledge of long noncoding RNAs (lncRNAs) for the detection and treatment of urologic cancers. We conclude that lncRNAs can be used as novel biomarkers in prostate, kidney, or bladder cancer. LncRNAs hold promise as future therapeutic targets, but more research is needed to gain a better understanding of their biologic function.

Introduction

Definition of noncoding RNA

The sequencing of the human genome led to the astonishing discovery that protein-coding genes compose <3% of human DNA. Yet >80% of our genome is actively transcribed to a versatile group of RNA transcripts without protein-coding potential (1, 2). Such transcripts are referred to as noncoding RNAs (ncRNAs). Based on their size and the arbitrary cut-off of 200 nucleotides (nts), they are classified into long ncRNAs (lncRNAs) and small ncRNAs. Although small ncRNAs, in particular microRNAs (miRNAs), have been extensively studied over the last two decades and many facets of their biology have been elucidated, still very little is known about the functional role of lncRNAs.

First evidence of the existence of long noncoding RNAs and their involvement in urologic cancers

The first suggestion that not all long RNA transcripts are messenger RNAs (mRNAs) that merely pass information from DNA to protein came >20 yr ago with the discovery of the paternally imprinted maternally expressed transcript (H19) gene, encoding a foetal-specific lncRNA deregulated in embryonic and adult tumours, particularly in bladder carcinoma (3, 4). Shortly after, the identification of the X inactive specific transcript (XIST) proposed a regulatory and structural function for ncRNAs (5). Evidence for the importance of protein-noncoding gene regions was also provided by the discovery of several transcripts, such as tumour suppressor growth arrest-specific 5 (GAS5), which is also a host gene for small nucleolar RNA (snoRNA) (6). The idea that lncRNAs may exhibit cancer-specific expression was strengthened by the discovery of prostate cancer antigen 3 lncRNA (PCA3 [DD3]) specifically overexpressed in malignant prostate tissue (7).

How many long noncoding RNAs are there?

The thorough annotation of the human genome by the ENCODE (1) and GENCODE (8) projects demonstrated that human DNA is pervasively transcribed, including regions that overlap protein-coding loci and regions previously assumed to be transcriptionally silent, and that many of these transcription products are in fact lncRNAs. Within 4 yr, the number of identified lncRNA genes increased from 6000 to >14 000. It is likely that hundreds of thousands if not millions of lncRNAs are yet to be discovered because 15% of the human genome remains to be annotated, and lncRNAs arising from overlapping protein-encoding loci still remain to be analysed (8).

Although the function of most lncRNAs is still unknown, their increasing numbers and the accumulating evidence for their involvement in many biologic processes provide compelling arguments in support of their importance in the normal and malignant cell.

Here we summarise existing knowledge on the systematics, biology, and function of lncRNAs, particularly these involved in prostate, kidney, and bladder cancer. We also discuss the possible utilisation of lncRNAs as novel biomarkers and potential therapeutic targets in

urologic malignancies and describe the major challenges and future perspectives of ongoing lncRNA research.

Evidence acquisition

The PubMed database was searched for articles in the English language published up to July 2013 with a combination of the following Medical Subject Headings terms: long non coding RNA, or long noncoding RNA, or long untranslated RNA, and cancer or neoplasms, and prostate or bladder or kidney.

Evidence synthesis

Features of long noncoding RNA

The GENCODE annotation project established the largest catalogue of human lncRNAs to date and provided comprehensive information regarding several common features of their structural organisation and genome processing (8). The following information is now known:

- LncRNAs are independent transcriptional units without protein-coding potential and not just unrecognized extensions of neighbouring protein-coding transcripts.
- Expressed lncRNA genes have the typical histone modifications associated with active transcription, but show generally lower and more tissue-specific expression compared with protein-coding genes.
- LncRNA- and protein-coding genes share similar length, processing, and splicing signals.
- LncRNA genes belong to evolutionary conserved families evolving faster than protein-coding genes where sequence similarity seems to be preserved mainly in regions involved in secondary structure formation

Overview of noncoding RNAs classes

The increasing numbers of newly discovered ncRNAs required the establishment of uniform nomenclature for long and small ncRNAs, which was introduced in 2011 by the HUGO Gene Nomenclature Committee (9). At present, lncRNAs are categorised based mainly on their location in respect to protein-coding genes because their functional classification is largely hampered by the lack of known function. In contrast, small ncRNAs are relatively well studied, and different subclasses are recognised based on their structural features and biologic function (10). For example, the best known small RNAs, miRNAs, are 20–22 nt in length and act as negative posttranscriptional regulators of gene expression (Table 1, Fig. 1).

Biologic functions of long noncoding RNAs in urologic cancers and their interaction with major cancer pathways

Accumulating evidence demonstrates that lncRNAs function as versatile regulators at each step during genetic information processing in the living cell. As such, lncRNAs interact with major cellular pathways controlling proliferation, differentiation, or apoptosis, and alterations in their function are involved in the pathogenesis of many human malignancies including prostate, kidney, and bladder cancer. Recent advances in transcriptome sequencing led to the discovery of many new lncRNAs (11, 12) associated with urologic malignancies and allowed the reexamination of other long-known cancer-associated lncRNAs with a function that has remained enigmatic for decades (Table 2).

Table 1 - Human noncoding RNA nomenclature

Type of ncRNA	Abbreviation	Symbol	Function/Description
Ribosomal RNA			**Protein synthesis**
Genomic	rRNA	RN18S, RN28S, RN5–8S, RN5S	
Mitochondrial	mit-rRNA		
Long ncRNA			
Antisense transcripts Opposite strand HOXA/B/C/D clusters	Antisense RNA	-AS -OS HOXA/B/C/D	Reside on the opposite strand of protein-coding genes and intersect their exons
Overlapping transcripts		-OT	Contain a coding gene within an intron on the same strand
Intronic transcripts		-IT	Reside within introns of a coding gene but do not intersect any exons
Host genes		-HG	Primary hosts of small ncRNA genes nested within their introns
Intergenic lncRNAs	lincRNA	LINC	Originate from protein noncoding genomic regions
lncRNA paralogues		TTTY, HCG, FAM, DGCR	Share homology with each other
Ultraconserved transcripts	ucRNA	Not assigned yet	Originate from genomic regions with 100% conservation between human, rat, and mouse
Circular RNAs	circRNA	Not assigned yet	Form during splicing by chemical bonding of two neighbouring exons. Function as miRNA sponges
Enhancer RNAs	eRNA	Not assigned yet	Originate at genomic enhancer regions. Boost gene transcription in tissue-specific and temporal manner.
Sno-related lncRNAs	Sno-lncRNA	Not assigned yet	Generated when the sequences between intronically encoded snoRNAs are not degraded. Sno-lncRNAs are flanked by snoRNAs instead of 5'-cap and 3'-poly(A) tail
Pseudogenes		-P	Highly similar to rotein-coding genes that have lost their coding potential. Generally untranscribed and/or untranslated but can be activated in different tissues or in cancer.
Small ncRNAs	**sncRNA**		
MicroRNA	miRNA	MIR	Posttranscriptional regulators
Transfer RNA Genomic, mitochondrial	tRNA	TRNA, MT-T	Protein synthesis

Spliceosomal RNA	splRNA	RNU	RNA splicing/maturation
Small nucleolar RNA H/ACA box, C/D box Cadjal body specific	snoRNA	SNORA, SNORD SCARNA	Ribosomal maturation, alternative splicing
PiWi-interacting RNA	piRNA	PIRC	Posttranscriptional retrotransposon silencing
RNase P/MRP RNA components		RPPH1; RMRP	tRNA and mitochondrial RNA processing
U7 snRNA		RNU7	Histone pre-mRNA processing
Vault RNA		VTRNA	Components of the vault RNP; possible role in drug resistance
7SK RNA		RN7SK	Regulates Pol II transcription
7SL RNA		RN7SL	Transmembrane transport of proteins
YRNAs		RNY	Assist the structural specificity of Ro PNPase
Telomerase RNA		TERC	Prevents erosion of chromosome ends
mRNA = messenger RNA; ncRNA = noncoding RNA.			

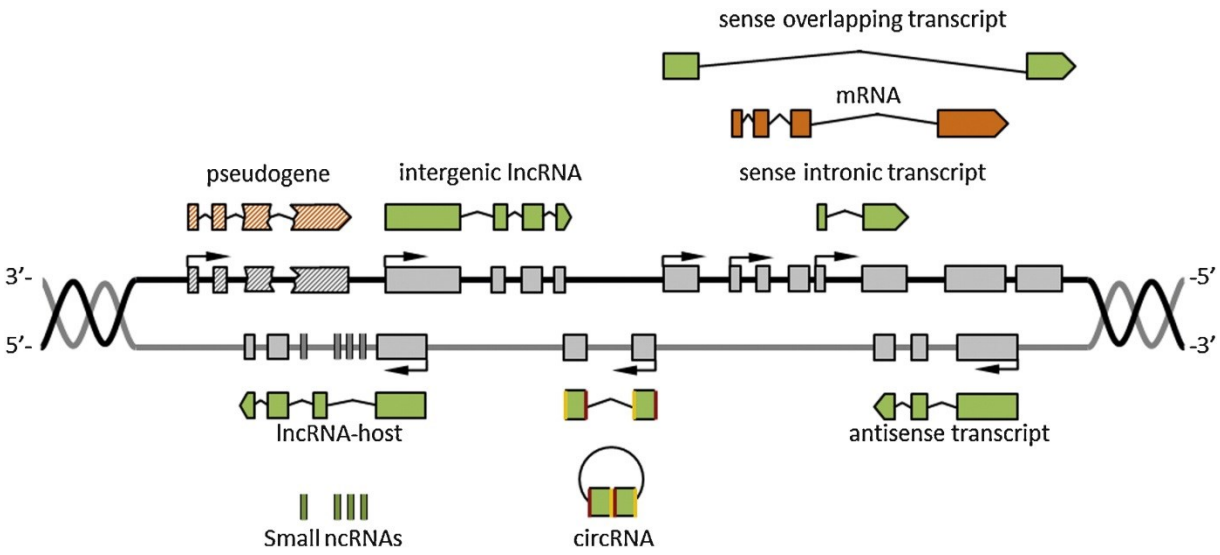


Figure 1: Genomic organisation of different long noncoding RNA (lncRNA) classes. A grey and black line represents DNA strands. Grey boxes represent protein- or lncRNA-coding genomic exons. Thin black lines represent spliced introns. Arrows indicate direction of transcription. Protein-coding transcripts (messenger RNAs) are orange. Noncoding transcripts (lncRNAs) are green. Pseudogenes have a diagonal stripe pattern. Intron boundaries of circular RNA precursors are shown in red (-5') and yellow (-3').

Long noncoding RNAs are epigenetic regulators of gene expression

Many lncRNAs reside in the nucleus where they actively interact with chromatin remodelling complexes (CRCs) to regulate the expression of genes residing on the same chromosome (in cis) or on another chromosome (in trans) through fine-tuning of chromatin architecture (2, 8).

Table 2 – Long noncoding RNAs in prostate, bladder, and kidney cancer

Name	Cytoband	Cancer type	Association with cancer
CBR3-AS1	21q22.2	Prostate	Oncogene; putative therapeutic target (70)
CTBP1-AS	4p16.3	Prostate	Oncogene (71)
GAS5	1q25.1	Kidney	Tumour suppressor (100)
		Prostate	Tumour suppressor; (34) putative oncogene host (38)
H19	11p15.5	Bladder	Prognostic marker (83); low-risk marker (85) Oncogene (4,45–47,50,51); targeted therapy agent (91)
		Kidney	Tumour suppressor (42); tumour suppressor host (37)
		Prostate	Putative susceptibility and diagnostic marker (43,44)
HIF1A-AS1, AS2	14q23.2	Kidney	Diagnostic and discriminative marker (81,82)
KCNQ1OT1	11p15	Kidney	Oncogene (40,41)
MALAT1	11q13.1	Bladder	Oncogene (59,60)
		Kidney	Oncogene (63,63)
		Prostate	Putative marker (61)
MEG3	14q32	Bladder	Tumour suppressor (54)
		Kidney	Tumour suppressor (55)
PCA3	9q21-q22	Prostate	Diagnostic marker (7,74)
PCAT1	8q24.21	Prostate	Putative marker and oncogene (11)
PCGEM	2q32	Prostate	High-risk (17,18,86) and predictive marker (78,79) Oncogene (19,21)
PRNCR1	8q24	Prostate	Susceptibility marker (20); oncogene (20,21)
PTENP1	9p21	Prostate	Oncogene; tumour suppressor (28,29)
SNHG16		Bladder	Putative diagnostic, prognostic, and predictive marker (80)
TUG1	22q12.2	Bladder	Putative diagnostic and prognostic marker; oncogene (65)
UCA1	19p13.12	Bladder	Diagnostic marker (66); oncogene (67–69)
ucRNAs	Multiple	Prostate	Putative oncogenes (73)
XIST	Xq13.2	Prostate	Putative diagnostic and prognostic marker (76,77)

ucRNA = ultraconserved RNA. An extended version of Table 2 is provided as Supplemental Table 1.

Cis-acting repressor long noncoding RNAs.

The most prominent epigenetic cis-regulatory lncRNA is the XIST that controls the X-linked gene dosage compensation between XY males and XX females. XIST is exclusively expressed from the future inactive X chromosome in females, accumulates in large quantities, and coats its host X chromosome (5). This causes the exclusion of RNA polymerase II and a rapid gain of repressive histone marks. In this way, XIST silences in cis the expression of the entire X chromosome. In male cancers with an X chromosome gain (eg, testicular germ cell tumours [TGCTs]), hypomethylation and reexpression of XIST can occur (14).

Trans-acting repressor long noncoding RNAs.

The first example of a trans-acting lncRNA came from the powerful breast cancer oncogene HOX transcript antisense RNA (HOTAIR). In normal cells, HOTAIR binds the chromatin remodelling polycomb repressor complex 2 (PRC2) and targets it to the HOXD locus (located on a different chromosome) where PRC2 performs the silencing of embryonic transcription factors. Overexpression of HOTAIR in cancer causes the genome-wide relocalisation of PRC2 and the epigenetic silencing of metastasis suppressor genes that drives cancer progression (15).

Similar to HOTAIR in breast cancer, in prostate cancer (PCa) the lincRNA prostate cancer associated transcript 1 (PCAT1) complexes with PRC2. PCAT1 is markedly overexpressed in a subset of metastatic cancers, suggesting that PCAT1 is a transcriptional regulator and may function as a prostate-specific transcriptional repressor of tumour suppressor genes controlling cell proliferation that may have an important role in PCa progression (11). By now, thousands of lincRNAs have been shown to complex with PRC2 and other repressive CRCs to supply them with the specificity needed to target distinct gene sets (16, 17).

Activating long noncoding RNAs.

LncRNAs can also function as transcriptional activators as demonstrated for prostate cancer non-coding RNA 1 (PRNCR1) and prostate-specific transcript 1 (PCGEM1). PCGEM1 is a highly prostate-specific, androgen-regulated lncRNA (18), with expression significantly higher in PCa specimens from African American and Chinese men (12, 19). Its overexpression in PCa cells promotes cell proliferation, attenuates doxorubicin-induced expression of p53 and p21, and inhibits apoptosis (19, 20). PRNCR1 is upregulated in precursor lesion prostatic intraepithelial neoplasia, and it is positively associated with the viability of PCa cells (21). Yang et al. recently demonstrated that PRNCR1 and PCGEM1 successively interact with the androgen receptor (AR) bound at DNA-enhancer regions in a ligand-dependent fashion and facilitate the chromosomal looping between AR-bound enhancers and the promoter sequences of androgen-responsive genes. Interestingly, in castration-resistant PCa cells, the overexpressed PCGEM1 and PRNCR1 can further cause the ligand-independent activation of (truncated) AR and promote cellular proliferation. This proposes both lncRNAs as possible therapeutic drug targets in advanced PCa (22).

Enhancer-originating RNAs (eRNAs) may also act as epigenetic cis-activators that maintain an active chromatin state at transcribed gene loci. In PCa cells, bidirectional eRNA production is induced after binding of AR to responsive enhancers, and it is concomitant with sustaining of open chromatin structure and indirect activation of gene expression (23) possibly via the

mechanism described above. Additionally, the transcription of unidirectional intergenic eRNAs can also lead to direct activation of adjacent genes (24). At present, a cis-activating role and an enhancer-like function is also anticipated for a large set of lncRNAs located adjacent to cancer-related protein-coding genes that act a step further in gene transcription by promoting transcriptional elongation (25).

Long noncoding RNAs and transcribed pseudogenes as competing endogenous RNA

The single-stranded nature and the native (self-)complementarity constitute the functionality of RNA. These features enable the rapid and specific contact between different RNAs (eg, between mRNA and miRNA during repression of protein synthesis). In this process, miRNAs in complex with effector proteins recognise and bind complementary sequence stretches (binding sites) in multiple mRNA targets. This leads to mRNA degradation or translational repression of the encoded protein. It iswell established that miRNAs have an essential regulatory role in virtually all cellular processes and that altered miRNA levels are implicated in many human cancers including urologic malignancies (26). Nevertheless, the mechanisms controlling the cellular levels of active miRNAs remained enigmatic until recently when several publications demonstrated that mRNAs, pseudogenes, and circular RNAs (circRNAs) compete endogenously for shared effector miRNAs. Such competing endogenous RNAs (ceRNAs; also called miRNA decoys or miRNA sponges) provide yet an additional regulatory level in the RNA network controlling gene expression (27).

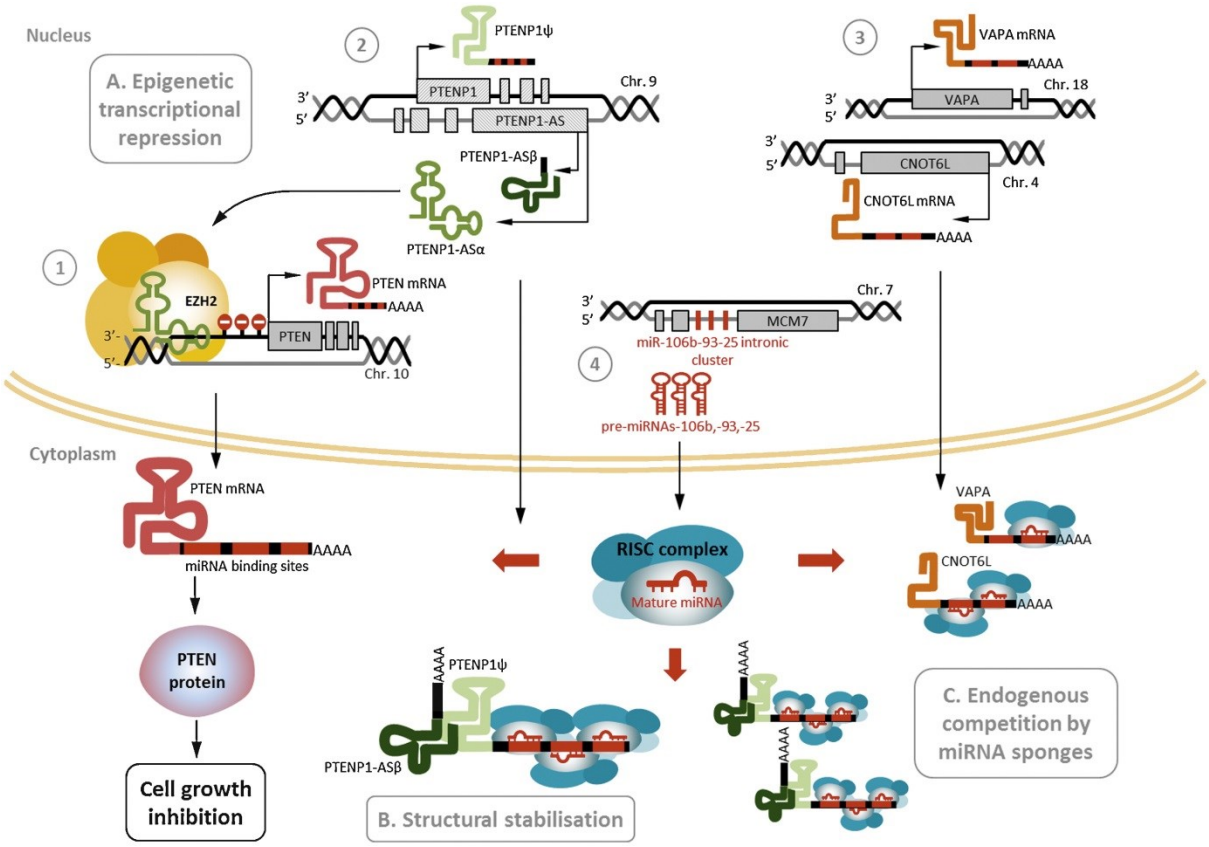


Figure 2: The expression of tumour suppressor phosphatase and tensin homolog (PTEN) is regulated by a complex noncoding RNA (ncRNA) network representing many long ncRNA (lncRNA) functions. 1) The PTEN gene encoded at chromosome 10 is transcribed to

PTEN messenger RNA (mRNA; in brown), which is exported to the cytoplasm and translated to PTEN protein that acts as a negative regulator of cell growth. The 30-untranslated region of PTEN mRNA contains binding sites (red/black dashed line) for microRNAs (miRNAs) from the miR-106b-93-25 cluster. **2)** PTENP1 pseudogene, highly homologous to PTEN, is encoded at chromosome 9 and coexpressed with PTEN in normal and malignant prostate tissues. Three lncRNAs (in green) are simultaneously transcribed from PTENP1: one in sense: PTENP1-C, and two in antisense: PTENP1-ASa and PTENP1-ASb. The PTENP1-C sequence is similar to PTEN mRNA and also contains binding sites for the miR-106b-93-25 cluster. **3)** Two protein-coding genes, CNOT6L and VAPA (in orange), encoded at chromosomes 4 and 18, respectively, contain binding sites for the miR-106b-93-25 cluster. **4)** The miRNA cluster miR-106b-93-25 (in red) targeting PTEN mRNA is intronically encoded at the MCM7 gene at chromosome 7. When overexpressed, miRNAs from the miR-106b-93-25 cluster are exported to the cytoplasm where they associate with the RNA-induced silencing complex (RISC) (in blue), cause the downregulation of PTEN mRNA, and promote prostate tumorigenesis. **A)** In the nucleus, PTENP1-ASa acts as a trans-acting epigenetic repressor that localises to the PTEN promoter and recruits the chromatin repressive protein complex EZH2 (in yellow). EZH2 silences the transcription of PTEN by introducing repressive histone marks (lollypops) at the PTEN promoter. **B)** The structure of PTENP1-c is stabilised by PTENP1-ASb by the formation of a double-stranded RNA:RNA tandem, which is exported to the cytoplasm. **C)** In the cytoplasm, this tandem functions as a miRNA sponge and sequesters miR-106b, miR-93, and miR-25. This leads to the de-repression of PTEN mRNA, higher levels of PTEN protein, and subsequent growth inhibition. In addition, independently of their coding potential, the VAPA and CNOT6L mRNAs can also function as miRNA sponges for PTEN.

Functional pseudogenes and coding-independent function of messenger RNAs.

A functional role for an expressed pseudogene acting as a ceRNA was first described for the regulatory RNA network that tidily controls the cellular levels of the phosphatase and tensin homolog (PTEN) (Fig. 2). PTEN is a haploinsufficient tumour suppressor, commonly lost in advanced PCa and other cancers, that antagonises cell growth signalling mediated by the PI3K-AKT pathway. PTEN mRNA is negatively regulated by the intronically encoded miRNA cluster miR-106b-93-25. Overexpression of miR-106b-93-25 causes downregulation of PTEN and initiates prostate tumourigenesis (28). Excess levels of miR-106b-93-25 can be sequestered by the miRNA sponge structure formed by two lncRNAs (PTENP1-ASb and PTENP1-C) produced from the pseudogene PTENP1 highly homologous to PTEN (29, 30), as well as by at least two other protein-coding RNAs (31).

The regulatory system that controls PTEN expression is probably not an exception because miRNA target sites in genes and their pseudogenes are well conserved and have been detected in the pseudogenes of gap junction protein, alpha 1, 43kDa (CX43), cyclin-dependent kinase 4 pseudogene (CDK4PS), forkhead box O3B pseudogene (FOXO3B), E2F transcription factor 3 pseudogene 1 (E2F3P1), POU class 5 homeobox 1 (OCT4), and Kirsten rat sarcoma viral oncogene homolog (KRAS). Notably, the four OCT4 pseudogenes are exclusively

expressed in cancer tissues and KRAS and Kirsten rat sarcoma viral oncogene homolog pseudogene 1 (KRAS1P) are positively correlated in PCa, suggesting a putative proto-oncogenic role for KRAS1P (29).

MicroRNA sponges.

Besides pseudogenes and mRNA, a newly emerging class of ceRNAs are the circRNAs. CircRNAs were first discovered in testis tissue (32), but their functionality was questioned until recently when transcriptome sequencing demonstrated the existence of thousands of well-expressed, tissue-specific stable circRNAs. Among these, a prominent example is the human circRNA antisense to the cerebellar degeneration-related protein 1 transcript CDR1 antisense RNA (CDR1-AS, ciRS-7). CDR1-AS harbours >70 conserved binding sites for the tumour suppressor miR-7 and binds miR-7 associated with its effector proteins with a capacity 10 times higher than any other known transcript (33, 34).

Sequence versus structure: protein-interacting competing endogenous RNAs.

The primary RNA sequence is the base of miRNA sponge activity. However, specific secondary RNA structures can be recognised by DNA/RNA interacting proteins (eg, nuclear transcription factors), even when the primary sequence is not quite conserved. This allows some lncRNAs to function as protein-interacting ceRNA exemplified by function of the GAS5 lncRNA. Although the primary sequence of the spliced GAS5 transcript is not preserved, its secondary structure mimics the genomic glucocorticoid receptor (GR) response element. In arrested cells, GAS5 binds GR and prevents it from interacting with responsive downstream genes. Thereby GAS5 functions as a riborepressor of cell survival that sensitises arrested cells to apoptosis. It has been suggested that GAS5 can also suppress the transcriptional activity of the progesterone receptor (PR) and androgen receptor (AR) in a ligand-dependent fashion (35).

Sno-related lncRNAs (sno-lncRNAs) can also serve as ceRNAs. In HeLa and human embryonic stem cells, sno-lncRNAs associate strongly with the RNA-binding FOX family splicing regulators and alter the splicing patterns of other transcripts (36). Of note, FOX2 binds the human oestrogen receptor a (ERa), as well as GR and PR, in a ligand-independent manner and inhibits tamoxifen-mediated Era transcriptional activation. At the same time, mutations in the FOX2 RNA binding domain allow ERα antagonists to manifest agonist activity (37). This suggests that sno-lncRNAs may be an important factor in the tissue-specific agonist activity of steroid receptor ligands.

Multifunctional long noncoding RNAs are host genes for small noncoding RNAs deregulated in cancer

The miR-106b-93-25 cluster described earlier is an example of a protein-coding miRNA host gene and efficient use of genome space where the activation of a single gene locus leads to the production of multiple transcripts with different functions. Similarly, lncRNAs can function as hosts of miRNAs and other small ncRNAs. It has been estimated that at least 4% of lncRNAs host small ncRNAs (8). For example, H19 is a primary precursor for the tumour suppressor miR-675 that inhibits cell proliferation in response to stress or oncogenic signals

(38); GAS5 hosts 10 different snoRNAs in its introns (6). Interestingly, some but not all of the GAS5-encoded snoRNAs are upregulated in progressing PCa (39), suggesting that separate mechanisms control the posttranscriptional levels of RNA products derived from the same precursor transcript.

Gene expression at complex long noncoding RNA loci is disturbed in bladder and kidney cancer

The 11p15.5 locus.

The 11p15.5 locus encodes several lncRNAs and transcription factors often affected in urologic malignancies. For example, KCNQ1 opposite strand/antisense transcript 1 (KCNQ1OT1) is a cis-regulatory lncRNA associated with multiple balanced chromosomal rearrangements in Beckwith-Wiedemann syndrome (BWS) (40). Loss of maternal-specific methylation is the most frequent defect in BWS, resulting in activation of KCNQ1OT1 and silencing of the negative regulator of cell proliferation and tumour suppressor cyclin-dependent kinase inhibitor 1C (CDKN1C, p57, Kip2) (41, 42).

H19, the oldest known lncRNA, is situated only few hundred kilobases away from KCNQ1OT1 next to the conversely imprinted insulin-like growth factor 2 (IGF2) gene. H19 is a paternally imprinted, maternally expressed transcript abundant in the developing embryo. Disturbed imprinting at the IGF2/H19 locus is associated with several urologic malignancies (Fig. 3). In Wilms tumours, childhood renal neoplasms often occurring in BWS, H19 is silenced on both chromosomes and that causes biallelic expression of IGF2, thereby conferring a growth advantage to the affected cells (43). Interestingly, loss of imprinting (LOI) and biallelic expression of IGF2 is also observed in the human prostate and in urothelial cellular models of aging and senescence (44) as well as in histologically normal human prostate tissues, and it is more extensive in men with associated cancer (45).

In contrast, hypomethylation of the paternal H19 allele is reported in bladder cancer (46) where H19 functions as a trans-acting repressor that promotes cell metastasis in vitro and in vivo by repressing the transcription of the cell–cell adhesion glycoprotein CDH1 (E-cadherin) and the antagonist of WNT-signalling naked cuticle homolog 1 (NKD1). This results in loss of cellular adhesion, and the indirect activation of the WNT-signalling pathway promoting epithelial-mesenchymal transition (47). Simultaneously, H19 increases bladder cancer growth by activating the inhibitor of DNA binding 2, dominant negative helix-loophelix protein (ID2) (48).

The 14q32.3 locus.

H19 expression is directly induced by the v-myc avian myelocytomatosis viral oncogene homolog (c-MYC) (49) and loss of the p53 tumour suppressor (50), further supporting the importance of H19 as a potent oncogene. Notably, hypoxia-induced H19 levels in hepatocellular and bladder carcinoma increase the levels of genes promoting angiogenesis, cell survival, and proliferation (48, 51, 52). In contrast, the H19-encoded miR-675 is a tumour suppressor that inhibits cellular proliferation in response to stress or oncogenic signals by

decreasing the levels of insulin-like growth factor 1 receptor Igf1r (38), demonstrating the multiple functionality of transcripts derived from the H19 locus.

The 14q32.3 locus is also associated with urologic malignancies and displays an organisation very similar to 11p15.5. Two coexpressed and reciprocally imprinted genes are located at 14q32.3: the maternally expressed 3 (MEG3) and the paternally expressed delta-like 1 homolog (DLK1). MEG3 is a tumour suppressor that activates p53 in cancer via inhibition of MDM2, stimulation of p53 promoter, and selective regulation of p53 targets (53). The expression of MEG3 lncRNA is lost in many primary tumours due to gene deletion, promoter hypermethylation, or hypomethylation of the corresponding imprinting control region. MEG3 expression is lost in bladder cancer cells and PCa cells (54). In bladder cancer tissues, MEG3 levels are significantly reduced compared with normal controls, causing autophagy activation and increased cellular proliferation (55).

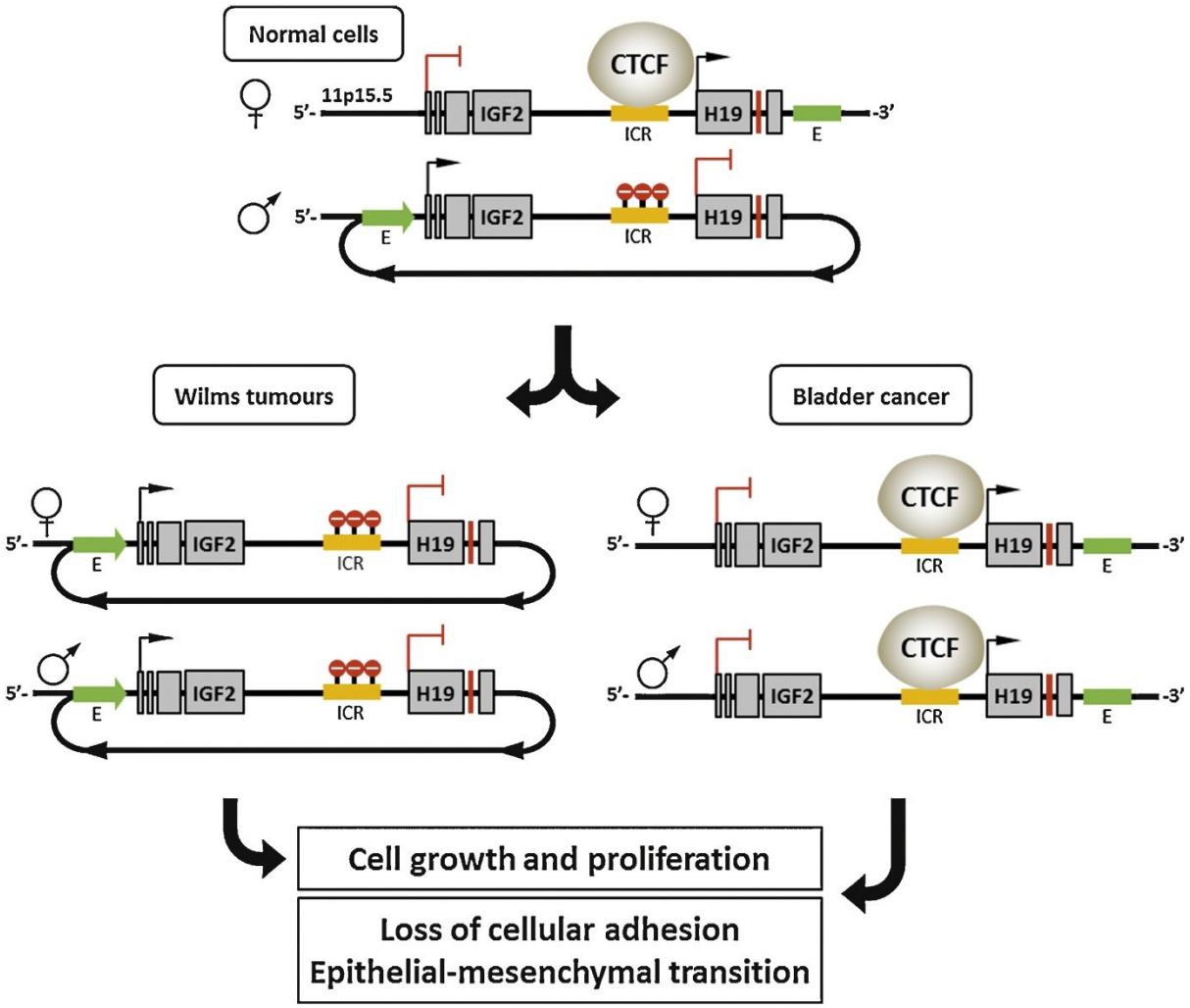


Figure 3: Genomic imprinting at the 11p15.5 locus harboring the insulin-like growth factor 2 (IGF2) and imprinted maternally expressed transcript (H19) genes in normal cells, Wilms tumours, and bladder carcinomas. Either silencing or overexpression of H19 can cause tumour growth. In normal cells, IGF2 and H19 are separated by the genomic

imprinting control region (ICR) recognised by the transcriptional regulator CTCF. IGF and H19 demonstrate monoallelic coexpression during embryogenesis and are under the control of the same enhancer elements (E) located downstream of H19. ICR is methylated on the paternal chromosome (lollypops), which blocks binding of CTCF, prevents transcription of H19, and allows activation of IGF2 by the distal enhancer. Instead, CTCF binds to the maternal chromosome and promotes H19 transcription. In this way, the methylation status of H19 balances in cis the expression of IGF2 residing on the same chromosome. In Wilms tumours, H19 is methylated at both chromosomes, which leads to expression of IGF2 from both alleles, accumulation of IGF2 protein, and cell growth stimulation. In bladder cancer, the H19 promoter region and ICR are hypomethylated, which blocks expression of IGF2 and causes the accumulation of H19 lncRNA.

The imprinted partner of MEG3, DLK1, is a candidate tumour suppressor in kidney cancer. DLK1 expression is maintained in normal kidney, but it is lost in most primary renal cell carcinomas (RCCs) and RCC-derived cell lines. Reintroduction of DLK1 increases anchorage-independent cell death and suppresses tumour growth in nude mice. The inactivation of DLK1 in RCCs is caused by gain of methylation upstream of MEG3 (56).

Downstream of MEG3, the 14q32.3 locus encodes one of the largest intergenic miRNA clusters (composed of 54 miRNAs) often deregulated in prostate, bladder, and other cancers, and several imprinted lncRNAs with unknown function(57). Interestingly, one of these lncRNAs, MEG8, hosts 31 snoRNAs and displays a genomic organisation resembling sno-lncRNAs (58).

Long noncoding RNAs are oncogenes in urologic cancers.

The metastasis associated lung adenocarcinoma transcript 1 (MALAT1/NEAT2) (59) is specifically upregulated in bladder, kidney, and a subset of PCas (12, 60–62). In urothelial carcinoma, MALAT1 is overexpressed; induces cell proliferation, migration, and survival; and promotes epithelial-mesenchymal transition by activating WNT signalling in vitro (60, 61). In RCC, MALAT1 is a fusion partner of TFEB, a transcription factor that regulates key developmental pathways in several cell lineages. The MALAT1 gene is fused to the TFEB gene, preserving the entire TFEB coding sequence and leading to a dramatic increase of TFEB protein levels and cancer progression (63, 64). MALAT1 is also involved in indirect activation of gene expression. Together with the lncRNA encoded by the taurine upregulated gene 1 (TUG1) (16), MALAT1 mediates the shuttling of the polycomb repressive complex between nuclear compartments leading to the activation or repression of growth-control genes (65). Interestingly, TUG1 is also overexpressed in urothelial carcinomas, and high TUG1 expression is associated with high grade and stage. Silencing of TUG1 in urothelial carcinoma cells causes proliferation inhibition and apoptosis induction, supporting an oncogenic function for TUG1 (66).

Urothelial cancer associated 1 (UCA1) is another lncRNA upregulated in bladder cancer (67–70). Overexpression of UCA1 enhances ERK1/2 MAPK and PI3-K/AKT kinase activity,

causing increased expression of the transcriptional coactivator p300 that regulates transcription via chromatin remodelling and both expression and phosphorylation of its interacting protein CREB, which in turn promotes cell cycle progression, carcinogenesis, and cancer invasion (68, 70). Overexpression of one of the UCA1 isoforms (UCA1a, CUDR) in bladder cancer cells antagonises cisplatin-induced apoptosis and promotes tumourigenicity in vivo, suggesting that UCA1a can serve as a new therapeutic target for bladder cancer (69).

In PCa, two androgen-responsive lncRNAs, CBR3-AS1 and CTBP1-AS, have been shown to indirectly regulate the expression of AR and downstream genes. CBR3-AS1 is one of three lncRNAs encoded in antisense to the carbonyl reductase 3 gene (CBR3). Expression of CBR3-AS1 is significantly elevated in primary tumours and PCa cells compared with normal tissues and benign prostatic hyperplasia. Silencing of CBR3-AS1 in androgen-responsive and nonresponsive LNCaP cells results in downregulation of AR, reduces cell proliferation, and induces apoptosis suggesting that CBR3-AS1 could be a novel therapeutic target in PCa (71).

CTBP1-AS is an androgen-responsive lncRNA encoded in antisense to the novel AR corepressor C-terminal binding protein 1 (CTBP1). CTBP1-AS is upregulated in PCa and promotes both hormone-dependent and castration-resistant tumour growth by antagonising the expression of CTBP1. In PCa cells, CTBP1-AS recruits the RNA-binding transcriptional repressor PSF that induces histone deacetylation of the CTBP1 promoter and transcriptional repression of CTBP1. Analysis of common CTBP1-AS/PSF target genes suggests that the CTBP1-AS/PSF tandem functions as a global androgen-controlled trans-repressor of tumour suppressor genes, thus promoting cell cycle progression (72).

Ultraconserved RNAs

Frequently originating from fragile genomic loci, ultraconserved RNAs (ucRNAs) are also candidate oncogenes with tissue- and disease-specific expression thought to function as transcriptional enhancers and regulators of alternative splicing that may interfere with the function of other RNAs through RNA-to-RNA loop interactions (73). In PCa some ucRNAs demonstrate altered expression associated with Gleason score and extraprostatic extension. In PCa cells, transcription of several ucRNAs is controlled by epigenetic mechanisms and/or androgens and correlates negatively with miRNA expression. UcRNA target analysis in PCa identified >1000 possible ucRNA-to-mRNA interactions, with enrichment of ucRNA targets in pathways related to calcium binding and RAS signalling (74).

Role of long noncoding RNAs in the treatment of patients with prostate, bladder, or kidney cancer

The altered expression of lncRNAs in urologic malignancies and their demonstrated involvement in cancer-associated cellular processes present them as attractive noninvasive biomarker candidates and open the possibility for the developing of novel therapeutic strategies.

Long noncoding RNA biomarkers

One of the oldest known lncRNAs, PCA3 (DD3), is an approved diagnostic urinary biomarker for PCa and shows potential in PCa diagnostics prior to biopsy (75). Likewise, nonmethylated DNA fragments originating from the XIST promoter have been proposed as a noninvasive serum marker in TGCT (76) and PCa (77, 78). Accumulation of UCA1 in urine sediments could be used as a sensitive and specific diagnostic and follow-up marker for patients with transitional cell carcinoma (67).

The highly tissue- and cancer-specific expression of lncRNAs and ucRNAs (11, 73) suggests that they possess diagnostic, prognostic, and/or predictive potential in urologic malignancies. For example, PCGEM1 has been characterised as a high-risk PCa marker and a potential biomarker for neoplasms responsive to chemoprevention by phytosterols (19, 79, 80). SNHG16 is positively associated with aggressive bladder cancer and chemotherapy resistance (81). Overexpression of hypoxia-inducible factor-1α antisense transcripts (HIF1A-AS) discriminates between papillary and nonpapillary RCC (82, 83). Finally, H19 in situ hybridisation in biopsies is prognostic for the early recurrence of bladder cancer (84); LOI at the H19/IGF2 locus is age associated with PCa susceptibility and could assist diagnosis (44, 45). Besides lncRNA expression, cancer risk is also associated with an enrichment of single nucleotide polymorphisms (SNPs) in lncRNA genes in PCa (85). A SNP in H19 is associated with a decreased risk of non–muscle-invasive bladder cancer (86); SNPs in PCGEM1 and PRNCR1 contribute to PCa susceptibility (21, 87).

Long noncoding RNA therapeutic concepts

Different therapeutic concepts targeting lncRNAs are currently under investigation. A direct method to "correct" the cellular levels of overexpressed oncogenic lncRNAs, widely used for research purposes, is silencing by small interfering RNA (siRNAs). SiRNAs designed to target lncRNA successfully reduce lncRNA expression (16, 73) and have been used to sensitise human fibroblasts to apoptosis in vitro and in vivo (88, 89). A similar approach implies the use of synthetic antisense DNA oligonucleotides (ASOs) complementary to lncRNA regions interacting with DNA, RNA, or proteins. ASOs can be effective agents that correct RNA gain-of-function effects and modulate the expression of expanded repeats, lncRNAs, and other mutant transcripts residing in the nucleus (89). Other possible lncRNA-targeting agents are DNAzymes, single-stranded DNA molecules able to cleave complementary sequences, engineered after the naturally occurring RNA-based ribozymes. Both classes carry future therapeutic potential as demonstrated by their use in muscle and brain diseases (90).

Although silencing oncogene expression is a logical approach for cancer therapy, one can also imagine using ceRNAs as molecular modulators of ''defective'' oncogenic miRNAs and transcription factors. Two examples described earlier are the miRNA sponge CDR1-AS (33, 34) and the protein decoy GAS5 sequestering GR (35). However, the successful application of ceRNAs in future clinical practice will require a detailed understanding of the secondary structure and functional elements of lncRNAs that is yet to be achieved.

Research has now demonstrated that regulatory elements that control lncRNA expression can be used in targeted anticancer therapy. For instance, the BC-819 DNA plasmid, which

contains H19 regulatory sequences, was evaluated as a promising and safe targeted therapy agent in phase 1 and 2 studies on pancreatic cancer (91). A similar vector is under consideration for bladder cancer therapy, where animal studies demonstrated effective tumour growth inhibition (92). Future research will have to elucidate whether any of these approaches can be successfully implemented in clinical practice.

Current challenges and future perspectives of long noncoding RNA research

The rapidly rising number of newly discovered lncRNAs and the accumulating experimental evidence elucidating their multifaceted functionality hold promise for a better understanding of cancer biology and future use in clinical practice. However, current investigations aiming at the comprehensive portrayal of lncRNA are confronted by several challenges.

At present, most novel transcripts are discovered via next-generation sequencing technologies that face computational limitations in terms of short sequence length, mapping, and de novo assembly of (fusion) transcripts originating from cancer genomes with complex structural rearrangements (eg, large deletions or insertions, chromosomal fusions, chromothripsis, or chromoplexy) (93–96). These challenges will be overcome in the near future because experimental sequencing protocols are already advancing towards their third and fourth generation where an individual transcript is analysed in its full length without the need of assembly and quantified in single cells on the background of a complex tissue (96). Recent developments in fluorescence probe design, imaging technology, and image processing enable the determination of (sub)cellular localisation and the measurement of absolute expression of endogenous transcripts in individual cells with single-molecule resolution in situ (97, 98).

The elucidation of lncRNAs function is hampered by their relatively low sequence conservation. However, the main functionality of RNA may reside in its tertiary structure determined by conserved sequence motives that assist RNA folding and are essential for protein binding. This is demonstrated by the sequential assembly of PRNCR1 and PCGEM1 with the AR (22), the structure of MEG3 imperative for its tumour suppressor functionality (54), or the multidomain organisation of steroid receptor RNA activator SRA (99). The identification and characterisation of such sequence motives and functionally active RNA domains, as opposed to individual transcripts, would provide the basis for better understanding and prediction of lncRNA activity (100).

Another key question that remains to be answered is what causes the strikingly specific expression of most lncRNAs in normal and cancerous tissues (8, 11, 73). At present, only a few reports have provided evidence for the genetic or epigenetic targeting of specific lncRNAs, for example by deletions, amplifications, fusion events, or methylation, some of which were discussed earlier. The genetic/epigenetic aberrations controlling lncRNA expression should be investigated in the future to apprehend the commonality of lncRNAs as drivers of tumourigenesis.

Conclusions

LncRNAs emerged rapidly as a diverse group of essential regulators of genetic information flow that interact with the epigenetic, transcriptional, and posttranscriptional pathways of cell proliferation, differentiation, and survival. Functional alterations of specific lncRNAs promote tumour formation, progression, and metastasis in many human malignancies including prostate, bladder, and kidney cancer. The tissue- and cancer-specific expression of lncRNAs demonstrates their potential as attractive noninvasive markers in urologic malignancies. A better understanding of the molecular nature of lncRNAs and the mechanisms by which they function in the normal and malignant cell will lead to better understanding of tumour biology and could provide novel therapeutic targets for the treatment of urologic cancers.

References

1. Consortium,E.P., Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

2. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

3. Brannan,C.I., Dees,E.C., Ingram,R.S. and Tilghman,S.M. (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, **10**, 28–36.

4. Elkin,M., Shevelev,A., Schulze,E., Tykocinsky,M., Cooper,M., Ariel,I., Pode,D., Kopf,E., de Groot,N. and Hochberg,A. (1995) The expression of the imprinted H19 and IGF-2 genes in human bladder carcinoma. *FEBS Lett*, **374**, 57–61.

5. Brown,C.J., Ballabio,A., Rupert,J.L., Lafreniere,R.G., Grompe,M., Tonlorenzi,R. and Willard,H.F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38–44.

6. Smith,C.M. and Steitz,J.A. (1998) Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol*, **18**, 6897–6909.

7. Bussemakers,M.J., van Bokhoven,A., Verhaegh,G.W., Smit,F.P., Karthaus,H.F., Schalken,J.A., Debruyne,F.M., Ru,N. and Isaacs,W.B. (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.*, **59**, 5975–9.

8. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, **22**, 1775–1789.

9. Wright,M.W. and Bruford,E.A. (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum Genomics*, **5**, 90–98.

10. Martens-Uzunova,E.S., Olvedy,M. and Jenster,G. (2013) Beyond microRNA--novel RNAs derived from small non-coding RNA and their implication in cancer. *Cancer Lett.*, **340**, 201–11.

11. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S., Kominsky,H.D., *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*, **29**, 742–749.

12. Ren,S., Peng,Z., Mao,J.H., Yu,Y., Yin,C., Gao,X., Cui,Z., Zhang,J., Yi,K., Xu,W., *et al.* (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res*, **22**, 806–821.

13. Qiao,H.P., Gao,W.S., Huo,J.X. and Yang,Z.S. (2013) Long Non-coding RNA GAS5 Functions as a Tumor Suppressor in Renal Cell Carcinoma. *Asian Pac J Cancer Prev*, **14**, 1077–1082.

14. Weakley,S.M., Wang,H., Yao,Q. and Chen,C. (2011) Expression and function of a large non-coding RNA gene XIST in human cancer. *World J Surg*, **35**, 1751–1756.

15. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.C., Hung,T., Argani,P., Rinn,J.L., *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.

16. Khalil,A.M., Guttman,M., Huarte,M., Garber,M., Raj,A., Rivea Morales,D., Thomas,K., Presser,A., Bernstein,B.E., van Oudenaarden,A., *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 11667–72.

17. Zhao,J., Ohsumi,T.K., Kung,J.T., Ogawa,Y., Grau,D.J., Sarma,K., Song,J.J., Kingston,R.E., Borowsky,M. and Lee,J.T. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–53.

18. Srikantan,V., Zou,Z., Petrovics,G., Xu,L., Augustus,M., Davis,L., Livezey,J.R., Connell,T., Sesterhenn,I.A., Yoshino,K., *et al.* (2000) PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 12216–21.

19. Petrovics,G., Zhang,W., Makarem,M., Street,J.P., Connelly,R., Sun,L., Sesterhenn,I.A., Srikantan,V., Moul,J.W. and Srivastava,S. (2004) Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene*, **23**, 605–611.

20. Fu,X., Ravindranath,L., Tran,N., Petrovics,G. and Srivastava,S. (2006) Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. *DNA Cell Biol*, **25**, 135–141.

21. Chung,S., Nakagawa,H., Uemura,M., Piao,L., Ashikawa,K., Hosono,N., Takata,R., Akamatsu,S., Kawaguchi,T., Morizono,T., *et al.* (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.*, **102**, 245–52.

22. Yang,L., Lin,C., Jin,C., Yang,J.C., Tanasa,B., Li,W., Merkurjev,D., Ohgi,K.A., Zhang,J., Evans,C.P., *et al.* (2013) lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature*, **500**, 598–602.

23. Wang,D., Garcia-Bassets,I., Benner,C., Li,W., Su,X., Zhou,Y., Qiu,J., Liu,W., Kaikkonen,M.U., Ohgi,K.A., *et al.* (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–4.

24. Wang,K.C., Yang,Y.W., Liu,B., Sanyal,A., Corces-Zimmerman,R., Chen,Y., Lajoie,B.R., Protacio,A., Flynn,R.A., Gupta,R.A., *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, **472**, 120–4.

25. Lai,F., Orom,U.A., Cesaroni,M., Beringer,M., Taatjes,D.J., Blobel,G.A. and Shiekhattar,R. (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature*, **494**, 497–501.

26. Iorio,M. V and Croce,C.M. (2012) microRNA involvement in human cancer. *Carcinogenesis*, **33**, 1126–33.

27. Salmena,L., Poliseno,L., Tay,Y., Kats,L. and Pandolfi,P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.

28. Poliseno,L., Salmena,L., Riccardi,L., Fornari,A., Song,M.S., Hobbs,R.M., Sportoletti,P., Varmeh,S., Egia,A., Fedele,G., *et al.* (2010) Identification of the miR-106b~25 microRNA cluster as a proto-oncogenic PTEN-targeting intron that cooperates with its host gene MCM7 in transformation. *Sci Signal*, **3**, ra29.

29. Poliseno,L., Salmena,L., Zhang,J., Carver,B., Haveman,W.J. and Pandolfi,P.P. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

30. Johnsson,P., Ackley,A., Vidarsdottir,L., Lui,W.O., Corcoran,M., Grander,D. and Morris,K. V (2013) A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat Struct Mol Biol*, **20**, 440–446.

31. Tay,Y., Kats,L., Salmena,L., Weiss,D., Tan,S.M., Ala,U., Karreth,F., Poliseno,L., Provero,P., Di Cunto,F., *et al.* (2011) Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, **147**, 344–357.

32. Capel,B., Swain,A., Nicolis,S., Hacker,A., Walter,M., Koopman,P., Goodfellow,P. and Lovell-Badge,R. (1993) Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell*, **73**, 1019–30.

33. Hansen,T.B., Jensen,T.I., Clausen,B.H., Bramsen,J.B., Finsen,B., Damgaard,C.K. and Kjems,J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–388.

34. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M., *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.

35. Kino,T., Hurt,D.E., Ichijo,T., Nader,N. and Chrousos,G.P. (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*, **3**, ra8.

36. Yin,Q.F., Yang,L., Zhang,Y., Xiang,J.F., Wu,Y.W., Carmichael,G.G. and Chen,L.L. (2012) Long noncoding RNAs with snoRNA ends. *Mol Cell*, **48**, 219–230.

37. Norris,J.D., Fan,D., Sherk,A. and McDonnell,D.P. (2002) A negative coregulator for the human ER. *Mol. Endocrinol.*, **16**, 459–68.

38. Keniry,A., Oxley,D., Monnier,P., Kyba,M., Dandolo,L., Smits,G. and Reik,W. (2012) The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat. Cell Biol.*, **14**, 659–65.

39. Martens-Uzunova,E.S., Jalava,S.E., Dits,N.F., van Leenders,G.J., Moller,S., Trapman,J., Bangma,C.H., Litman,T., Visakorpi,T. and Jenster,G. (2012) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*, **31**, 978–991.

40. Mitsuya,K., Meguro,M., Lee,M.P., Katoh,M., Schulz,T.C., Kugoh,H., Yoshida,M.A., Niikawa,N., Feinberg,A.P. and Oshimura,M. (1999) LIT1, an imprinted antisense RNA in the human

KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. *Hum. Mol. Genet.*, **8**, 1209–17.

41. Pandey,R.R., Mondal,T., Mohammad,F., Enroth,S., Redrup,L., Komorowski,J., Nagano,T., Mancini-Dinardo,D. and Kanduri,C. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell*, **32**, 232–46.

42. Chiesa,N., De Crescenzo,A., Mishra,K., Perone,L., Carella,M., Palumbo,O., Mussa,A., Sparago,A., Cerrato,F., Russo,S., *et al.* (2012) The KCNQ1OT1 imprinting control region and non-coding RNA: new properties derived from the study of Beckwith-Wiedemann syndrome and Silver-Russell syndrome cases. *Hum Mol Genet*, **21**, 10–25.

43. Frevel,M.A., Sowerby,S.J., Petersen,G.B. and Reeve,A.E. (1999) Methylation sequencing analysis refines the region of H19 epimutation in Wilms tumor. *J Biol Chem*, **274**, 29331–29340.

44. Fu,V.X., Schwarze,S.R., Kenowski,M.L., Leblanc,S., Svaren,J. and Jarrard,D.F. (2004) A loss of insulin-like growth factor-2 imprinting is modulated by CCCTC-binding factor down-regulation at senescence in human epithelial cells. *J Biol Chem*, **279**, 52218–52226.

45. Fu,V.X., Dobosy,J.R., Desotelle,J.A., Almassi,N., Ewald,J.A., Srinivasan,R., Berres,M., Svaren,J., Weindruch,R. and Jarrard,D.F. (2008) Aging and cancer-related loss of insulin-like growth factor 2 imprinting in the mouse and human prostate. *Cancer Res*, **68**, 6797–6802.

46. Takai,D., Gonzales,F.A., Tsai,Y.C., Thayer,M.J. and Jones,P.A. (2001) Large scale mapping of methylcytosines in CTCF-binding sites in the human H19 promoter and aberrant hypomethylation in human bladder cancer. *Hum Mol Genet*, **10**, 2619–2626.

47. Luo,M., Li,Z., Wang,W., Zeng,Y., Liu,Z. and Qiu,J. (2013) Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. *Cancer Lett*, **333**, 213–221.

48. Luo,M., Li,Z., Wang,W., Zeng,Y., Liu,Z. and Qiu,J. (2013) Upregulated H19 contributes to bladder cancer cell proliferation by regulating ID2 expression. *Febs J*, **280**, 1709–1716.

49. Barsyte-Lovejoy,D., Lau,S.K., Boutros,P.C., Khosravi,F., Jurisica,I., Andrulis,I.L., Tsao,M.S. and Penn,L.Z. (2006) The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.*, **66**, 5330–7.

50. Dugimont,T., Montpellier,C., Adriaenssens,E., Lottin,S., Dumont,L., Iotsova,V., Lagrou,C., Stéhelin,D., Coll,J. and Curgy,J.J. (1998) The H19 TATA-less promoter is efficiently repressed by wild-type tumor suppressor gene product p53. *Oncogene*, **16**, 2395–401.

51. Ayesh,S., Matouk,I., Schneider,T., Ohana,P., Laster,M., Al-Sharef,W., De-Groot,N. and Hochberg,A. (2002) Possible physiological role of H19 RNA. *Mol Carcinog*, **35**, 63–74.

52. Matouk,I.J., DeGroot,N., Mezan,S., Ayesh,S., Abu-lail,R., Hochberg,A. and Galun,E. (2007) The H19 non-coding RNA is essential for human tumor growth. *PLoS One*, **2**, e845.

53. Zhou,Y., Zhong,Y., Wang,Y., Zhang,X., Batista,D.L., Gejman,R., Ansell,P.J., Zhao,J., Weng,C. and Klibanski,A. (2007) Activation of p53 by MEG3 non-coding RNA. *J Biol Chem*, **282**, 24731–24742.

54. Zhou,Y., Zhang,X. and Klibanski,A. (2012) MEG3 noncoding RNA: a tumor suppressor. *J. Mol. Endocrinol.*, **48**, R45–53.

55. Ying,L., Huang,Y., Chen,H., Wang,Y., Xia,L., Chen,Y., Liu,Y. and Qiu,F. (2013) Downregulated MEG3 activates autophagy and increases cell proliferation in bladder cancer. *Mol Biosyst*, **9**, 407–411.

56. Kawakami,T., Chano,T., Minami,K., Okabe,H., Okada,Y. and Okamoto,K. (2006) Imprinted DLK1 is a putative tumor suppressor gene and inactivated by epimutation at the region upstream of GTL2 in human renal cell carcinoma. *Hum Mol Genet*, **15**, 821–830.

57. Benetatos,L., Hatzimichael,E., Londin,E., Vartholomatos,G., Loher,P., Rigoutsos,I. and Briasoulis,E. (2013) The microRNAs within the DLK1-DIO3 genomic region: involvement in disease pathogenesis. *Cell. Mol. Life Sci.*, **70**, 795–814.

58. Cavaille,J., Seitz,H., Paulsen,M., Ferguson-Smith,A.C. and Bachellerie,J.P. (2002) Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum Mol Genet*, **11**, 1527–1538.

59. Ji,P., Diederichs,S., Wang,W., Boing,S., Metzger,R., Schneider,P.M., Tidow,N., Brandt,B., Buerger,H., Bulk,E., *et al.* (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, **22**, 8031–8041.

60. Han,Y., Liu,Y., Nie,L., Gui,Y. and Cai,Z. (2013) Inducing cell proliferation inhibition, apoptosis, and motility reduction by silencing long noncoding ribonucleic acid metastasis-associated lung adenocarcinoma transcript 1 in urothelial carcinoma of the bladder. *Urology*, **81**, 209 e1–7.

61. Ying,L., Chen,Q., Wang,Y., Zhou,Z., Huang,Y. and Qiu,F. (2012) Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. *Mol Biosyst*, **8**, 2289–2294.

62. Lin,R., Maeda,S., Liu,C., Karin,M. and Edgington,T.S. (2007) A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene*, **26**, 851–8.

63. Davis,I.J., Hsi,B.-L., Arroyo,J.D., Vargas,S.O., Yeh,Y.A., Motyckova,G., Valencia,P., Perez-Atayde,A.R., Argani,P., Ladanyi,M., *et al.* (2003) Cloning of an Alpha-TFEB fusion in renal tumors harboring the t(6;11)(p21;q13) chromosome translocation. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 6051–6.

64. Kuiper,R.P., Schepens,M., Thijssen,J., van Asseldonk,M., van den Berg,E., Bridge,J., Schuuring,E., Schoenmakers,E.F.P.M. and van Kessel,A.G. (2003) Upregulation of the transcription factor TFEB in t(6;11)(p21;q13)-positive renal cell carcinomas due to promoter substitution. *Hum. Mol. Genet.*, **12**, 1661–9.

65. Yang,L., Lin,C., Liu,W., Zhang,J., Ohgi,K.A., Grinstein,J.D., Dorrestein,P.C. and Rosenfeld,M.G. (2011) ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell*, **147**, 773–788.

66. Han,Y., Liu,Y., Gui,Y. and Cai,Z. (2013) Long intergenic non-coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder. *J. Surg. Oncol.*, **107**, 555–9.

67. Wang,X.S., Zhang,Z., Wang,H.C., Cai,J.L., Xu,Q.W., Li,M.Q., Chen,Y.C., Qian,X.P., Lu,T.J., Yu,L.Z., *et al.* (2006) Rapid identification of UCA1 as a very sensitive and specific unique marker for human bladder carcinoma. *Clin Cancer Res*, **12**, 4851–4858.

68. Wang,F., Li,X., Xie,X., Zhao,L. and Chen,W. (2008) UCA1, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. *FEBS Lett*, **582**, 1919–1927.

69. Wang,Y., Chen,W., Yang,C., Wu,W., Wu,S., Qin,X. and Li,X. (2012) Long non-coding RNA UCA1a(CUDR) promotes proliferation and tumorigenesis of bladder cancer. *Int J Oncol*, **41**, 276–284.

70. Yang,C., Li,X., Wang,Y., Zhao,L. and Chen,W. (2012) Long non-coding RNA UCA1 regulated cell cycle distribution via CREB through PI3-K dependent pathway in bladder carcinoma cells. *Gene*, **496**, 8–16.

71. Cui,Z., Ren,S., Lu,J., Wang,F., Xu,W., Sun,Y., Wei,M., Chen,J., Gao,X., Xu,C., *et al.* (2012) The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor. *Urol. Oncol.*, 10.1016/j.urolonc.2011.11.030.

72. Takayama,K.-I., Horie-Inoue,K., Katayama,S., Suzuki,T., Tsutsumi,S., Ikeda,K., Urano,T., Fujimura,T., Takagi,K., Takahashi,S., *et al.* (2013) Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. *EMBO J.*, 10.1038/emboj.2013.99.

73. Calin,G.A., Liu,C.G., Ferracin,M., Hyslop,T., Spizzo,R., Sevignani,C., Fabbri,M., Cimmino,A., Lee,E.J., Wojcik,S.E., *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, **12**, 215–229.

74. Hudson,R.S., Yi,M., Volfovsky,N., Prueitt,R.L., Esposito,D., Volinia,S., Liu,C.G., Schetter,A.J., Van Roosbroeck,K., Stephens,R.M., *et al.* (2013) Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer. *Mol Cancer*, **12**, 13.

75. Crawford,E.D., Rove,K.O., Trabulsi,E.J., Qian,J., Drewnowska,K.P., Kaminetsky,J.C., Huisman,T.K., Bilowus,M.L., Freedman,S.J., Glover,W.L., *et al.* (2012) Diagnostic performance of PCA3 to detect prostate cancer in men with increased prostate specific antigen: a prospective study of 1,962 cases. *J. Urol.*, **188**, 1726–31.

76. Kawakami,T., Okamoto,K., Ogawa,O. and Okada,Y. (2004) XIST unmethylated DNA fragments in male-derived plasma as a tumour marker for testicular cancer. *Lancet (London, England)*, **363**, 40–2.

77. Laner,T., Schulz,W.A., Engers,R., Muller,M. and Florl,A.R. (2005) Hypomethylation of the XIST gene promoter in prostate cancer. *Oncol Res*, **15**, 257–264.

78. Song,M.A., Park,J.H., Jeong,K.S., Park,D.S., Kang,M.S. and Lee,S. (2007) Quantification of CpG methylation at the 5'-region of XIST by pyrosequencing from human serum. *Electrophoresis*, **28**, 2379–2384.

79. Ifere,G.O. and Ananaba,G.A. (2009) Prostate cancer gene expression marker 1 (PCGEM1): a patented prostate- specific non-coding gene and regulator of prostate cancer progression. *Recent Pat DNA Gene Seq*, **3**, 151–163.

80. Ifere,G.O., Barr,E., Equan,A., Gordon,K., Singh,U.P., Chaudhary,J., Igietseme,J.U. and Ananaba,G.A. (2009) Differential effects of cholesterol and phytosterols on cell proliferation, apoptosis and expression of a prostate specific gene in prostate cancer cell lines. *Cancer Detect Prev*, **32**, 319–328.

81. Zhu,Y., Yu,M., Li,Z., Kong,C., Bi,J., Li,J. and Gao,Z. (2011) ncRAN, a newly identified long noncoding RNA, enhances human bladder tumor growth, invasion, and survival. *Urology*, **77**, 510 e1–5.

82. Thrash-Bingham,C.A. and Tartof,K.D. (1999) aHIF: a natural antisense transcript overexpressed in human renal cancer and during hypoxia. *J Natl Cancer Inst*, **91**, 143–151.

83. Bertozzi,D., Iurlaro,R., Sordet,O., Marinello,J., Zaffaroni,N. and Capranico,G. (2011) Characterization of novel antisense HIF-1alpha transcripts in human cancers. *Cell Cycle*, **10**, 3189–3197.

84. Ariel,I., Sughayer,M., Fellig,Y., Pizov,G., Ayesh,S., Podeh,D., Libdeh,B.A., Levy,C., Birman,T., Tykocinski,M.L., *et al.* (2000) The imprinted H19 gene is a marker of early recurrence in human bladder carcinoma. *Mol Pathol*, **53**, 320–323.

85. Jin,G., Sun,J., Isaacs,S.D., Wiley,K.E., Kim,S.T., Chu,L.W., Zhang,Z., Zhao,H., Zheng,S.L., Isaacs,W.B., *et al.* (2011) Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis*, **32**, 1655–1659.

86. Verhaegh,G.W., Verkleij,L., Vermeulen,S.H., den Heijer,M., Witjes,J.A. and Kiemeney,L.A. (2008) Polymorphisms in the H19 gene and the risk of bladder cancer. *Eur Urol*, **54**, 1118–1126.

87. Xue,Y., Wang,M., Kang,M., Wang,Q., Wu,B., Chu,H., Zhong,D., Qin,C., Yin,C., Zhang,Z., *et al.* (2013) Association between lncrna PCGEM1 polymorphisms and prostate cancer risk. *Prostate Cancer Prostatic Dis*.

88. Hung,T., Wang,Y., Lin,M.F., Koegel,A.K., Kotake,Y., Grant,G.D., Horlings,H.M., Shah,N., Umbricht,C., Wang,P., *et al.* (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat. Genet.*, **43**, 621–9.

89. Wheeler,T.M., Leger,A.J., Pandey,S.K., MacLeod,A.R., Nakamori,M., Cheng,S.H., Wentworth,B.M., Bennett,C.F. and Thornton,C.A. (2012) Targeting nuclear RNA for in vivo correction of myotonic dystrophy. *Nature*, **488**, 111–5.

90. Mastroyiannopoulos,N.P., Uney,J.B. and Phylactou,L.A. (2010) The application of ribozymes and DNAzymes in muscle and brain. *Molecules*, **15**, 5460–72.

91. Hanna,N., Ohana,P., Konikoff,F.M., Leichtmann,G., Hubert,A., Appelbaum,L., Kopelman,Y., Czerniak,A. and Hochberg,A. (2012) Phase 1/2a, dose-escalation, safety, pharmacokinetic and preliminary efficacy study of intratumoral administration of BC-819 in patients with unresectable pancreatic cancer. *Cancer Gene Ther.*, **19**, 374–81.

92. Amit,D. and Hochberg,A. (2010) Development of targeted therapy for bladder cancer mediated by a double promoter plasmid expressing diphtheria toxin under the control of H19 and IGF2-P4 regulatory sequences. *J Transl Med*, **8**, 134.

93. Berger,M.F., Lawrence,M.S., Demichelis,F., Drier,Y., Cibulskis,K., Sivachenko,A.Y., Sboner,A.,

Esgueva,R., Pflueger,D., Sougnez,C., *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–20.

94. Teles Alves,I., Hiltemann,S., Hartjes,T., van der Spek,P., Stubbs,A., Trapman,J. and Jenster,G. (2013) Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. *Hum. Genet.*, 10.1007/s00439-013-1308-1.

95. Shen,M.M. (2013) Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell*, **23**, 567–9.

96. McGinn,S. and Gut,I.G. (2013) DNA sequencing - spanning the generations. *N. Biotechnol.*, **30**, 366–72.

97. Kwon,S. (2013) Single-molecule fluorescence in situ hybridization: quantitative imaging of single RNA molecules. *BMB Rep*, **46**, 65–72.

98. Itzkovitz,S. and van Oudenaarden,A. (2011) Validating transcripts with probes and imaging technology. *Nat Methods*, **8**, S12–9.

99. Novikova,I. V, Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res*, **40**, 5034–5051.

100. Mercer,T.R. and Mattick,J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*, **20**, 300–307.

Chapter 3

Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer

René Böttcher[1,3], A. Marije Hoogland[2], Natasja Dits[1], Esther I. Verhoef[2], Charlotte Kweldam[2], Piotr Waranecki[1], Chris H Bangma[1], Geert J.L.H. van Leenders[2], Guido Jenster[1]

1 Dept. of Urology, Erasmus MC, Rotterdam, The Netherlands
2 Dept. of Pathology, Erasmus MC, Rotterdam, The Netherlands
3 Dept. of Bioinformatics, Technical University of Applied Sciences Wildau, Wildau, Germany

Published in

Oncotarget. 2015;6(6):4036-50

Supplementary Material is available via

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4414171

Abstract

Current prostate cancer (PCa) biomarkers such as PSA are not optimal in distinguishing cancer from benign prostate diseases and predicting disease outcome. To discover additional biomarkers, we investigated PCa-specific expression of novel unannotated transcripts. Using the unique probe design of Affymetrix Human Exon Arrays, we identified 334 candidates (EPCATs), of which 15 were validated by RT-PCR. Combined into a diagnostic panel, 11 EPCATs classified 80% of PCa samples correctly, while maintaining 100% specificity. High specificity was confirmed by *in situ* hybridization for EPCAT4R966 and EPCAT2F176 (SChLAP1) on extensive tissue microarrays. Besides being diagnostic, EPCAT2F176 and EPCAT4R966 showed significant association with pT-stage and were present in PIN lesions. We also found EPCAT2F176 and EPCAT2R709 to be associated with development of metastases and PCa-related death, and EPCAT2F176 to be enriched in lymph node metastases. Functional significance of expression of 9 EPCATs was investigated by siRNA transfection, revealing that knockdown of 5 different EPCATs impaired growth of LNCaP and 22RV1 PCa cells. Only the minority of EPCATs appear to be controlled by androgen receptor or ERG. Although the underlying transcriptional regulation is not fully understood, the novel PCa-associated transcripts are new diagnostic and prognostic markers with functional relevance to prostate cancer growth.

Introduction

Despite continuous research efforts over the past decades, prostate cancer (PCa) remains one of the leading causes of male cancer deaths, with an estimated 70,100 deaths in Europe in 2014. Incidence rates are highest in countries of the western hemisphere including Europe, North America and Oceania, which can be partly explained by the widely applied blood test for prostate specific antigen (PSA) [1,2]. Although the serum PSA level offers high sensitivity for PCa detection, its specificity is limited as PSA levels can also be elevated in benign prostate diseases such as benign prostate hyperplasia (BPH) and prostatitis. Thus, the most important drawback of PSA screening is a high number of false positives leading to unnecessary biopsies and overtreatment of patients due to a lack of prognostic markers. Up to date this remains a challenge and additional prognostic factors, such as disease associated genes, are needed [3].

Earlier studies discovered several other PCa- associated genes, among them two long non-coding RNAs (lncRNAs) that show disease-associated overexpression, PCGEM1 and PCA3 (DD3) [4,5]. The latter has since been extensively studied as diagnostic urine marker for PCa, offering better performance for detecting PCa when compared to PSA [6]. With the introduction of high throughput technologies, such as tiling arrays and next generation sequencing, several other PCa-associated lncRNAs such as PRNCR1, PCAT1, PCAT18, PCAT29 and SChLAP1 were identified [7–14].

LncRNAs have been associated with several functions, including epigenetic regulation of gene expression by acting as regulatory factors in *cis*, as well as in *trans* by involvement in chromatin remodeling [15–18]. Additionally, direct binding to active androgen receptor (AR) and recruitment of additional factors for AR-mediated gene expression has been reported [19]. However, a recent study found contradicting evidence for these findings and thus further research is required to clarify lncRNA involvement in AR activity [20]. Still, many functional relationships of lncRNAs as well as their tissue-specific regulation remain unclear. Currently, lncRNAs are gaining more interest as potential biomarkers for various malignant diseases, due to their highly tissue-specific expression profiles [17,21].

In this study, we set out to discover novel PCa-specific lncRNAs based on Affymetrix Human Exon Arrays by adapting a cancer outlier profile analysis (COPA, [22]). Our approach made use of the unique design of these arrays, which include probes against predicted sequences (‘full’) next to probes targeting known sequences (‘core’ and ‘extended’). This type of microarray has recently been successfully adapted for lncRNA profiling, showing the general potential of the platform in lncRNA studies [11]. To increase reliability of our results, we combined three Affymetrix Human Exon Array datasets and searched for reoccurring outlier patterns indicating novel transcripts. We then used RNA-sequencing (RNA-seq) data to refine our transcript definitions and subsequently validated them via RT-PCR. Computational evaluation of the validated transcripts confirmed absence of protein coding potential, suggesting that these transcripts are indeed lncRNAs. Two transcripts were chosen for staining of tissue microarrays using *in situ* hybridization and successfully discriminated PCa

from normal adjacent prostate (NAP) and benign prostate tissue.

Results

334 candidate PCa-associated transcripts were identified

Novel transcript candidates were identified by searching for unannotated Affymetrix Human Exon Array transcript clusters (TCs) that showed a PCa-specific outlier profile using a COPA transformation [22]. After removing all TCs targeting known genes, we discarded TCs with fewer than 5% outliers in cancerous samples and with outliers in control groups. All remaining TCs were then grouped into 'EPCATs' (Erasmus MC PCa-associated transcripts) based on proximity, strand and similarity in expression (see Figure 1). EPCAT names were assigned to directly indicate genomic location and are based on chromosome, strand and a unique identifier. For instance, EPCAT2F176 (SChLAP1) is located on the forward strand of chromosome 2. EPCATs had to be present in at least two datasets to be considered for further analysis. Differences between datasets (i.e. missing parts in one or the other) were resolved by a union of all TCs involved in a particular EPCAT to maximize its size. Our meta-analysis of three available Exon Array datasets resulted in 334 EPCATs comprising 2086 TCs that exhibited a prostate cancer-specific expression profile (see Supplementary Tables 1 – 2). We observed that combining several datasets severely reduced the number of EPCATs identified by one dataset alone, suggesting a reduction in false positives in doing so (see Figure 2a). Next, we classified the identified EPCATs based on their genomic origin with regard to UCSC known genes, and observed that 75 EPCATs were being classified as intergenic or antisense transcripts. The majority of EPCATs (259) overlapped / extended either 5' or 3' ends or was located in intronic regions of genes known to LNCipedia [23] or UCSC (see Figure 2b).

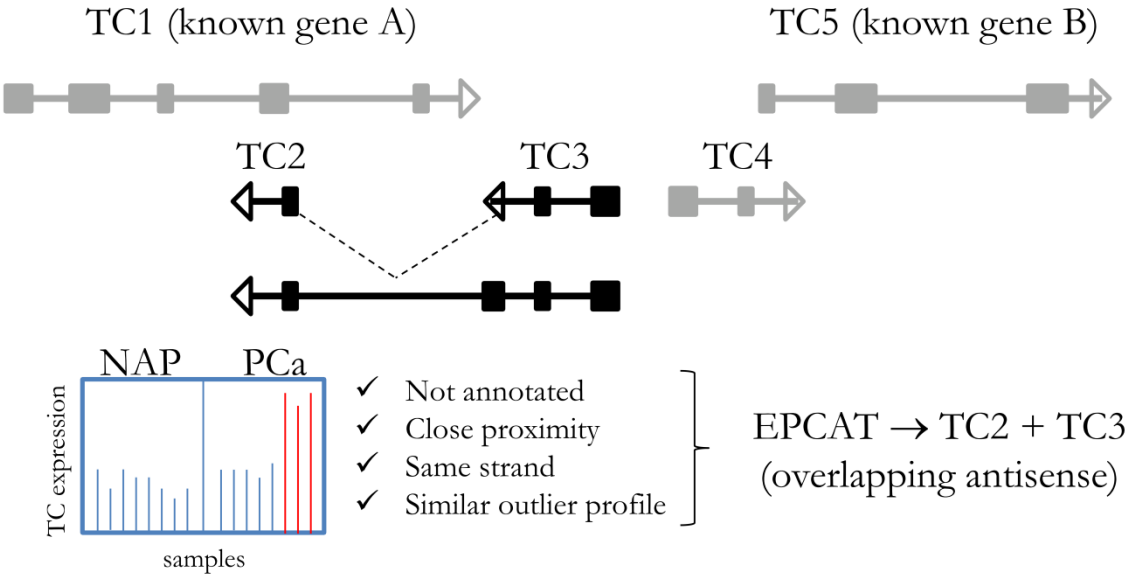


Figure 1: **Principle steps of EPCAT identification**. Affymetrix transcript clusters that had no annotation assigned were grouped into one locus if they were located on the same strand in close proximity (<250 kb) and showed a similar PCa-specific outlier profile (transcript clusters TC2 and TC3). Transcript clusters that did not meet these criteria were not included in the particular EPCAT.

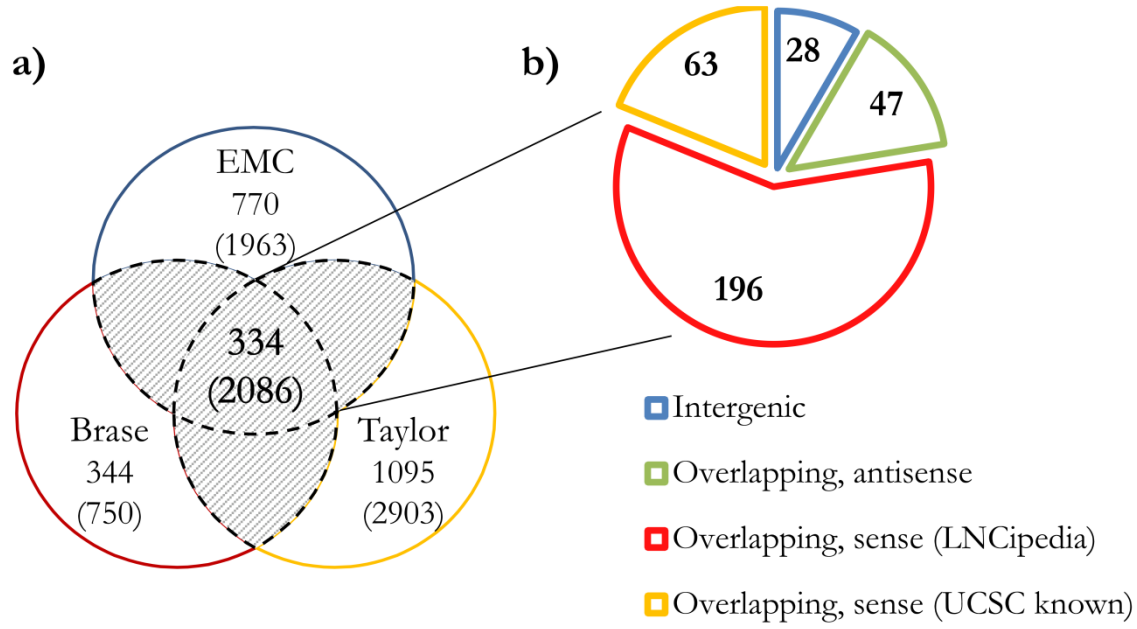


Figure 2: **Total number and classification of EPCATs**. a) Total number of EPCATs identified by each individual dataset as well as a combination of at least two datasets (shaded area, 334 EPCATs). b) Classification of these 334 EPCATs based on their relative position to LNCipedia [23] genes. UCSC known gene annotations were selected if no overlap with LNCipedia was found. Overlaps include cases in which an EPCAT overlaps and extends the 5' or 3' ends of known genes or resides in an intron.

Visual inspection of these results confirmed that similar PCa-specific expression patterns occurred in all three datasets with TCs grouped into one EPCAT following the same PCa-specific outlier profile (see Figure 3 and Supplementary Figures 1 – 6 for a subset of 15 EPCATs that were subsequently PCR-validated). We also inspected EPCAT expression in other publicly available datasets comprising samples from lung, brain, breast, colorectal and gastric cancer tissue as well as several normal tissues. For most of the EPCATs, expression was very low in virtually all samples, indicating a PCa-specific expression of these transcripts similar to other previously reported lncRNAs ([12,24], see Supplementary Figures 3 – 6). However, some EPCATs such as EPCAT5R633 and EPCATXR234 were detected in multiple lung, colorectal and breast tumors and appear deregulated in different cancer types. To gain insight into their transcriptional regulation, we tested whether any EPCATs are androgen regulated by incorporating a publicly available dataset of R1881 treated LNCaP cells. We observed that out of 301 EPCATs expressed in LNCaP 31 were significantly associated with androgen treatment and showed more than 50% increase or decrease in expression ($p < 0.05$; 13 up-, 18 downregulated, see Supplementary Figure 7). In addition, we tested for coexpression with known outlier genes ERG and ETV1 [22] by Spearman's correlation coefficient, and found that 17 EPCATs showed significant correlation with ERG (Spearman's $\rho \geq 0.5$ and $p < 0.05$, see Supplementary Figure 8), while no significant coexpression with ETV1 was observed. Public ChIP-seq data [25] targetting AR and ERG was used as second

source of evidence for AR and ERG regulation. We found that 15 of the 33 differentially expressed EPCATs (including 50 kb flanks) had overlapping AR peaks, whereas ERG peaks were found for 4 of the 17 coexpressed EPCATs (see methods).

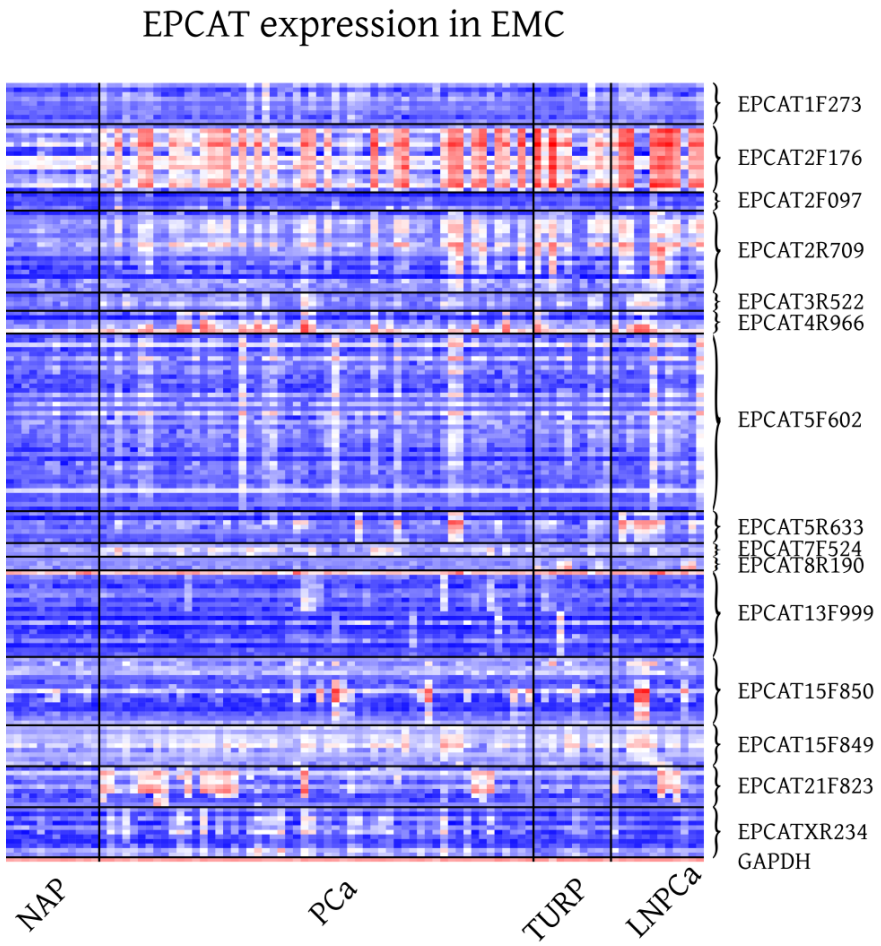


Figure 3: **Expression of 15 RT-PCR validated EPCATs in EMC Exon Array samples.** EMC (GSE41408, [27]), comprised localized prostate cancer obtained via radical prostatectomy (PCa), transurethral resection of the prostate (TURP), lymph node metastasis (LNPCa) and normal adjacent prostate (NAP) tissue.

To gather more evidence for the existence of our transcript candidates, we performed a reference guided assembly of RNA-seq data obtained from 18 patients with localized PCa as well as 5 samples from lymph node metastases. We used Cufflinks [26] to predict intron-exon boundaries in the genomic regions of the EPCATs while masking known annotated genes, which resulted in 222 predicted transcripts. We chose 20 well defined candidates that showed high expression and added additional candidate exons after manual evaluation of several genomic loci. We also included EPCAT8R190, which was initially filtered out due to its presence in only one dataset (EMC), but was subsequently discovered as a candidate due to its high expression in castration resistant prostate cancer (CRPC). We were able to design working RT-PCR primers for 15 out of these 21 candidates and validated their expression in 6

prostate cancer cell lines (see Figure 4 and Supplementary Table 3). The primers were designed intron spanning, allowing us to PCR from exon to exon, and validated exons were Sanger sequenced. Individual exons of an EPCAT showed the same expression pattern throughout our cell line panel, whereas expression patterns differed between different EPCATs, indicating independent expression and regulation. To obtain full length sequences, a λgt11 library containing cDNA from the LNCaP cell line was used (see Materials and Methods).

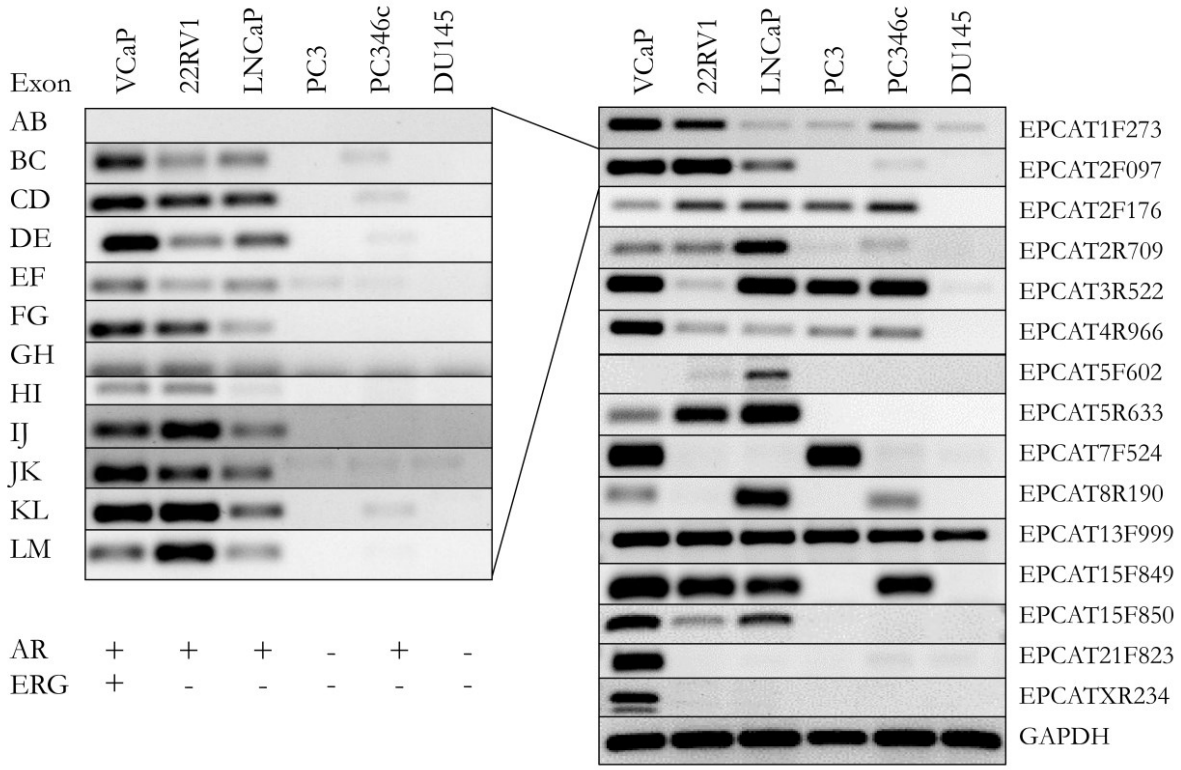


Figure 4: **Validation of 15 EPCATs in 6 prostate cancer cell lines**. Intron-spanning primers were designed for each EPCAT. Exons of one transcript followed similar expression patterns (left side). Only the most representative and optimal primer set for an EPCAT is shown in the right panel. These primers were also used to design Taqman probes (see Supplementary Table 9 – 10). AR and TMPRSS2-ERG status for each cell line are indicated as present (+) or absent (-).

EPCATs can serve as diagnostic markers in patient tissues

TaqMan RT-PCR was used to quantify expression of the 15 EPCATs in two separate patient cohorts, however, only 11 EPCATs had working TaqMan probes and were subsequently quantified. The first cohort comprised a subset of patients also present in the EMC Exon Array dataset and allowed comparison between qRT-PCR and Exon Arrays for the EPCATs. Therefore, we treated this cohort as a training set and used the second, independent cohort as validation set. Comparing expression measurements of qRT-PCR with the averaged expression values of all TCs of an EPCAT yielded varying concordance between both

techniques (average R^2 = 0.58, see Supplementary Figure 9). These results indicated that not all EPCATs were sufficiently represented by Affymetrix TCs and that RNA-seq data is essential for defining gene structures. Next, a receiver operator characteristic (ROC) was created using the test cohort to maximize area under curve (AUC) by weighing each EPCAT in the diagnostic panel. When applying the same panel to the validation cohort, an AUC of 0.87 confirmed high specificity and sensitivity for PCa diagnosis (see Supplementary Figure 10).

Two lncRNAs in 2q31.3 are associated with prostate cancer progression

To evaluate possible prognostic value of the 15 validated EPCATs from our EMC Exon Array dataset, we characterized their expression profiles in 54 patients with clinical follow-up (see [27] for further information). We performed a retrospective analysis for prediction of prostate cancer-related death (PCaD), development of clinical metastases (PCaMets) after radical prostatectomy (RP) as well as biochemical recurrence (BCR) defined by PSA progression after RP. Samples were clustered into two groups using Partition Around Medoids (PAM) and significant association with clinical endpoints was tested using a bootstrapping analysis and label permutation to calculate p-values (see methods). Using FDR correction, we observed that EPCAT2R709 and EPCAT2F176 (SChLAP1) showed significant association with PCaMets and PCaD. To evaluate whether any EPCAT could discriminate poor clinical outcome, we used a Kaplan-Meier analysis for the same clinical endpoints. Again, EPCAT2F176 and EPCAT2R709 showed a significant association with PCaMets and PCaD (see Supplementary Figures 11 – 14 and Supplementary Tables 4a – 4d). Interestingly, both EPCAT loci are located in chromosome 2q31.3, with EPCAT2R709 being found on the antisense strand, approximately 120 kb upstream of the first exon of EPCAT2F176. Additionally, both EPCATs show similar expression profiles (Spearman's $\rho = 0.79$ for all samples analyzed via qRT-PCR, $\rho = 0.93$ for EMC Exon Arrays).

Evaluation of coding potential and conservation

We evaluated if any of the 15 PCR-validated EPCATs exhibits protein coding potential using two approaches: iSeeRNA and PhyloCSF [28,29]. iSeeRNA classified all processed EPCATs as non-coding, however, EPCAT13F999 did not pass minimum length requirements (200 bp). We used all known coding RefSeq genes (36,818) as positive control, of which 34,476 (93.64%) were classified as protein coding and 2342 (6.36%) as non-coding. For PhyloCSF, known coding genes GAPDH and ERG were used as positive controls. Both genes were assigned high positive scores by PhyloCSF, as compared to negative scores for all EPCATs indicating no coding potential (see Supplementary Figure 15). Sequence conservation of the EPCATs was evaluated using per-base conservation scores from UCSC (PhyloP) for several genome panels. 1000 randomly picked coding genes in the UCSC RefSeq table as well as 1000 Repeat regions served as controls. The results illustrate that EPCAT sequences are overall less conserved than protein coding sequences, while being more conserved than most Repeat regions, which is concordant with previous findings ([17,24], see Supplementary Figure 16).

In situ hybridization revealed diagnostic power and prognostic value

To investigate whether EPCATs can serve as potential pathological tissue markers markers and specifically distinguish cancerous from normal prostate tissues, we stained tissue microarrays (TMAs) for presence of the two EPCATs showing highest expression among our 11 qRT-PCR quantified transcripts (EPCAT2F176 / SChLAP1 and EPCAT4R966). Due to their non-coding nature, we used *in situ* hybridization (ISH) to directly target the RNA molecules. All four TMAs comprised a total of 418 PCa samples from RPs, 120 transurethral resections of the prostate (TURP, 65 hormone refractory, 55 hormone sensitive), 119 lymph node metastasis (LNPCa) and 113 normal adjacent prostate samples (NAP), as well as normal prostate obtained via 81 TURPs, 5 total pelvic exenterations (TE) and 48 radical cystoprostatectomies (RCP). Normal tissue samples from kidney, liver, placenta as well as a sample containing urothelial cell carcinoma served as control (see Supplementary Tables 5a – 5b). After TMA scoring, we observed that all 4 control tissues on TMA 1 and 2 were indeed negative (score = 0) for both EPCATs, which showed PCa-specific expression as expected from our previous findings (see Figure 5a – 5j and Supplementary Table 6). Moreover, we found significant association with pathological stage, whereas other clinical parameters (Gleason score, surgical margins, pre-treatment PSA) were not significantly associated (see Supplementary Table 7a – 7d). Normal prostate samples of patients without prostate cancer showed complete absence of EPCAT expression (see Supplementary Figure 16), whereas 12 NAP samples (10.62%) exhibited higher expression levels compared to samples from normal prostate (Figure 6). In a ROC analysis, both EPCATs showed high specificity and limited sensitivity in distinguishing cancerous samples when used individually (28.61% PCa samples positive, AUC = 0.66 for EPCAT2F176 / SChLAP1 and 28.01% PCa samples positive, AUC = 0.65 for EPCAT4R966). Combining both EPCATs, we were able to correctly classify 39.4% of the cancer samples in our cohort while maintaining a specificity of 100% (AUC = 0.71).

Using ISH also allowed us to study subcellular localization of the EPCATs, revealing that both transcripts are present in the cytoplasm as well as the nucleus, with EPCAT2F176 showing a tendency to be more nuclear than cytoplasmic, consistent with previous findings [12]. Furthermore, we also identified several prostate intraepithelial neoplasia (PIN) lesions that showed positive staining for the EPCATs (7 / 21 lesions for EPCAT2F176 (33.3%), 1 / 21 lesion for EPCAT4R966 (4.8%), see Figure 6g – 6j).

We used our third TMA comprising 119 samples to evaluate EPCAT expression in lymph nodes of patients undergoing a lymph node exploration in addition to RP. We found that out of 73 samples containing tumor tissue, 46 were positive for EPCAT2F176 (63.0%), representing a significant increase in number of positive samples compared to localized PCa ($p = 0.0404$, Fisher's exact test). For EPCAT4R966, tumor was present in 71 of the sliced cores, of which 16 were stained positive (22.5%; $p = 0.3866$, Fisher's exact test). Furthermore, all tumor free samples were found to be negative.

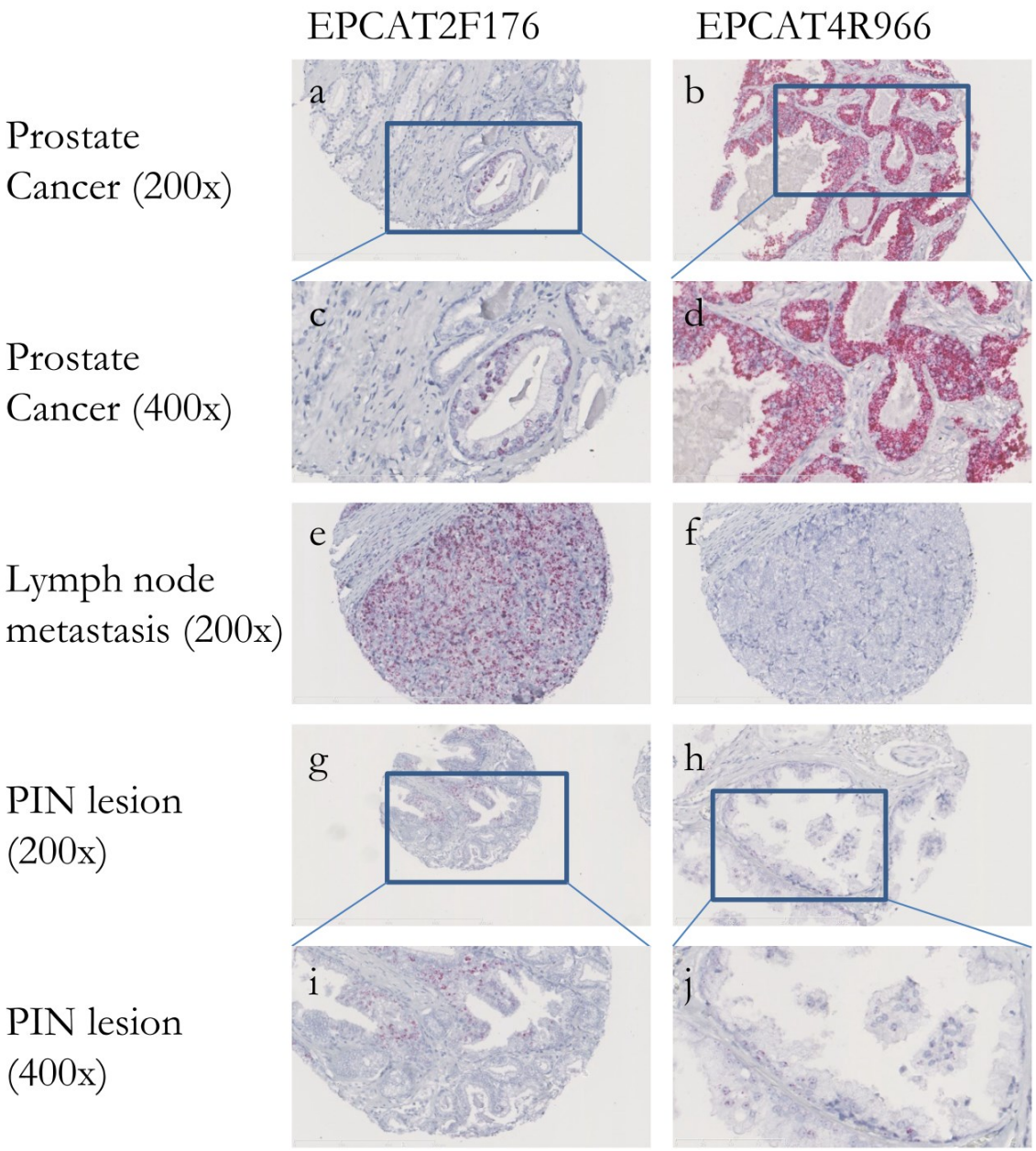


Figure 5: *In situ* hybridization of two EPCATs in prostate cancer tissues. (**a – d**) Both EPCAT2F176 as well as EPCAT4R966 show highly specific expression in PCa cells, whereas surrounding stromal tissue scored negative. (**e – f**) Lymph node metastases also scored positive for both EPCATs and complementary expression could be observed when comparing the same tissue cores, highlighting their added diagnostic potential. (**g – j**) PIN lesions were also found positive, indicating EPCAT expression as an early event in cancer development.

As for our fourth TMA comprising hormone refractory and hormone sensitive patient samples, we did not observe a significant correlation of hormonal status with any EPCAT nor a combination of both. EPCAT2F176 was found positive in 61 out of 109 TURP samples containing tumor tissue (55.9%), whereas 41 out of 103 tumor containing samples were positive for EPCAT4R966 (39.8%, see Supplementary Table 6).

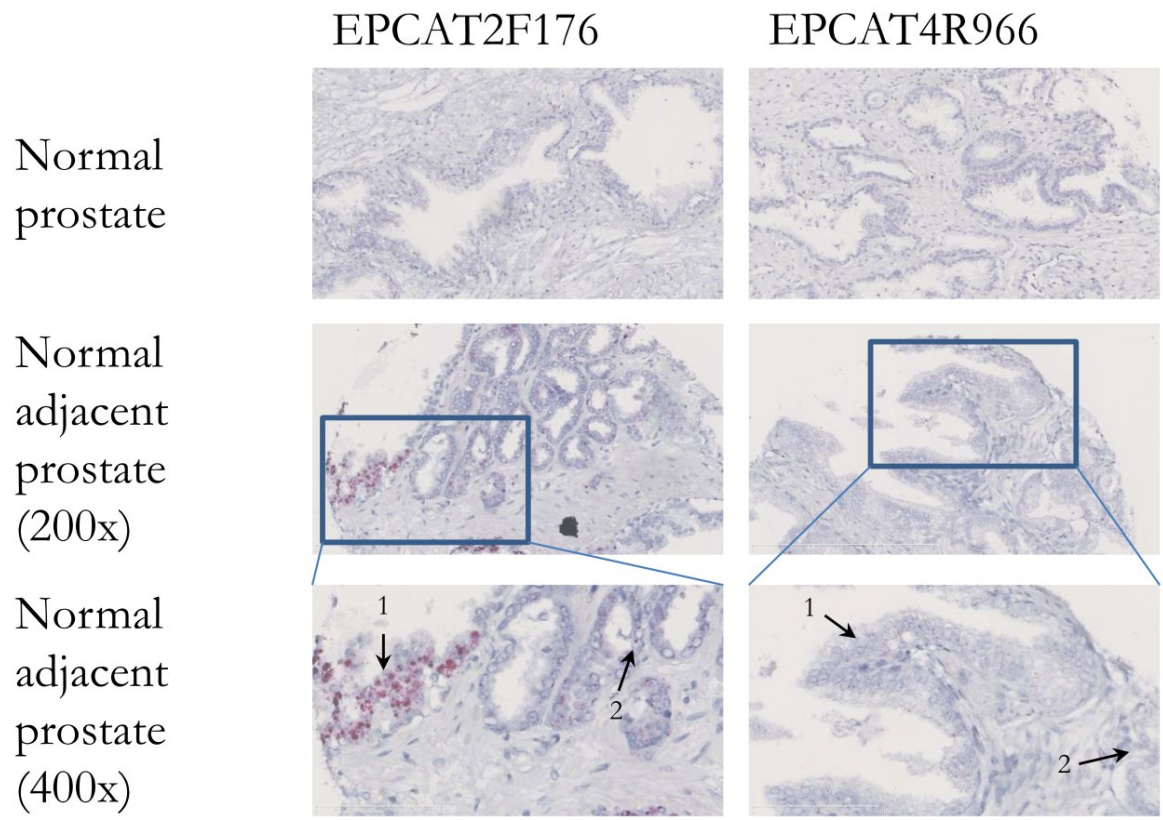


Figure 6: *In situ* hybridization of two EPCATs in normal prostate tissues. Both EPCAT2F176 as well as EPCAT4R966 showed no expression in normal prostate tissue obtained via radical cystoprostatectomy. However, normal cells (1) adjacent to prostate cancer (2) were found positive for both EPCATs.

Knock-down of EPCATs impedes growth of prostate cancer cells

To investigate their functional impact on PCa growth, we performed siRNA-directed knockdown of 9 PCR-validated EPCATs (EPCAT1F273, EPCAT2F176, EPCAT2R709, EPCAT3R522, EPCAT4R966, EPCAT5R633, EPCAT8R190, EPCAT15F850, EPCATXR234) in LNCaP and 22RV1 cells. Cell viability was assessed by MTT-assay, and transfections with two scrambled RNAs were used to evaluate unspecific treatment effects of siRNA transfection. We observed significant reductions in cell viability for 6 of these 9 EPCATs (EPCAT1F273, EPCAT3R522, EPCAT4R966, EPCAT8R190, EPCAT15F850, EPCATXR234), 5 of which were showing consistent effects in both LNCaP and 22RV1 (see Figure 7).

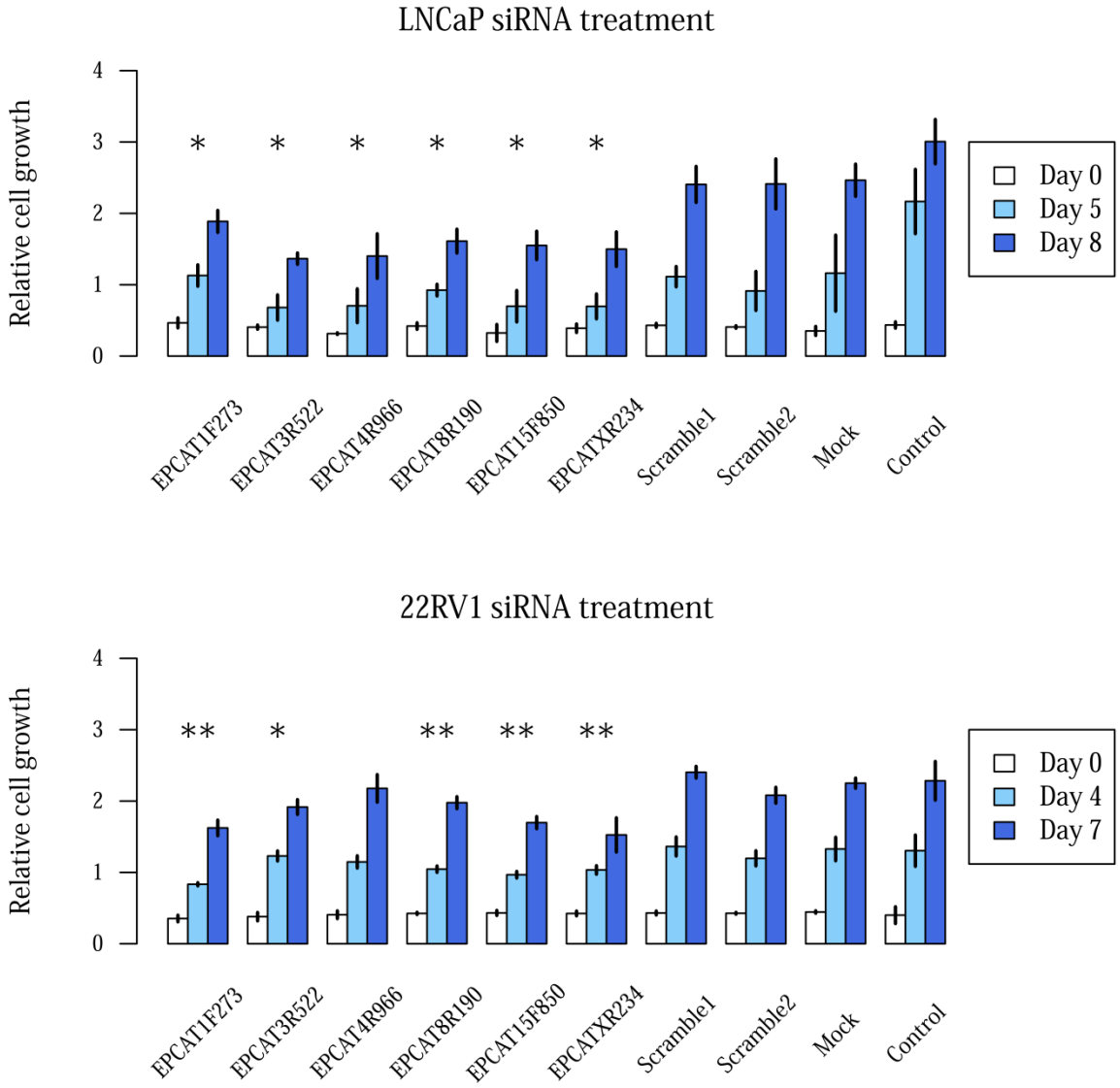


Figure 7: Cell viability measured by MTT assay after treatment of LNCaP and 22RV1 cells. All measurements were performed in triplicates and a t-test was used to determine significant differences ($p < 0.05$) between treatment and scrambled control RNA. * denotes a significant difference at day 7 / 8, ** at both day 5 and 8 / day 4 and 7 for LNCaP and 22RV1, respectively. Experiment were performed twice and representative results are displayed.

Discussion

We successfully set out to identify novel transcripts with PCa-specific expression profiles using unannotated transcript clusters of Affymetrix Human Exon Arrays. The large number of transcript candidates identified shows that we do not yet have a full overview of all the transcribed genomic regions. With efforts such as ENCODE and GENCODE, it has become clear that the number of protein coding genes is reaching a plateau of about 21,000 [30]. On the contrary, the number of non-coding transcripts is increasing rapidly, as particularly deep RNA-sequencing of many normal and diseased tissues reveals a wealth of novel small and long transcripts. Our 334 EPCATs add to this pool of newly identified RNAs. 10 EPCATs were also identified by Prensner *et al*., while 196 EPCATs, of which 9 validated transcripts overlapped with the 32,183 human transcripts present in the LNCipedia [23] database.

In previous studies, several lncRNAs have been associated with PCa development and progression, emphasizing their role as potential markers and therapy targets in cancers [17]. Various mechanisms of lncRNA dependent activation and repression of expression have been reported in PCa, among them are post-transcriptional regulation of BRCA2 by PCAT-1 [31], post-translational regulation of SNF5 protein by SChLAP1 binding [12] as well as mediation of enhancer-promoter looping by interaction with AR (PCGEM1 and PRNCR1, [19]), which is currently disputed and requires further research for clarification [20]. Other described mechanisms include regulation of alternative splicing by MALAT1 and silencing of antisense genes by CDKN2B-AS1 / ANRIL [32]. Furthermore, PCAT29 (EPCAT15F849) has been recently suggested as tumor suppressor in PCa, although its mechanism of action is still unclear [13].

Despite these promising findings, the value of the newly identified lncRNAs in PCa prognostic profiles has not yet been established. To address the need for novel prognostic markers, we investigated whether EPCAT expression on three Affymetrix Exon Array cohorts is related to poor prognostic outcome and found that at least two transcripts (EPCAT2F176 / SChLAP1 and EPCAT2R709) are associated with development of metastasis and PCa-related death. EPCAT2R709 is located approximately 100 kb upstream in antisense direction to EPCAT2F176, making the genomic region on chromosome 2q31.3 a highly interesting target for further studies. Using the RNAscope ISH technology, we independently validated the diagnostic accuracy and power to predict pathological stage of EPCAT2F176 and EPCAT4R966. The association of EPCAT2F176 with development of metastasis and PCa-related death was not confirmed using the TMA, which could be due to differences in sample cohorts and detection technologies. Nevertheless, we did observe a significant increase in number of positive LNPCa samples compared to localized PCa for EPCAT2F176, which could indicate an involvement in formation of metastasis and supports our earlier results.

Both EPCAT2F176 and EPCAT4R966 were found expressed in some PIN lesions by ISH, suggesting that their expression might be an early event in PCa development. Moreover, both transcripts were expressed in approximately 10% of NAP tissue samples, whereas normal prostate controls were completely negative, suggesting that normal adjacent tissue might

differ from truly normal tissue as previously reported [33–35]. Therefore, lncRNA biomarkers such as our EPCATs enable a morphology-independent, molecular-based identification of potentially malignant prostate tissue. Taken together, these findings highlight the high specificity of EPCAT expression and pose questions as to how these lncRNAs are regulated and why they are expressed in subsets of patients only.

We chose three transcription factors with known involvement in PCa to investigate EPCAT regulation, namely AR, ERG and ETV1. Using public Affymetrix Exon Array [36] and ChIP-seq data [25] we found evidence for 4 ERG and 15 AR regulated EPCATs, of which 3 had been PCR-validated. Since the majority of EPCATs does not appear to be AR or ERG regulated, other regulatory mechanisms such as DNA methylation, chromatin restructuring or combinations of transcription factors could play a role. Thus, whether an interplay between these factors will explain the outlier PCa-specific expression of EPCATs is a new and challenging field of research.

In addition to their reported diagnostic and prognostic potential, siRNA-directed knockdown in combination with an MTT-assay revealed that 6 EPCATs (EPCAT1F273, EPCAT3R522, EPCAT4R966, EPCAT8R190, EPCAT15F850, EPCATXR234) are involved in PCa cell viability and growth. Like the recently identified PCAT1, SChLAP1 and PCAT29, the expression of some of the novel EPCATs is functionally relevant and therefore, cancer-associated lncRNAs should not entirely be seen as transcriptional noise due to aberrant regulation.

Despite unknown regulation and of most EPCATs, they offer high specificity in discriminating malignant disease from benign prostate tissues. With the exemplary lncRNA PCA3 being used as clinical diagnostic marker in a urine-based test [6], one can envision that a combination of EPCATs can supplement PCA3 and TMPRSS2-ERG based diagnostic panels. If EPCATs are present in urine, such an assay might help to improve specificity of diagnosis of current markers and reduce the number of unnecessary prostate biopsies.

In conclusion, we present evidence for the existence of novel prostate cancer-specific transcripts that demonstrate diagnostic and prognostic value and might serve important roles in tumor development and progression. A subset of EPCATs is Androgen Receptor or ERG regulated, but for most novel transcripts their unique transcriptional regulation in cancer is still not fully resolved and poses a new challenging research question.

Methods

Public Exon array datasets

We used three independent publicly available datasets of Affymetrix Human Exon Arrays to discover novel prostate cancer-associated transcripts; referred to as 'Taylor' (GSE21034, [37]) and 'Brase' (GSE29079, [38]) and 'EMC'. 'EMC' contains 48 previously published prostate cancer samples (GSE41408, [27]) as well as additional cancerous and control samples, accessible via GEO accession number GSE59745. The datasets comprised samples from normal adjacent prostate (NAP), localized prostate cancer obtained via radical prostatectomy (PCa) and transurethral resection of the prostate (TURP, EMC only), as well as metastasis in lymph node (LNPCa, EMC and Taylor) and other tissues (MetPCa, Taylor only).

Public datasets of other tissues were used for validation of PCa-specific expression and contained samples of lung cancer (GSE12236, [39]), gastric cancer (GSE13195), brain cancer (GSE9385, [40]) as well as breast, colorectal and lung cancer tissue (GSE16534, [41,42]). Androgen regulation of novel transcripts was investigated using a public dataset of LNCaP cells grown in androgen depleted medium or in presence of 10 nM R1881 (GSE32875, [36]).

Patient samples used for gene expression microarray, qRT-PCR and tissue microarray analysis

We used normal and tumor samples of patients from the frozen tissue bank of the Erasmus Medical Center (Rotterdam, the Netherlands, obtained between 1984 and 2001). Further information concerning these patient samples were previously published [43,44]. Experimental protocols were approved by the Erasmus MC Medical Ethics Committee following the Medical Research Involving Human Subjects Act.

For usage on Exon Arrays, 12 NAP and 8 PCa samples were obtained via radical prostatectomies (RP) and histologically evaluated by an uropathologist after haematoxylin/eosin staining of tissue sections. 10 cancer samples obtained by TURP and 12 LNPCa samples obtained via lymphadenectomy were also added to the cohort.

For quantitative real-time RT-PCR, an additional 40 PCa, 43 TURP, 1 LNPCa and 5 NAP samples were chosen along with 3 PCa-negative TURP and 2 lymph node samples that served as controls (see Supplementary Table 2).

Hybridization of exon arrays for clinical samples from normal adjacent prostate

RNA isolation from snap-frozen PCa and NAP samples was performed using RNAbee (Campro Scientific, Berlin, Germany). GeneChip Human Exon 1.0 ST arrays (Affymetrix, Santa Clara, CA, USA) were used to determine expression profiles of each sample. Experiments were performed at the Center for Biomics, Erasmus MC, Rotterdam, the Netherlands and at ServiceXS, Leiden, the Netherlands, according to the manufacturer's instructions [27].

Discovery of novel prostate cancer-associated transcripts

All datasets were normalized via RMA as implemented in the aroma.affymetrix Bioconductor R-package ([45]; CDF used: HuEx-1_0-st-v2,fullR3,A20071112,EP.CDF, see http://www.aroma-project.org/) and summarized transcript cluster (TC) expression values were obtained for the "full" evidence level. An adapted COPA [22] was performed on log2 expression values and a threshold of $\frac{2 \cdot MAD\ (transcript\ cluster\ z-scores)}{0.6745}$ was used to detect outlier samples (as suggested by [46,47]). TCs with known gene assignment based on Affymetrix NetAffx annotation (NA32, based on hg19), outliers in normal tissue samples or less than 5% outliers in cancer samples were removed. All remaining TCs were grouped based on proximity (less than 250 kb apart), same strand and similarity in outlier profile (Spearman's $\rho \geq 0.5$), after which the combined TCs are referred to as EPCATs (see Figure 1). EPCATs that were detected in only one dataset or that comprised less than 12 physical probes on the array were removed. In case EPCATs differed between datasets, all involved TCs were merged into a single EPCAT in order to maximize size and complete the transcript.

Independent validation via RNA-seq data

Independent validation was performed using RNA-seq data of 27 organ-confined PCa samples from 18 patients obtained via laser capture micro dissection and 5 LNPCa samples. RNA-sequencing was performed on a Genome Analyzer II platform using TruSeq adapters (Illumina, San Diego, CA, USA) at Aros Applied Biosciences (Aarhus, Denmark). Sequencing reads were aligned to a pre-indexed hg19 human reference genome using TopHat 2.0.4 [48]. Resulting BAM files were pooled based on tissue type (PCa and LNPCa) to increase resolution for less abundant transcripts and genomic regions covered by EPCATs including 10 kb flanks were extracted. Cufflinks 2.0.2 was executed in reference guided fashion [26,49] and results were curated manually using IGV [50], linking single exons into transcripts and further adding candidates that were missed by Cufflinks. Curated exon-intron boundaries were used to design junction spanning PCR primers.

cDNA synthesis and RT- PCR analysis

RNA-Bee reagent (Campro Scientific, Veenendaal, The Netherlands) was used for total RNA isolation according to manufacturer's protocol. RNA quality was checked on 1% agarose gel and cDNA was synthesized using MMLV-reverse transcriptase kit, according to manufacturer's instructions. EPCAT expression was validated in 6 cell lines (VCaP, 22RV1, LNCaP, PC3, PC346c, DU145 [51–56]) using RT-PCR. Custom PCR primers and TaqMan probes were designed using Primer 3 [57]. Primers were ordered by Sigma Aldrich (St. Louis, MO, USA), probes were ordered at IBA-Lifesciences (Göttingen, Germany, see Supplementary Tables 9 – 10). ABsolute QPCR ROX Mix from Thermo Scientific (Waltham, MA, USA) was used to perform TaqMan real-time PCR analysis on a 7500 Fast Real-Time PCR System from Applied Biosystems (Foster City, CA, USA). Two housekeeping genes, GAPDH (assay ID Hs99999905_m1, Applied Biosystems Foster City, CA, USA) and HMBS were used as endogenous references and a mixture of cDNAs from prostate carcinoma xenografts as calibrator. Quantification of HMBS was performed using 0.33 µM of primer

solution (forward: 5’ CATGTCTGGTAACGGCAATG 3’ and reverse: 5’ GTACGAGGCTTTCAATGTTG 3’) in Power SybrGreen PCR Master Mix (Applied Biosystems), according to thermocycling protocol recommended by the manufacturer. Transcript quantities for each sample were normalized against the average of two endogenous references and relative to a calibrator.

Determining full length sequences of novel transcripts

RT-PCR validated exons were Sanger sequenced using ABI Prism BigDye Terminator v3.1 Ready Reaction Cycle Sequencing Kit. After PCR processing, samples were analyzed using ABI Prism 3100 Genetic Analyzer (Applied Biosystems, Foster City, California, United States).

To identify the 5' and 3' ends of PCR-validated EPCATs, a nested primer approach was used on a λgt11 full length cDNA library of the LNCaP prostate cancer cell line. The λgt11 outer primers were: 5’ TTCAACATCAGCCGCTACA 3’ (forward) and 5’ AAATCCATTGTACTGCCGGA 3’ (reverse). The λgt11 inner primers were: 5’ ACTGATGGAAACCAGCCATC 3’ (forward) and 5’ CCGTATTTCGCTAAGGAAA 3’ (reverse). For amplification of the 5' end of an EPCAT, 0.15 µl of outer forward λgt11 primer and 0.15 µl outer reverse EPCAT primer were used. For amplification of the 3' end of an EPCAT, 0.15 µl of the outer reverse λgt11 primer and 0.15 µl outer forward EPCAT primer were used. The first reaction template was a 1:10 diluted λgt11 cDNA library preheated to 95°C for 5 minutes. For the second reaction, all quantities were doubled and inner primers as well as 1 µl of PCR product from first reaction were used. PCR products were loaded on 1% agarose gel in 1x TBE and the specific band was extracted using GeneJETGel extraction kit (Thermo Fisher Scientific Inc, Waltham, Massachusetts) following manufacturer’s instructions. Specific products were directly used for sequencing and product concentration was determined using a Nanodrop Spectrophotometer ND-1000 (Thermo Fisher Scientific Inc, Waltham, Massachusetts). Sequencing reaction was the same as for RT-PCR products.

Investigation of transcriptional regulation of EPCATs

Androgen regulation of EPCATs was investigated via a public dataset comprising LNCaP cells grown in androgen depleted medium (DCC) or in 10 nM R1881 supplemented medium (GSE32875, [36]). Averaged log2 transformed expression values of all TCs for each EPCAT were used for all analyses. Welch’s t-test was used for comparison of both conditions and p-values were corrected using Benjamini & Hochberg [58]. ERG and ETV1 regulation was evaluated using Spearman’s correlation coefficient. AR and ERG binding in EPCAT regions was further investigated using public ChIP-seq data [25]. Peaks called by Yu *et al*. were converted to hg19 using liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and overlapped with previously identified candidate EPCATs via bedtools [59] including 50 kb flanks. Potential regulation was assumed if at least one peak was falling into the candidate region. For coexpression analysis of genes overlapping EPCATs on the same strand, genes from the UCSC known genes table were intersected with EPCAT regions using bedtools. HGNC symbols for overlapping genes were obtained via biomaRt [60] and median expression

values of associated TCs were correlated with EPCAT expression (Spearman's correlation coefficient).

Computational evaluation of coding potential

Evaluation of coding potential was performed for hg19 build sequences using iSeeRNA (1.2.1) [28] and PhyloCSF (downloaded 22.11.2013) [29]. For iSeeRNA, all RT-PCR validated exon locations were supplied in BED12 format and known coding genes retrieved from the UCSC RefSeq table served as positive controls. For PhyloCSF, a FASTA file containing multiple species alignments for each EPCAT was obtained via the Galaxy 'Stitch Gene blocks' tool (http://usegalaxy.org/). Alignments were based on a 46 way Multiz alignment of hg19. All genome builds were converted to common names and intersected with a panel of 29 mammals offered by PhyloCSF. After splitting the FASTA file by gene, PhyloCSF was run using options --frames=3 –aa for each gene. Two known coding genes, GAPDH and ERG, served as controls.

Computational evaluation of conservation

For each EPCAT's exons, we downloaded base-wise conservation scores (PhyloP) based on Multiz alignments of 100 vertebrates from the UCSC Genome Browser (http://genome.ucsc.edu). Per EPCAT, PhyloP basewise scores were averaged in 50 bp windows and the highest of these averages was used as overall representative score of the gene locus. 1000 randomly selected coding RefSeq genes as well as 1000 randomly selected Repetitive elements (RepeatMasker, UCSC Genome Browser) served as controls.

Tissue microarray construction

A total of four tissue microarrays (TMAs) was used to evaluate expression of two EPCATs (EPCAT4R966 and EPCAT2F176) in patient tissues, xenografts and cell lines (see Supplementary Tables 5a – 5b).

The first TMA consisted of 481 patient samples from radical prostatectomies for PCa and several control specimens as described previously [61]. Controls comprised normal prostate tissues from radical cystoprostatectomies (RCP, n = 7), urothelial cell carcinomas (n = 5), invasive ductal mammary adenocarcinomas (n = 5), palliative transurethral resection of the prostate (TURP, n = 10), prostate cancer lymph node metastasis (LNPCa, n = 10) and placenta (n = 1). Additionally, PCa cell lines (n = 7) and prostate cancer xenografts models (n = 22) were included.

The second TMA, comprised 127 triplicate patient samples of nonneoplastic prostate tissue. We performed a search in PALGA (Pathologisch anatomisch landelijk geautomatiseerd archief, Houten, the Netherlands) and selected 53 patients who had undergone RCP or pelvic exenteration (PE), due to bladder cancer. TURP samples from 74 patients with clinical BPH were included in the TMA as well. All operations had taken place between 2003 and 2013. In RCP and PE specimen, we selected prostate glands from the peripheral zone, whereas transition zone was selected in TURP samples. All slides were histopathologically reviewed to

exclude presence of prostate adenocarcinoma. Several tissues were added to the TMA as landmarks: placenta (n = 1), kidney (n = 1), ovary (n = 1) and spleen (n = 1).

The third TMA contained 119 LNPCa samples from patients who underwent RP combined with a lymph node exploration, obtained between 1989 and 2006 at the Erasmus MC.

The fourth TMA comprised a total of 120 PCa samples, operated between 1982 and 2009 in the Erasmus MC. 35 samples were obtained after RP and 85 samples contained TURP material. 65 of 120 patients were hormone refractory prostate cancers (CRPC), 55 patients were hormone sensitive. After patient selection, all TMAs were constructed using an automated TMA constructor (ATA-27 Beecher Instruments, Sun Prairie, WI, USA) available at the Department of Pathology, Erasmus MC.

In situ hybridisation and quantification - RNAscope

RNA *in situ* hybridisation on FFPE tissue was performed with RNAscope (Advanced Cell Diagnostics, Inc, Hayward, California). One week old 5 μm sections were dewaxed and treated with heat and protease antigen retrieval according to manufacturer's protocol. Specific target probes for EPCAT2F176 (targeting 466 nt) and EPCAT4R966 (targeting 1152 nt) provided by Advanced Cell Diagnostics were hybridized on the tissue (see Supplementary Table 8 for EPCAT sequences). Signal amplification on the probe was followed by visualisation with fast-red and counterstaining with haematoxylin. Probes for housekeeping gene ubiquitin C and bacterial gene dapB served as positive and negative controls. Scoring of TMAs was performed in-house by a trained uropathologist. Only counts above 0 were considered as positive.

Assessment of diagnostic potential

Diagnostic potential was assessed by creating a receiver operator characteristic for 11 EPCATs for which working TaqMan probes were available. Samples that were present in the EMC Exon Array dataset were used as discovery cohort, while the remaining 47 samples (40 PCa, 5 NAP) were used for validation. The R package 'optAUC' was used for AUC maximization in the test cohort and ROC-curves were created using the 'ROC'-package.

Kaplan-Meier survival analysis and evaluation of prognostic potential

Samples of localized PCa from the 'EMC' dataset were used to determine prognostic potential of the 15 validated EPCATs. For each EPCAT, TC intensity values were averaged and used as representative measures of gene expression. Partition Around Medoids (PAM, R-package 'cluster') was used to define two groups of samples with high and low expression of an EPCAT. Overrepresentation of three clinical endpoints was evaluated for 54 patients with available clinical information using a bootstrapping approach. The clinical endpoints were: (i) biochemical recurrence, defined as a rise in serum PSA level from undetectable to ≥0.2 ng/ml in at least two consecutive measurements (at least three months apart) after RP; (ii) clinical progression, defined by occurrence of metastasis in lymph nodes or other organs (iii) prostate cancer related death. For bootstrapping, class labels (clinical endpoints of patients) were

permuted, sampled and assigned to two groups with PAM defined sizes. Sampling was repeated 10,000 times for each EPCAT to create a sample distribution and p-values were calculated as the number of samplings having more positive associations with a clinical endpoint than the original EPCAT entry, divided by the number of iterations. In addition, Kaplan-Meier curves (R package 'survival') were created for each EPCAT and clinical endpoint.

siRNA knockdown and cell viability

Silencer Select siRNA probes were designed by and purchased from Ambion (Life Technologies, Carlsbad, CA, USA). SiRNA probes consisted of a sense and an antisense siRNA for each target transcript with the following sequences:

EPCAT1F273: GGGAAGCAUUGAAAUAGUAtt (sense siRNA), UACUAUUUCAAUGCUUCCCag (antisense siRNA); EPCAT3R522: CAGCUAAGCUGAAAAAGCAtt (sense siRNA), UGCUUUUUCAGCUUAGCUGtc (antisense siRNA); EPCAT4R966: GGCUUGUCGUGUGAUCUAAtt (sense siRNA), UUAGAUCACACGACAAGCCta (antisense siRNA); EPCAT8R190: CCAUGUCCUUGAGAUAAAAtt (sense siRNA), UUUUAUCUCAAGGACAUGGga (antisense siRNA); EPCAT15F850: GAAUGAGAGUCAUCAUGUAtt (sense siRNA), UACAUGAUGACUCUCAUUCag (antisense siRNA); EPCATXR234: CCUUAACAAUGGAUCUGCAtt (sense siRNA), UGCAGAUCCAUUGUUAAGGtt (antisense). PCa cells LNCaP ($12*10^3$ cells) and 22RV1 ($8*10^3$ cells) were transferred to 96 wells plates and kept in RPMI 1640 and 5% FCS. After one day, cells were transfected in triplicate with 500 nM siRNA using DharmaFECT 3 Transfection Reagent (GE Healthcare, Little Chalfont, UK) according to the manufacturers' instructions (20 µl siRNA mix and 80 µl 5% DCC medium per well). 100 µl 5% FCS medium was added to all wells not measured at day 0. Proliferation was subsequently measured using 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) at indicated time points (LNCaP: 0, 5, 8 days; 22RV1: 0, 4, 7 days). All experiments were performed twice.

Acknowledgements

We would like to thank Martijn van der Schoor for the preliminary work to find the EPCATs using Affymetrix Human Exon Arrays and Peter Beyerlein for support of this project. We would also like to thank Theo van der Kwast for his contribution of histopathological evaluation of PCa samples that were used to extract RNA for RT-PCR evaluation.

References

1. Malvezzi M, Bertuccio P, Levi F, La Vecchia C, Negri E. European cancer mortality predictions for the year 2014. Ann Oncol. 2014; 25(8):1650–1656.

2. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. N Engl J Med. 1987; 317(15):909–916.

3. Roobol MJ, Carlsson S V. Risk stratification in prostate cancer screening. Nat Rev Urol. 2013; 10(1):38–48.

4. Srikantan V, Zou Z, Petrovics G, Xu L, Augustus M, Davis L, Livezey JR, Connell T, Sesterhenn IA, Yoshino K, Buzard GS, Mostofi FK, McLeod DG, Moul JW, Srivastava S. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proc Natl Acad Sci U S A. 2000; 97(22):12216–12221.

5. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res. 1999; 59(23):5975–5979.

6. Dijkstra S, Mulders PFA, Schalken JA. Clinical use of novel urine and blood based prostate cancer biomarkers: A review. Clin Biochem. 2014; 47(10-11):889–896.

7. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL. A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet. 2004; 36(3):299–303.

8. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5(7):621–628.

9. Chung S, Nakagawa H, Uemura M, Piao L, Ashikawa K, Hosono N, Takata R, Akamatsu S, Kawaguchi T, Morizono T, Tsunoda T, Daigo Y, Matsuda K, Kamatani N, Nakamura Y, Kubo M. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci. 2011; 102(1):245–252.

10. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011; 29(8):742–749.

11. Du Z, Fei T, Verhaak RGW, Su Z, Zhang Y, Brown M, Chen Y, Liu XS. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol. 2013; 20(7):908–913.

12. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche TJ, Malik R, Bedenis R, McGregor N, Ma T, Chen W, Han S, Jing X, Cao X, Wang X, Chandler B, Yan W, Siddiqui J, Kunju LP, Dhanasekaran SM, Pienta

KJ, Feng FY, Chinnaiyan AM. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013; 45(11):1392–1398.

13. Malik R, Patel L, Prensner JR, Shi Y, Iyer M, Subramaniyan S, Carley A, Niknafs YS, Sahu A, Han S, Ma T, Liu M, Asangani I, Jing X, Cao X, Dhanasekaran SM, Robinson D, Feng FY, Chinnaiyan AM. The lncRNA PCAT29 Inhibits Oncogenic Phenotypes in Prostate Cancer. Mol Cancer Res. 2014;

14. Crea F, Watahiki A, Quagliata L, Xue H, Pikor L, Parolia A, Wang Y, Lin D, Lam WL, Farrar WL, Isogai T, Morant R, Castori-Eppenberger S, Chi KN, Wang Y, Helgason CD. Identification of a long non-coding RNA as a novel biomarker and potential therapeutic target for metastatic prostate cancer. Oncotarget. 2014; 5(3):764–774.

15. Lee JT. Epigenetic regulation by long noncoding RNAs. Science. 2012; 338(6113):1435–1439.

16. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013/03/07 ed. 2013; 20(3):300–307.

17. Martens-Uzunova ES, Böttcher R, Croce CM, Jenster G, Visakorpi T, Calin GA. Long Noncoding RNA in Prostate, Bladder, and Kidney Cancer. Eur Urol. 2014; 65(6):1140–1151.

18. Gesualdo F Di, Capaccioli S, Lulli M. A pathophysiological view of the long non-coding RNA world. Oncotarget. 2014.

19. Yang L, Lin C, Jin C, Yang JC, Tanasa B, Li W, Merkurjev D, Ohgi KA, Zhang J, Evans CP, Rosenfeld MG. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature. 2013; 500(7464):598–602.

20. Prensner JR, Sahu A, Iyer MK, Malik R, Asangani IA, Poliakov A, Vergara IA, Jenkins RB, Davicioni E, Feng FY, Arul M. The lncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer. Oncotarget. 2014; 5(6):1434–1438.

21. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012/09/08 ed. 2012; 22(9):1775–1789.

22. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (80-). 2005; 310(5748):644–648.

23. Volders P-J, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdagh P. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res. 2013; 41(Database issue):D246–51.

24. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT,

Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011; 29(8):742–749.

25. Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, Gong Y, Cheng H, Laxman B, Vellaichamy A, Shankar S, Li Y, Dhanasekaran SM, Morey R, Barrette T, Lonigro RJ, Tomlins SA, Varambally S, Qin ZS, Chinnaiyan AM. An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression. Cancer Cell. 2010; 17(5):443–454.

26. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010/05/04 ed. 2010; 28(5):511–515.

27. Boormans JL, Korsten H, Ziel-van der Made AJC, van Leenders GJLH, de Vos C V, Jenster G, Trapman J. Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. Int J Cancer. 2013; 133(2):335–345.

28. Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics. 2013; 14 Suppl 2(Suppl 2):S7.

29. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011; 27(13):i275–i282.

30. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012/09/08 ed. 2012; 22(9):1760–1774.

31. Prensner JR, Chen W, Iyer MK, Cao Q, Ma T, Han S, Sahu A, Malik R, Wilder-Romans K, Navone N, Logothetis CJ, Araujo JC, Pisters LL, Tewari AK, Canman CE, Knudsen KE, Kitabayashi N, Rubin MA, Demichelis F, Lawrence TS, Chinnaiyan AM, Feng FY. PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. Cancer Res. 2014; 74(6):1651–1660.

32. Walsh AL, Tuzova A V, Bolton EM, Lynch TH, Perry AS. Long noncoding RNAs and prostate carcinogenesis: the missing “linc”? Trends Mol Med. 2014;

33. Haaland CM, Heaphy CM, Butler KS, Fischer EG, Griffith JK, Bisoffi M. Differential gene expression in tumor adjacent histologically normal prostatic tissue indicates field cancerization. Int J Oncol. 2009; 35(3):537–546.

34. Chandran UR, Dhir R, Ma C, Michalopoulos G, Becich M, Gilbertson J. Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. BMC Cancer. 2005; 5(1):45.

35. Braakhuis BJM, Leemans CR, Brakenhoff RH. Using tissue adjacent to carcinoma as a normal control: an obvious but questionable practice. J Pathol. 2004; 203(2):620–621.

36. Rajan P, Dalgliesh C, Carling PJ, Buist T, Zhang C, Grellscheid SN, Armstrong K, Stockley J, Simillion C, Gaughan L, Kalna G, Zhang MQ, Robson CN, Leung HY, Elliott DJ. Identification of novel androgen-regulated pathways and mRNA isoforms through genome-wide exon-specific profiling of the LNCaP transcriptome. PLoS One. 2011; 6(12):e29088.

37. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18(1):11–22.

38. Brase JC, Johannes M, Mannsperger H, Fälth M, Metzger J, Kacprzyk LA, Andrasiuk T, Gade S, Meister M, Sirma H, Sauter G, Simon R, Schlomm T, Beissbarth T, Korf U, Kuner R, Sültmann H. TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. BMC Cancer. 2011; 11:507.

39. Xi L, Feber A, Gupta V, Wu M, Bergemann AD, Landreneau RJ, Litle VR, Pennathur A, Luketich JD, Godfrey TE. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. Nucleic Acids Res. 2008; 36(20):6535–6547.

40. French PJ, Peeters J, Horsman S, Duijm E, Siccama I, Van Den Bent MJ, Luider TM, Kros JM, Van Der Spek P, Sillevis Smitt PA. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. Cancer Res. 2007; 67(12):5635–5642.

41. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VEH, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, Seshagiri S. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature. 2010; 466(7308):869–873.

42. Lin E, Li L, Guan Y, Soriano R, Rivers CS, Mohan S, Pandita A, Tang J, Modrusan Z. Exon Array Profiling Detects EML4-ALK Fusion in Breast, Colorectal, and Non-Small Cell Lung Cancers. Mol Cancer Res. 2009; 7(9):1466–1476.

43. Van Der Heul-Nieuwenhuijsen L, Hendriksen PJM, Van Der Kwast TH, Jenster G. Gene expression profiling of the human prostate zones. BJU Int. 2006; 98:886–897.

44. Martens-Uzunova ES, Jalava SE, Dits NF, van Leenders GJ, Moller S, Trapman J, Bangma CH, Litman T, Visakorpi T, Jenster G. Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. Oncogene. 2012; 31(8):978–991.

45. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk A V, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics. 2008; 24(15):1707–1714.

46. Daszykowski M, Kaczmarek K, Vanderheyden Y, Walczak B. Robust statistics in data analysis - A review: Basic concepts. Chemom Intell Lab Syst. 2007; 85(2):203–219.

47. Li J-W, Schmieder R, Ward RM, Delenick J, Olivares EC, Mittelman D. SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics. 2012; 28(9):1272–1273.

48. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14(4):R36.

49. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011; 27(17):2325–2329.

50. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011; 29(1):24–26.

51. Korenchuk S, Lehr JE, MClean L, Lee YG, Whitney S, Vessella R, Lin DL, Pienta KJ. VCaP, a cell-based model system of human prostate cancer. In Vivo. 2001; 15(2):163–168.

52. Sramkoski RM, Pretlow TG, Giaconia JM, Pretlow TP, Schwartz S, Sy MS, Marengo SR, Rhim JS, Zhang D, Jacobberger JW. A new human prostate carcinoma cell line, 22Rv1. In Vitro Cell Dev Biol Anim. 1999; 35(7):403–409.

53. Horoszewicz JS, Leong SS, Chu TM, Wajsman ZL, Friedman M, Papsidero L, Kim U, Chai LS, Kakati S, Arya SK, Sandberg AA. The LNCaP cell line--a new model for studies on human prostatic carcinoma. Prog Clin Biol Res. 1980; 37:115–132.

54. Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). Invest Urol. 1979; 17(1):16–23.

55. Marques RB, Erkens-Schulze S, de Ridder CM, Hermans KG, Waltering K, Visakorpi T, Trapman J, Romijn JC, van Weerden WM, Jenster G. Androgen receptor modifications in prostate cancer cells upon long-termandrogen ablation and antiandrogen treatment. Int J Cancer. 2005; 117(2):221–229.

56. Stone KR, Mickey DD, Wunderli H, Mickey GH, Paulson DF. Isolation of a human prostate carcinoma cell line (DU 145). Int J Cancer. 1978; 21(3):274–281.

57. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000; 132:365–386.

58. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995; 57(1):289–300.

59. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842.

60. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005; 21(16):3439–3440.

61. Hoogland AM, Jenster G, van Weerden WM, Trapman J, van der Kwast T, Roobol MJ, Schröder FH, Wildhagen MF, van Leenders GJ. ERG immunohistochemistry is not predictive for PSA recurrence, local recurrence or overall survival after radical prostatectomy for prostate cancer. Mod Pathol. 2012; 25(3):471–479.

Chapter 4

Human phosphodiesterase 4D7 (PDE4D7) expression is increased in TMPRSS2-ERG positive primary prostate cancer and independently adds to a reduced risk of post-surgical disease progression

René Böttcher[1,7], David JP Henderson[2,7,8], Kalyan Dulla[3], Dianne van Strijp[3], Leonie F Waanders[3], Gregor Tevz[3,4], M L Lehman[4], Dennis Merkle[3], Geert JLH van Leenders[5], George S Baillie[2], Guido Jenster[1], Miles D Houslay[6] and Ralf Hoffmann[2,3]

1. Department of Urology, Erasmus Medical Center, Rotterdam 3000 CA, The Netherlands;
2. Institute of Cardiovascular and Medical Science, University of Glasgow, Glasgow G12 8TA, Scotland;
3. Departments of Oncology Solutions and Precision Diagnostics, Philips Research Europe, Eindhoven 5656 AE, The Netherlands;
4. Australian Prostate Cancer Research Centre—Institute of Health and Biomedical Innovation, University of Technology, and Translational Research Institute, Brisbane, Queensland 4102, Australia;
5. Department of Pathology, Erasmus Medical Center, Rotterdam 3000 CA, The Netherlands
6. Institute of Pharmaceutical Science, King's College London, London WC2R 2LS, UK
7. These authors contributed equally to this work.
8. Current address: The Salk Institute, Regulatory Biology Lab, La Jolla, CA 92307, USA

Published in

British Journal of Cancer (2015) 113, 1502–1511

Supplementary Material is available via

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815894

Abstract

Background: There is an acute need to uncover biomarkers that reflect the molecular pathologies, underpinning prostate cancer progression and poor patient outcome. We have previously demonstrated that in prostate cancer cell lines PDE4D7 is downregulated in advanced cases of the disease. To investigate further the prognostic power of PDE4D7 expression during prostate cancer progression and assess how downregulation of this PDE isoform may affect disease outcome, we have examined PDE4D7 expression in physiologically relevant primary human samples.

Methods: About 1405 patient samples across 8 publically available qPCR, Affymetrix Exon 1.0 ST arrays and RNA-sequencing data sets were screened for PDE4D7 expression. The TMPRSS2-ERG gene rearrangement status of patient samples was determined by transformation of the exon array and RNA-seq expression data to robust z-scores followed by the application of a threshold 43 to define a positive TMPRSS2-ERG gene fusion event in a tumour sample.

Results: We demonstrate that PDE4D7 expression positively correlates with primary tumour development. We also show a positive association with the highly prostate cancer-specific gene rearrangement between TMPRSS2 and the ETS transcription factor family member ERG. In addition, we find that in primary TMPRSS2-ERG-positive tumours PDE4D7 expression is significantly positively correlated with low-grade disease and a reduced likelihood of progression after primary treatment. Conversely, PDE4D7 transcript levels become significantly decreased in castration resistant prostate cancer (CRPC).

Conclusions: We further characterise and add physiological relevance to PDE4D7 as a novel marker that is associated with the development and progression of prostate tumours. We propose that the assessment of PDE4D7 levels may provide a novel, independent predictor of post-surgical disease progression.

Introduction

Prostate cancer is the most commonly occurring non-skin malignancy in men, with an estimated 900 000 new cases diagnosed world-wide in 2013 (1). However, reactive clinical intervention after routine diagnosis often leads to significant overtreatment of non-aggressive tumours. This has severe negative impacts on both patient quality of life and the medical resources of healthcare institutions (2, 3). Therefore, the characterisation of new biomarkers and methods of clinical assessment is of significant importance when assessing the need for different forms of clinical intervention.

Previous studies have shown that signalling pathways mediated by the second messenger cAMP have various roles in the development and progression of prostate cancer (4). Cyclic nucleotide phosphodiesterases (PDEs) (5, 6) provide the sole means of degrading cAMP and cGMP in cells, and are pivotally placed to regulate cAMP signalling by virtue of their intracellular location and post-translational modification (7, 8). Each of the 11 PDE genes encode for a series of isoform variants, thereby greatly increasing the diversity of unique regulatory mechanisms, intracellular targeting and kinetic properties, which define functionally independent and unique signalling roles within the cell (8–10). This diversity underpins a paradigm of compartmentalised, temporally gated cyclic nucleotide signalling. Due to the complexity of these orchestrated signalling events, any change in PDE isoform expression or regulation can functionally contribute to disease onset (11–15). The molecular characterisation of these changes can be expected to provide means for the development of novel therapeutics and diagnostics (8, 16).

Members of the PDE4D subfamily have been implicated as underpinning the molecular pathology of various diseases including prostate cancer (17, 18), stroke (19), acrodysostosis (14) and COPD (15). The PDE4D gene encodes a cohort of isoforms that are classified as long, short and super-short. Long isoforms possess two conserved regulatory domains, called UCR1 and UCR2, which allow long isoforms to be phosphorylated and activated by PKA (30,50 cAMP-dependent protein kinase) after cAMP elevation in cells (20), as well as being functionally regulated through phosphorylation by activated forms of ERK, MK2 and AMPK (21, 22). PDE4D7 is a long isoform member of this subfamily (23). We have demonstrated that PDE4D7 exhibits a specific pattern of intracellular localisation in prostate cancer cells, where it is functionally targeted to the sub-plasma membrane compartment (18). Spatially constrained PDE4D7 appears to perform a pivotal role in these cells by desensitising sub-plasma membrane-localised cAMP signalling (18), as well as providing a node for crosstalk with signalling pathways that elicit the activation of Erk, MK2 and AMPK (21, 22, 24, 25). PDE4D7 activity is also regulated by PKA phosphorylation within its unique N-terminal region (26). Interestingly, susceptibility markers for ischaemic stroke also map to the region of Chr5q12, where PDE4D7 and the androgen-regulated PART1 exons are located (19). We have previously demonstrated that PDE4D7 is highly expressed in androgen-responsive prostate cancer cell lines and xenografts, while being downregulated in castration resistant samples (18). Indeed, the ectopic overexpression of PDE4D7 in castration resistant prostate cancer (CRPC) cell lines reduced cellular proliferation, while specific knockdown of the PDE

isoform in androgen-sensitive cells lead to an increase in cellular proliferation, indicating a functional role of PDE4D7 downregulation during the progression to CRPC growth. Here, we set out to assess whether the changes in PDE4D7 expression we observed in model systems have clinical relevance. To do this, we analysed 1405 tumour samples sourced from 8 independent patient cohorts that were enrolled at different clinical centres (Supplementary Table 1). Our analyses of clinical samples highlight an increase in PDE4D7 expression during initial tumorigenesis and further support our contention that PDE4D7 levels then fall profoundly in CRPC, suggesting that PDE4D7 transcripts may provide a potentially useful biomarker and therapeutic target.

Methods

Human tissue samples.

Human tissues samples were obtained under local laws and regulation to obtain and handle patient material for research purposes. Sample descriptions are depicted in Figure 1A.

Molecular biology (RNA extraction, cDNA synthesis and primer design).

If not otherwise indicated RNA isolation, cDNA conversion and Real-Time PCR were performed using RNeasy Kit (QIAGEN GmbH, Hilden, Germany, 74004), iScript cDNA synthesis kit (Bio-Rad Inc, Hercules, CA, USA, 170–8890), GeneAmp Fast PCR Master Mix (Applied Biosystems Inc, Foster City, CA, USA, 4362070) respectively, according to the manufacturer's instruction. Real-Time PCR probe and primer sets were developed by targeting isoform-specific intron-spanning regions of genetic code (Supplementary Table 3).

Quantitative RT–PCR (qRT–PCR).

To enable the comparison of qPCR data across different experiments, we normalised the Ct value for PDE4D7 against the mean of the Ct values for the reference genes (Supplementary Table 3) to generate a normalized PDE4D7 expression value. We use the following formula to normalise the raw Ct values:

$$N\left(Ct_{gene\ of\ interest}\right) = Mean\left(Ct_{ref\ gene}\right) - \left(Ct_{gene\ of\ interest}\right)$$

Where $N(Ct_{gene\ of\ interest})$ is normalised gene expression value for a gene of interest; where $Mean(Ct_{ref\ gene})$ is the arithmetic mean of the PCR Cq values of the selected combination of reference genes; where $(Ct_{gene\ of\ interest})$ is the PCR Cq value of the gene of interest. Note: in case DNA microarray or RNA-seq technologies was used to measure PDE4D7 expression, the qPCR Ct value was replaced by a normalised measurement of the respective technology, for example, an robust multi-array average (RMA) normalised gene expression value for DNA microarrays, or a TPM (transcript per million) normalised gene expression value for RNA-sequencing.

Analysis of Affymetrix Human Exon Arrays.

Raw CEL files were downloaded from Gene Expression Omnibus for the publically available data sets (Supplementary Table 1). Data processing and RMA normalisation were performed using the aroma.affymetrix R-package (Affymetrix Inc, Santa Clara, CA, USA;(27)) and transcript isoform expression was measured by averaging log2-transformed intensity values of the following isoform-specific probe sets: PDE4D7 (2858406, 2858407 and 2858408); Note: for data set Erho et al. (2013) (Supplementary Table 1) only probe set 2858408 was used in the analysis as probe sets 2858406 and 2858407 showed relatively limited signal intensities compared with probe set 2858408.

RNA-seq data analysis.

RNA-seq data of 193 prostate cancer clinical samples (36 normal, 157 tumour) was downloaded from The Cancer Genome Atlas (TCGA) Data Portal (4 September 2013) and the expression value of genes and isoforms (TPM-transcript per million) was estimated as previously described (28).

Positive TMPRSS2-ERG fusion status was estimated in general by transformation to robust z-scores. Positive TMPRSS2-ERG fusion status was estimated by transformation to robust z-scores, utilising robust statistical measures, namely median and median absolute deviation, to replace mean and SD, which are sensitive to outliers. Thus, log2-transformed expression values were converted by z-score = (expression-median(expression))/(MAD(expression)), and a threshold of >3 was applied to define samples with positive fusion events. Subsequently, a threshold of >3 was applied to define samples with positive fusion events. For the Erho et al. (2013) data set, we applied a supervised clustering algorithm (Partitioning Around Medoids) to assign prostate cancer samples in one of the two clusters (high ERG or low ERG) based on the log2-transformed expression values of ERG. High ERG expression was subsequently assumed as representative for the presence of a positive TMPRSS2-ERG fusion event.

To assess whether any evidence of ERG binding in the genomic region of PDE4D could be observed, we utilised public ChIP-seq data (GSE14092) from the VCaP prostate cancer cell line after liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to hg19 and found 43 peaks overlapping PDE4D when including 50-kb flanking regions. One of these peaks overlapped the PDE4D7 promoter region, while another was located in close proximity (<200 bases distance), which may hint towards an involvement of ERG binding in regulation of PDE4D7 expression.

Statistical data analysis

For ROC analysis, calculation of AUC under the ROC, ROC P-values and Box-and-Whisker plots the statistical software package MedCalc (MedCalc Software BVBA, Ostend, Belgium) was used. P-values for differences of mean expression were calculated by using Wilcoxon–Mann–Whitney testing unless mentioned otherwise.

Kaplan–Meier Survival curves have been generated by the medical statistical software package MedCalc based on the time to event for those patients who experienced the respective event (e.g., biochemical recurrence (BCR) or clinical recurrence (CR) of disease after surgery) and for those patients who did not suffer from the event at the time of follow-up (censored data). Further, to segregate the analysed patient cohort into two survival groups we determined a cut-off of PDE4D7 expression from a ROC curve analysis. The respective cut-off was objectively determined from the ROC curve at the unique point in the curve, where the sum of sensitivity and specificity reached a maximum.

Results

We have recently provided evidence, suggesting that PDE4D7 may play an important role in regulating cAMP signalling during prostate cancer progression (18). To further explore this finding, we have evaluated the expression of PDE4D7 in a total of eight clinically relevant patient data sets. These data sets comprised a total of 1405 patient samples stratified into 8 sample categories listed in Figure 1A. Three different technology platforms were also leveraged to ensure reproducibility and significance of the gene expression data for PDE4D7, namely: (1) qPCR; (2) Affymetrix Human Exon Array 1.0 ST; (3) RNA-seq (see Supplementary Table 1). More details of the data sets used within this study can be found in Supplementary Tables 1 and 2.

PDE4D7 expression correlates with primary localised prostate tumours and is significantly downregulated in CRPC.

Our previous investigation in cell lines and xenograft material found that PDE4D7 was differentially expressed between androgen sensitive/responsive and CRPC cells (18). To assess if this finding is physiologically relevant, we thought it prudent to examine PDE4D7 transcript expression in primary patient samples. We selected three prostate cancer exon array data sets (Taylor et al., 2010; Boormans et al., 2013; Böttcher et al., 2015; J Schalken, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands, Personal Communication; Supplementary Table 1) and analysed a range of primary prostate cancer samples including tissues collected from patients who developed biochemical or clinical tumour progression after primary treatment, as well as metastases and CRPC (Figure 1B–D; Supplementary Table 5). We observed a striking downregulation in PDE4D7 expression between primary prostate cancer without tumour progression (Primary PCa, NP) and primary prostate cancer tissue with either progression to BCR (Primary PCa, BCR) or CR (Primary PCa, CR). The ROC analysis for the group-wise comparisons revealed AUCs are between 0.61 and 0.82 (Supplementary Table 5). In line with our previous findings, the most significant downregulation was observed between tissues representing primary prostate cancer vs. CRPC (data sets Taylor et al., 2010 and J Schalken, Personal Communication; P-values for differential PDE4D7 expression 5.80E-04, and 1.90E-05, respectively; AUCs for PDE4D7 ROC analysis 0.82, 95% CI 0.73–0.88 and 0.81, 95% CI 0.71–0.90, respectively; Supplementary Table 5). In contrast to the comparison between primary tumours and CRPC, a differential expression of PDE4D7 between primary prostate cancer and metastatic tissue could not be confirmed in the data set from Taylor et al. (2010) (P=1.60E-01; AUC=0.65; 95% CI 0.55–0.74) nor in Boormans et al. (2013); Böttcher et al. (2015) (P=1.10E-01; AUC=0.67; 0.49–0.82); however, in the data set produced by J Schalken, Personal Communication the expression difference was significant (P=4.6E-04) with a very large AUC (0.91; 95% CI 0.81–0.97). Overall this data confirms our original observation made in in vitro models of prostate cancer; PDE4D7 is significantly downregulated in aggressive and advanced forms of prostate cancer.

A

Category	Description
NAT	Normal adjacent tissue
BL	Benign lesions due to infection or inflammation
BPH	Benign prostate hyperplasia
Primary PCa, NP	Primary tumour, no progression
Primary PCa, BCR	Primary tumour, biochemical recurrence
Primary PCa, CR	Primary tumour, clinical recurrence
Metastases	Metastatic tissue (e.g. lymph node)
CRPC	Castration resistant prostate cancer

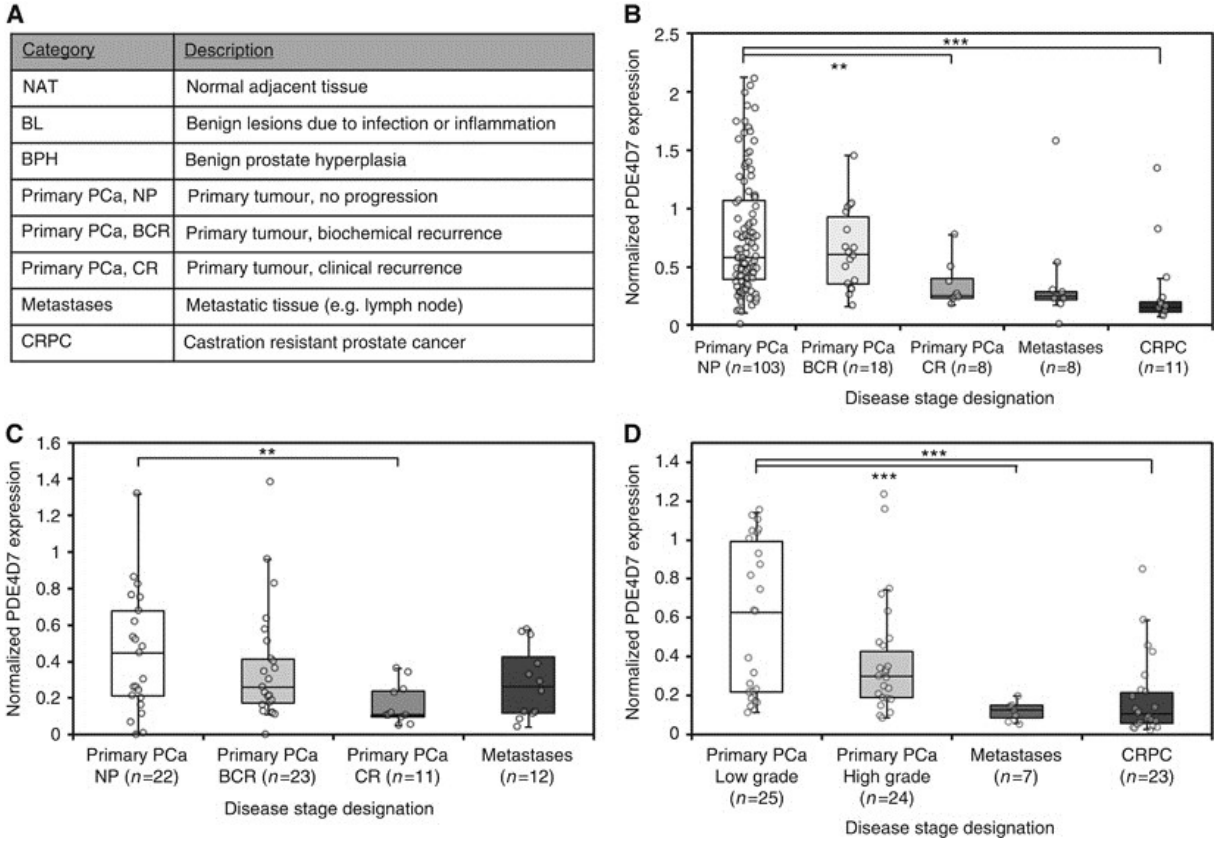


Figure 1: Expression of PDE4D7 splice variant in primary, metastatic and castration resistant cancerous prostate tissues. Box-and-Whisker plots of the normalised PDE4D7 transcript expression in various prostate cancer tissues. For all data sets, and all P-values and AUCs see Supplementary Table 5. (**A**) Disease stage annotation for this study. (**B**) Data: Taylor et al, 2010; P-values of group comparison for difference of mean PDE4D7 expression: (Primary PCa, NP) vs (Primary PCa, BCRandCR), P=7.2E−02; (Primary PCa, NP&BCR) vs (Primary PCa, CR), P=5.90E−03; (Primary PCa, all) vs (Metastases), P=1.60E−01; (Primary PCa, all) vs (CRPC), P=5.8E−04; (**C**) Data: Boormans et al, 2013; Böttcher et al, 2015; P-values of group comparison for difference of mean PDE4D7 expression: (Primary PCa, NP) vs (Primary PCa, BCR&CR), P=6.50E−02; (Primary PCa, NP) vs (Primary PCa, CR), P=1.30E−03; (Primary PCa, all) vs (Metastases), P=1.1E−01; (**D**) Data: J Schalken, Personal Communication; P-values of group comparison for difference of mean PDE4D7 expression: (Primary PCa, low grade) vs (Primary PCa, high grade), P=2.0E−01; (Primary PCa, all grades) vs (Metastases), P=4.60E−04; (Primary PCa, all grades) vs (CRPC), P=1.90E−05. **P<0.01 and ***P<0.001.

PDE4D7 expression is upregulated in localised primary prostate tumours and correlates with TMPRSS2-ERG gene fusion.

To assess the significance of PDE4D7 expression within the context of the normal prostate epithelia, we extended the exon array analysis to include patient tissue taken from areas adjacent to prostate tumours (NAT). We examined 850 patient samples across seven independent data sets (Supplementary Tables 1 and 2). Interestingly, we observed a

significant upregulation of PDE4D7 in primary prostate cancer vs. NAT (Figure 2A–D; Supplementary Table 6). This suggests that PDE4D7 upregulation in prostate tissue may be involved with initial tumorigenesis.

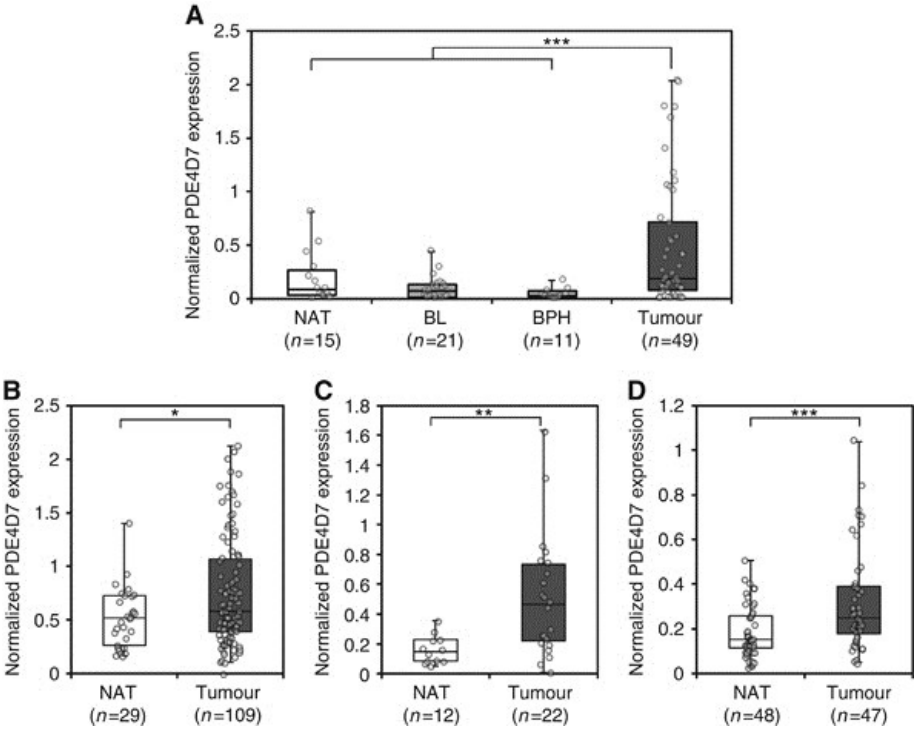


Figure 2: Expression of PDE4D7 splice variant in normal, benign vs cancerous prostate tissues. Box-and-Whisker plots of the normalised PDE4D7 transcript expression in various prostate cancer tissues. For all data sets and all P-values see Supplementary Table 6. (**A**) Data: Origene; P-values of group comparison for difference of mean PDE4D7 expression: (NAT) vs (Primary PCa), P=6.86E−02; (NAT&BL&BPH) vs (Primary PCa), P=1.31E−04; (BL&BPH) vs (Primary PCa), P=4.0E−04; (BPH) vs (Primary PCa), P=3.2E−02; (**B**) Data: Taylor et al, 2010; P-values of group comparison for difference of mean PDE4D7 expression: (NAT) vs (Primary PCa), P=3.30E−02; (**C**) Data: Boormans et al, 2013; Böttcher et al, 2015; P-values of group comparison for difference of mean PDE4D7 expression: (NAT) vs (Primary PCa), P=3.50E−03; (**D**) Data: Brase et al, 2011; P-values of group comparison for difference of mean PDE4D7 expression: (NAT) vs (Primary PCa), P=1.00E−03. *P<0.05, **P<0.01 and ***P<0.001.

To investigate this further, we set out to establish if there was any correlation between PDE4D7 expression and factors known to regulate initial tumorigenesis in the prostate. The TMPRSS2-ERG gene fusion has previously been reported as a clinical indicator for prostate cancer formation. Since its discovery, this prostate cancer-specific fusion event has been described in B50% of prostate cancer patients and has become a molecular hallmark of prostatic tumours (29). Given the status of TMRSS2-ERG as the most relevant genomic fusion event so far identified in prostate cancer, we tested the expression of PDE4D7 in 1106 patients with (Primary PCa, TMPRSS2-ERG positive; Figure 3) and without (Primary PCa,

TMPRSS2-ERG negative; Figure 3) this gene fusion. Figure 3A–C shows the PDE4D7 expression across three exon array data sets (data sets (30–33); Supplementary Table 1), which we picked for graphical illustration (all data sets where we had information on TMRPSS2-ERG rearrangement information available can be found in Supplementary Table 7). Intriguingly, we observed a significantly higher PDE4D7 expression in tumour samples that harboured the TMPRSS2-ERG gene fusion when compared with TMPRSS2-ERG negative samples or when compared with NAT (2-fold median increase, with some samples in excess of 30-fold upregulation; P-values of group-wise comparisons between TMPRSS2-ERG negative vs. positive tumours: 3.33E-08; 8.60E-03; 3.80E-06, respectively). At the same time there was no significant expression difference observed between TMPRSS2- ERG gene fusion negative cancer samples and NAT (Figure 3A–C; Supplementary Table 7).

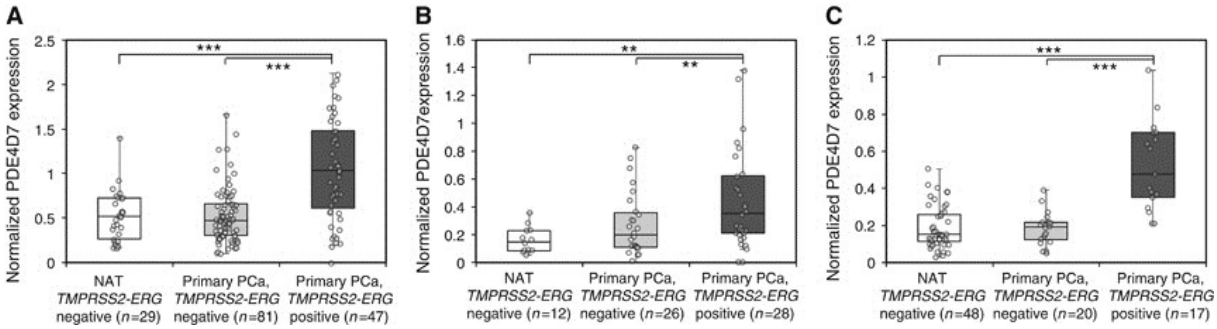


Figure 3: Correlation of PDE4D7 expression in normal and cancerous human prostate tissues to TMPRSS2-ERG gene fusion status. (**A**) Box-and-Whisker plots of the normalised PDE4D7 transcript expression in various prostate cancer tissues. For all data sets, and all P-values see Supplementary Table 7. Positive TMPRSS2-ERG fusion status was estimated by transformation to robust z-scores (Materials and Methods). Subsequently, a threshold of >3 was applied to define samples with positive fusion events. Samples were divided into three different groups: (1) normal adjacent tissue without TMPRSS2-ERG fusion events (NAT TMPRSS2-ERG negative); (2) prostate tumour tissue without TMPRSS2-ERG fusion events (Primary PCa, TMPRSS2-ERG negative), and (3) prostate tumour tissue with TMPRSS2-ERG fusion events (Primary PCa, TMPRSS2-ERG positive). (**A**) Data: Taylor et al, 2010; P-values of group comparison for difference of mean PDE4D7 expression: (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG negative), P=9.00E−01; (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=1.10E−05; (Primary PCa, TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=3.33E−08. (**B**) Data: Boormans et al, 2013; Böttcher et al, 2015; P-values of group comparison for difference of mean PDE4D7 expression: (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG negative), P=5.90E−01; (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=5.60E−03; (Primary PCa, TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=8.60E−03. (**C**) Data: Brase et al, 2011; P-values of group comparison for difference of mean PDE4D7 expression: (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG negative), P=7.80E−01; (NAT TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=5.10E−07; (Primary PCa, TMPRSS2-ERG negative) vs (Primary PCa, TMPRSS2-ERG positive), P=3.80E−06. **P<0.01 and ***P<0.001.

PDE4D7 expression is positively correlated with low-grade TMPRSS2-ERG-positive prostate tumours.

Having discovered a strong correlation between TMPRSS2-ERG fusion and PDE4D7 expression, we then set out to ascertain if cancer aggressiveness is correlated with PDE4D7 expression. We compared the transcript levels of PDE4D7 against pathology-graded cancer samples utilising three exon array data sets (Taylor et al., 2010; Brase et al., 2011; J Schalken, Personal Communication; Supplementary Table 1), as well as the TCGA prostate adenocarcinoma RNA-seq Data Set Prostate Cancer (Release September 2013). We categorised Gleason score (pGleason) into the following four groups of increasing grade: (1) pGleason 3+3, (2) pGleason 3+4, (3) pGleason 4+3, (4) pGleason ≥4+4. A total of 264 patients were included in this stratification, and Supplementary Table 8 provides an overview of various group-wise comparisons of these different pGleason groups. Amazingly, a significant downregulation of PDE4D7 between low grade (pGleason≤ 3-4) vs. high grade (pGleason ≥4+3) tumours was only observed in patients possessing the TMPRSS2-ERG gene fusion (Figure 4A and B; Supplementary Table 8). The initial increase in PDE4D7 expression in low-grade prostate cancer is in keeping with our observations from Figure 3. It is significant that in TMPRSS2- ERG-positive tumour samples the expression of PDE4D7 is negatively correlated with increasing pGleason, highlighting the transient nature of PDE4D7 upregulation. This finding bears a striking resemblance to our previous observations in cell lines and xenografts (18).

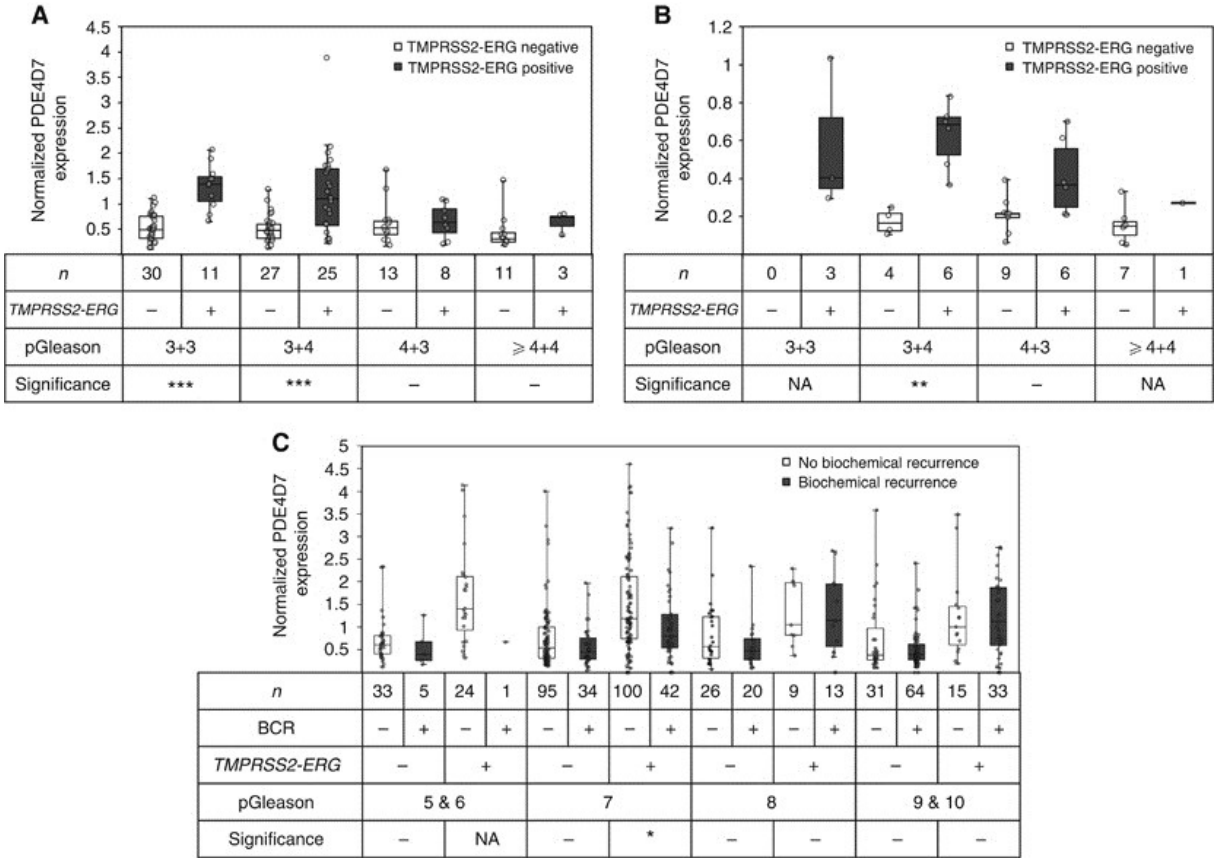


Figure 4: Correlation of PDE4D7 expression to pathology gleason score. (**A** and **B**) Box-and-Whisker plots of the normalised PDE4D7 transcript expression in various prostate cancer

tissues. For all data sets, and all P-values see Supplementary Table 8 (data sets Taylor et al, 2010; Brase et al, 2011) and Supplementary Table 9 (data set Erho et al, 2013). For estimation of positive TMPRSS2-ERG fusion status see Materials and Methods. Patient cohorts were categorised according to their pGleason on histology as indicated. (**A**) Data Taylor et al, 2010; P-values of group comparison for difference of mean PDE4D7 expression: (pGleason 3+3 & 3+4 (TMPRSS2-ERG negative)) vs (pGleason 4+3 & greater than or equal to4+4 (TMPRSS2-ERG negative)), P=4.80E−01; (pGleason 3+4 & 3+4 (TMPRSS2-ERG positive)) vs (pGleason 4+3 & greater than or equal to4+4 (TMPRSS2-ERG positive)), P=2.40E−03; (**B**) Data Brase et al, 2011; P-values of group comparison for difference of mean PDE4D7 expression: (pGleason 3+3 & 3+4 (TMPRSS2-ERG negative)) vs (pGleason 4+3 & greater than or equal to4+4 (TMPRSS2-ERG negative)), P=8.20E−01; (pGleason 3+4 & 3+4 (TMPRSS2-ERG positive)) vs (pGleason 4+3 & greater than or equal to4+4 (TMPRSS2-ERG positive)), P=4.20E−02; (**C**) Data Erho et al, 2013; progression after primary treatment (i.e., surgery) is indicated as BCR (+) or absence (−) of BCR. P-values of group comparison for difference of mean PDE4D7 expression: (pGleason 7 (TMPRSS2-ERG negative), NP) vs (pGleason 7 (TMPRSS2-ERG negative, BCR)), P=1.10E−01; (pGleason 7 (TMPRSS2-ERG positive), NP) vs (pGleason 7 BCR (TMPRSS2-ERG positive), BCR), P=4.60 E−02.

PDE4D7 expression is correlated with clinical outcome in patients expressing the TMPRSS2-ERG gene fusion.

To test our hypothesis that PDE4D7 expression can predict clinical outcome in patients with positive TMPRSS2-ERG gene rearrangement, we used an exon array data sets covering 527 eligible patient samples where longitudinal outcome data was available (30, 32, 34). The data allowed for prediction of BCR after primary treatment. The patients were grouped according to their TMPRSS2-ERG gene fusion status, as well as according to pGleason (5 and 6, 7, 8, and 9 and 10). We then compared the PDE4D7 expression in patient groups with vs. without BCR during 5-years follow-up after primary treatment (Figure 4C and Supplementary Table 9). We could not detect a significant change in the expression of PDE4D7 in any of the TMPRSS2-ERG-negative pGleason groupings. However, for patient significant differential expression in the pGleason 7 group between no progression and BCR during follow-up, while this was not the case for the pGleason scores 47. Unfortunately, there is only a single patient sample in the pGleason 5 and 6 group with positive TMPRSS2-ERG status and progression to BCR so we could not calculate a P-value. However, this particular sample shows a very low PDE4D7 expression value compared with the samples in this pGleason group but without post-treatment progression (Figure 4C). We concluded from this that low PDE4D7 expression values in patient samples with low pGleason scores (6 and 7) are associated with an increased likelihood of biochemical failure after primary intervention.

A graphical representation of PDE4D7 expression in various cell and tissue types including AR negative/AR positive cell lines and xenografts, primary prostate cancer with and without progression to biochemical or CR, metastases and CRPC is shown in Figure 5A (cell lines and xenograft samples) and Figure 5B (patient samples; Supplementary Table 4). The samples are ordered based on their normalised PDE4D7 expression. For the cell lines, xenografts, primary

tumours without progression and primary tumours with progression to BCR or CR, as well as CRPC tumours, the status of the TMPRSS2-ERG rearrangement is indicated. In general, the more aggressive type of samples are represented by low expression levels of PDE4D7, while less aggressive samples show elevated PDE4D7 expression. It is evident from the depicted cell lines and xenografts that the expression level of PDE4D7 is largely influenced by its TMPRSS2-ERG rearrangement status rather than its AR expression status, where AR positive cell lines without gene fusion show low PDE4D7 expression, while cell lines of the same category but positive gene translocation demonstrate high PDE4D7 expression levels (Figure 5A). It is also of importance to note that this effect seems to be very specific to the ERG translocation as cell lines or xenograft samples with ETV1 or ETV4 translocations do not show elevated PDE4D7 transcription (Figure 5A). Also, looking at the samples collected from patients without disease progression during follow-up reveals that those samples that were positively tested for TMPRSS2-ERG in general show increased expression of PDE4D7 (Figure 5B). This was also the case for primary tumour samples where patients progressed to either biochemical or CR as well as for CRPC. We further annotated for patients who experienced a biochemical relapse the time to PSA recurrence as two categories—relapse <24 months vs. relapse >24 months after primary treatment. We observed a clear association between an increased PDE4D7 expression level and an elevated time to recurrence (P=1.72E-02; eight out of nine patients with normalised PDE4D7 expression >0 had a BCR recurrence event >24 months after primary therapy; Figure 5B). Furthermore, we noticed that from eight patients with clinical disease recurrence during follow-up seven patients showed normalised PDE4D7 expression values <0 (Figure 5B) while only in one patient tissue we could measure PDE4D7 expression values 40 (Figure 5B).

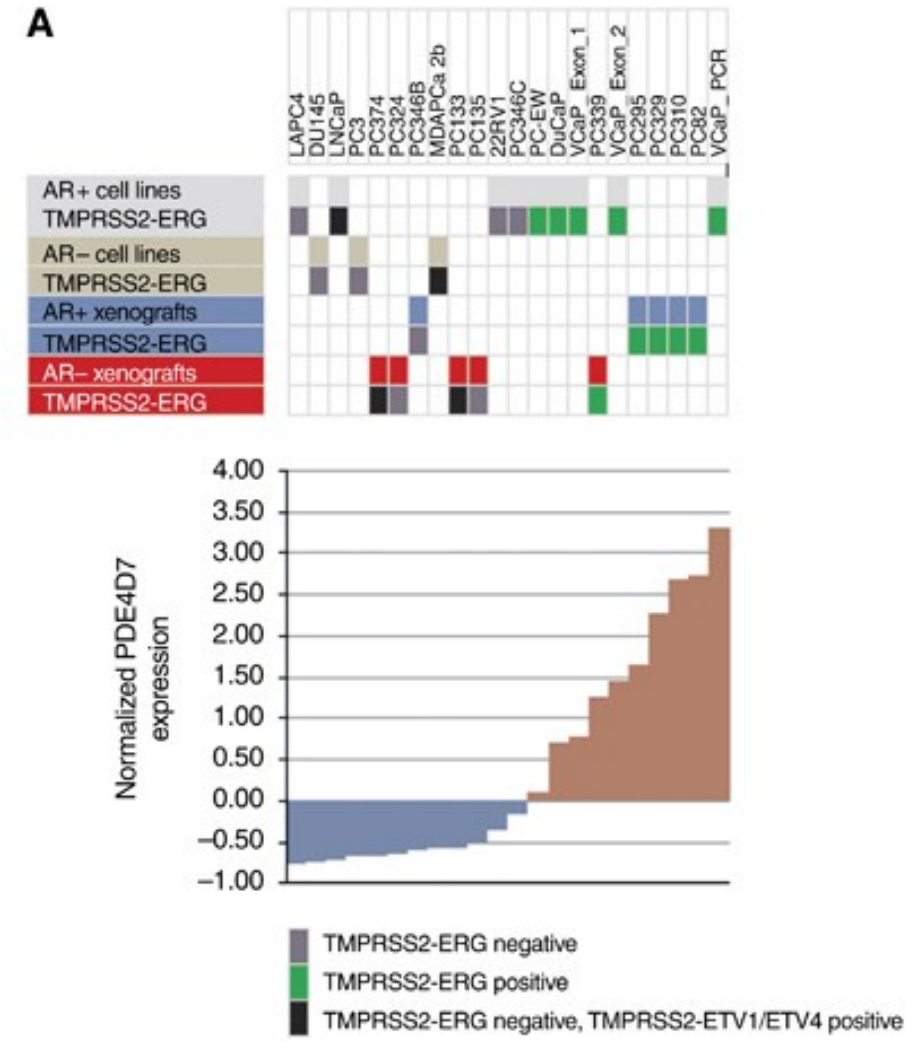
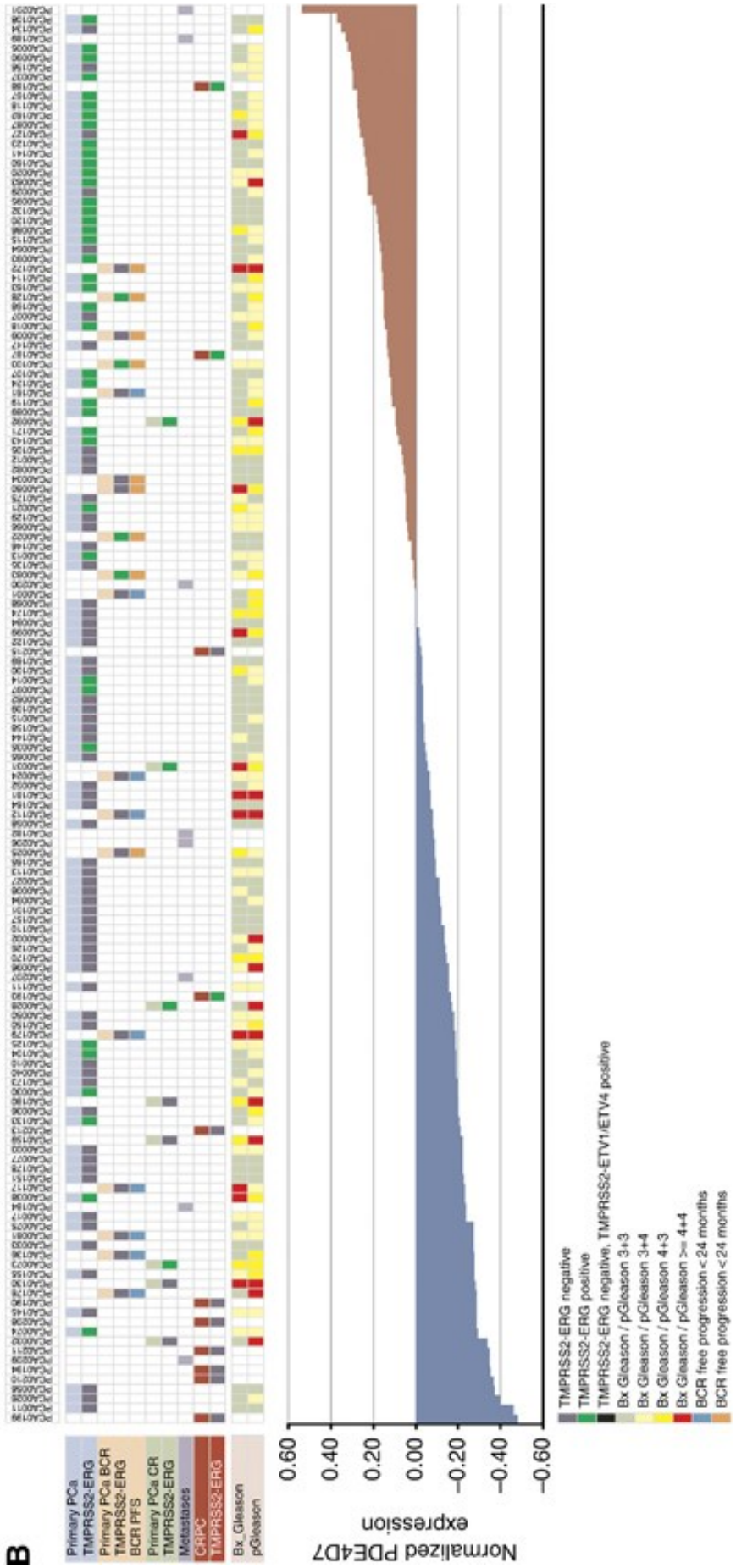


Figure 5: Correlation of PDE4D7 expression in cancerous human prostate tissues to patient outcome. A range of prostate Cancer Cell lines, xenografts, as well as patient prostate cancer tissues (data Taylor et al, 2010; Boormans et al, 2013; Böttcher et al, 2015; Supplementary Table 1) are ranked according to PDE4D7 expression in the respective cells or tissues. The normalised PDE4D7 expression value of each sample was adjusted by subtracting the mean of all expression values of the sample set. Details of cell lines, xenografts and patient samples can be found in Supplementary Table 4. (**A**) PDE4D7 expression in cell lines and xenograft tissues;


B
Primary PCa
TMPRSS2-ERG
Primary PCa BCR
TMPRSS2-ERG
BCR PFS
Primary PCa CR
TMPRSS2-ERG
Metastases
CRPC
TMPRSS2-ERG
Bx_Gleason
pGleason
Normalized PDE4D7
expression
0.60
0.40
0.20
0.00
−0.20
−0.40
−0.60
TMPRSS2-ERG negative
TMPRSS2-ERG positive
TMPRSS2-ERG negative, TMPRSS2-ETV1/ETV4 positive
Bx Gleason / pGleason 3+3
Bx Gleason / pGleason 3+4
Bx Gleason / pGleason 4+3
Bx Gleason / pGleason >= 4+4
BCR free progression <24 months
BCR free progression <24 months

(**B**) PDE4D7 expression in patient samples. For all samples its rank as well as the TMPRSS2-ERG, −ETV1 or −ETV4 fusion status is indicated. The BCR progression free survival (BCR PFS) after surgery (<24 months vs >24 months) is indicated. Further the biopsy Gleason score (Bx_Gleason) as well as the pathology Gleason (pGleason) is given. Samples are categorised into the following groups: AR+ cell lines—androgen-sensitive cell lines; AR− cell lines—androgen-insensitive cell lines; AR+ xenografts—androgen-sensitive xenografts; AR− xenografts—androgen-insensitive xenografts; CRPC—castration resistant prostate cancer; metastases—metastatic tumour; primary PCa—primary prostate cancer, no progression during follow-up; primary PCa BCR—primary prostate cancer, progression to BCR during follow-up; primary PCa CR—primary prostate cancer, progression to CR during follow-up.

To further confirm this, we investigated the PDE4D7 expression in samples of patients that all underwent BCR during follow-up in one data set (30). To segregate the patients into two survival groups, we applied a PDE4D7 expression value which was derived from a ROC analysis between patients who had BCR <24 months vs. patients with BCR >24 months. We determined the unique point of PDE4D7 expression in the ROC curve where the sum of the sensitivity and the specificity becomes a maximum (i.e., <0.51) and used this factor for the Kaplan–Meier analysis. By this we could separate two patient cohorts (HR=0.29; P=6.0E-04) with a median time to BCR after primary treatment of <10 months vs. a median time to BCR of >30 months (Figure 6A). When applying the same cut-off of <0.51 in an analysis of an independent data set (32), we could verify this correlation to time to BCR after surgery (HR=0.36; P=1.6E-03) in this patient cohort with either a median time to BCR of o10 months, or a median time to recurrence >50 months (Figure 6B). The correlation of low PDE4D7 expression to time to BCR after primary treatment was further re-enforced in the second data set (32), where time to CR demonstrated a fivefold increased risk of reaching the endpoint of metastatic disease within a median of 18 months after surgery when applying a cut-off <0.26 for PDE4D7 expression compared with a median time to CR of 95 months if PDE4D7 expression was >0.26 (HR=0.2; P=2.0E-03) (Figure 6C). This data strongly supports our hypothesis that low expression of PDE4D7 correlates with increased short-term biochemical reoccurrence, as well as manifestation of metastatic disease. Most samples collected from CRPC patients demonstrated low PDE4D7 expression levels while again those samples that were positive for the TMPRSS2-ERG fusion gene were measured with increased PDE4D7 transcription (Figure 5B). Whether CRPC patients with positive gene fusion and PDE4D7 expression 40 will survive longer compared with patients with negative TMRPSS2-ERG fusion and PDE expression <0 is a very interesting subject for further research.

Discussion

Analysis of data from large scale genome sequencing projects like TCGA has uncovered a potential role of the PDE4D gene in various types of cancer (35). Indeed, loss of PDE4D was noted as one of the 10 most relevant gene deletion events in 1 study cohort (36). Although PDE4D copy number and, to a lesser degree, mutational status correlates with cancer incidence the role of PDE4D isoform expression has not been studied in a clinical context.

Recent studies have implicated individual PDE4D transcripts in the development of prostate cancer (17, 18). Specifically, we reported for the first time the downregulation of PDE4D7 in hormone-refractory prostate disease represented by a wide range of both cellular and xenograft models (18). Here, we set out to discern whether the differential regulation of PDE4D7 could be verified in human tissue samples collected from primary, as well as metastatic and castration resistant tumours. Encouragingly, across multiple data sets we were able to detect a clear and significant downregulation of PDE4D7 transcript abundance correlating with increasing prostate disease aggressiveness (as assessed by increasing pGleason score and disease stage).

We previously demonstrated that selective knockdown of PDE4D7 expression in androgen-sensitive cell line models led to a more aggressive phenotype, while its overexpression in CRPC cells had the opposite effect (18). The precise details of the cAMP signalling pathways regulated by PDE4D7 during the development of aggressive prostate cancer remain to be uncovered and are subject to future research. However, we would like to propose that PDE4D7 has a contributing role in initial prostate cancer cell states rather than having a 'passenger effect' occurring as a consequence of the molecular changes induced by other factors. To understand the baseline for PDE4D7 expression, and thereby contextualise the differential regulation of this particular PDE isoform during prostate cancer development and progression, we examined its expression status in normal prostate tissue compared with primary and advanced prostate cancers. Notably, the expression of the PDE4D7 transcript was significantly lower in normal, as well as tissue of benign origin compared with low-grade prostate tumours. This leads us to propose a model, where PDE4D7 expression becomes upregulated in primary disease. This, perhaps, reflects an attempt by cells to counteract the proliferative phenotype, before the failure/overcoming of this response leads to PDE4D7 downregulation, which characterises the more aggressive prostate tumours. Thus PDE4D7 appears to be functionally involved in the primary development of prostatic tumours. However, our data suggests that future cellular and molecular studies could usefully be directed to ascertain whether the initial upregulation of PDE4D7 is intimately involved in the initial stage of prostate tumorigenesis.

Interestingly, we uncover here a novel link between AR signalling and PDE4D7 expression by correlating the incidence of TMPRSS2-ERG gene fusion and PDE4D7 transcript levels. The TMPRSS2-ERG gene fusion between the prostate specific serine protease TMPRSS2 and the ETS transcription factor family member ERG was first detected in 2005 by a statistical outlier approach (29). Subsequently, this gene fusion has been shown to be present in B50%

of prostate cancer patients and is, consequently, one of the most prominent genomic fusion events reported in prostate cancer (37). This translocation results in androgen-regulated ERG expression such that the androgen-responsive promoter of TMPRSS2 now drives TMPRSS2-ERG expression, resulting in an upregulation in both the expression and activity of the transcription factor, ERG (29). However, despite numerous studies the clinical implications and functional consequences of the genomic fusion remain to be fully understood (38–41). Here, we uncover a remarkably significant difference in PDE4D7 expression between TMPRSS2-ERG-negative and TMPRSS2-ERG-positive tumour samples. Indeed, when stratified by TMPRSS2-ERG incidence it is clear that PDE4D7 is most significantly upregulated in low-grade TMPRSS2-ERG-positive tumours. This raises the possibility that PDE4D7 expression may be directly or indirectly regulated by the aberrant transcriptional activity of the TMPRSS2-ERG fusion protein. Inspection of the PDE4D gene reveals several putative binding sites for ERG, one within the promoter region of PDE4D7 (Materials and Methods). It would therefore seem logical that if PDE4D7 is regulated by ERG transcription, an increase in the expression of the androgen-regulated TMPRSS2-ERG factor would lead to a concurrent androgen-driven increase in PDE4D7 expression.

To date, most newly detected prostate cancer cases are clinically classified low-risk diseases (42). It is crucial to understand the natural history of these tumours as it is under considerable debate whether and to what extent low-risk Gleason 6 tumours are able to progress to higher grade tumours leading to metastatic spread or even cancer-specific death (43, 44). Interestingly, our data may indicate that reduced expression of PDE4D7 in low to intermediate Gleason tumours is correlated to progression after primary treatment. Although initially positively correlated with tumour development, the expression of PDE4D7 actually appears to be protective against further disease progression, which is in line with the data previously obtained regarding the cellular functioning of PDE4D7 (18). As new strategies for targeted pharmacological manipulation of specific PDE4D transcripts become available then PDE4D7 likely provides a promising future target in the treatment of primary and/ or advanced prostate cancer. Our data indicate that during tumour progression the risk of fast recurrence to clinical endpoints like biochemical or clinical disease is correlated to the level of PDE4D7 expression in the primary tumour. Consequently, patients with a low expression level of PDE4D7 in their primary cancers after surgical resection may very well be candidates for immediate adjuvant treatment like radiotherapy and/or androgen ablation. Furthermore, the manipulation of PDE4D7 suggests a strategy to selectively treat TMPRSS2-ERG fusion-positive prostate cancers. However, the success of such strategy may depend on the stratification into molecular sub-types according to the status of the TMPRSS2-ERG gene translocation.

The data presented here demonstrates the relevance of PDE4D7 as a potential biomarker for more accurate prostate cancer diagnostics. In particular, we have demonstrated the potential role of this specific splice variant of the PDE4D gene for prognosis of aggressive prostate cancer in the molecular sub-type of TMPRSS2-ERG-positive prostate tumours as well as its role as a putative target gene for therapy of primary vs. late-stage, hormone-refractory disease.

Acknowledgements

This study was supported by the framework of CTMM (The Netherlands), the Center for Translational Molecular Medicine, PCMM project (grant 03O-203). We would like to acknowledge support from the Biotechnology and Biological Sciences Research Council (UK) for award of a CASE studentship to DJPH. BBSRC funding to GSB (BB/G01647X/1). We also like to thank Drs Chris Bangma and Mark Wildhagen for their support regarding patient sample collection and annotation.

References

1. Ferlay,J., Soerjomataram I,I., Dikshit,R., Eser,S., Mathers,C., Rebelo,M., Parkin,D.M., Forman D,D. and Bray,F. (2014) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, **136**, E359–86.

2. Andriole,G.L., Crawford,E.D., Grubb,R.L., Buys,S.S., Chia,D., Church,T.R., Fouad,M.N., Gelmann,E.P., Kvale,P.A., Reding,D.J., *et al.* (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.

3. Schröder,F.H., Hugosson,J., Roobol,M.J., Tammela,T.L.J., Ciatto,S., Nelen,V., Kwiatkowski,M., Lujan,M., Lilja,H., Zappa,M., *et al.* (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.

4. Merkle,D. and Hoffmann,R. (2011) Roles of cAMP and cAMP-dependent protein kinase in the progression of prostate cancer: Cross-talk with the androgen receptor. *Cell. Signal.*, **23**, 507–515.

5. Conti,M. and Beavo,J. (2007) Biochemistry and physiology of cyclic nucleotide phosphodiesterases: essential components in cyclic nucleotide signaling. *Annu. Rev. Biochem.*, **76**, 481–511.

6. Maurice,D.H., Ke,H., Ahmad,F., Wang,Y., Chung,J. and Manganiello,V.C. (2014) Advances in targeting cyclic nucleotide phosphodiesterases. *Nat. Rev. Drug Discov.*, **13**, 290–314.

7. Lugnier,C. (2006) Cyclic nucleotide phosphodiesterase (PDE) superfamily: A new target for the development of specific therapeutic agents. *Pharmacol. Ther.*, **109**, 366–398.

8. Houslay,M.D. (2010) Underpinning compartmentalised cAMP signalling through targeted cAMP breakdown. *Trends Biochem. Sci.*, **35**, 91–100.

9. Houslay,M.D., Baillie,G.S. and Maurice,D.H. (2007) cAMP-Specific phosphodiesterase-4 enzymes in the cardiovascular system: a molecular toolbox for generating compartmentalized cAMP signaling. *Circ. Res.*, **100**, 950–66.

10. Francis,S.H., Blount,M.A. and Corbin,J.D. (2011) Mammalian cyclic nucleotide phosphodiesterases: molecular mechanisms and physiological functions. *Physiol. Rev.*, **91**, 651–90.

11. Lee,H., Graham,J.M., Rimoin,D.L., Lachman,R.S., Krejci,P., Tompson,S.W., Nelson,S.F., Krakow,D. and Cohn,D.H. (2012) Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am. J. Hum. Genet.*, **90**, 746–751.

12. Michot,C., Le Goff,C., Goldenberg,A., Abhyankar,A., Klein,C., Kinning,E., Guerrot,A.M., Flahaut,P., Duncombe,A., Baujat,G., *et al.* (2012) Exome sequencing identifies PDE4D mutations as another cause of acrodysostosis. *Am. J. Hum. Genet.*, **90**, 740–745.

13. Apuhan,T., Gepdiremen,S., Arslan,A.O. and Aktas,G. (2013) Evaluation of patients with nasal polyps about the possible association of desmosomal junctions, RORA and PDE4D gene. *Eur. Rev. Med. Pharmacol. Sci.*, **17**, 2680–3.

14. Kaname,T., Ki,C.-S., Niikawa,N., Baillie,G.S., Day,J.P., Yamamura,K.-I., Ohta,T., Nishimura,G.,

Mastuura,N., Kim,O.-H., *et al.* (2014) Heterozygous mutations in cyclic AMP phosphodiesterase-4D (PDE4D) and protein kinase A (PKA) provide new insights into the molecular pathology of acrodysostosis. *Cell. Signal.*, **26**, 2446–59.

15. Yoon,H.-K., Hu,H.-J., Rhee,C.-K., Shin,S.-H., Oh,Y.-M., Lee,S.-D., Jung,S.-H., Yim,S.-H., Kim,T.-M. and Chung,Y.-J. (2014) Polymorphisms in PDE4D are associated with a risk of COPD in non-emphysematous Koreans. *COPD*, **11**, 652–8.

16. Houslay,M.D. (2005) The long and short of vascular smooth muscle phosphodiesterase-4 as a putative therapeutic target. *Mol. Pharmacol.*, **68**, 563–7.

17. Rahrmann,E.P., Collier,L.S., Knutson,T.P., Doyal,M.E., Kuslak,S.L., Green,L.E., Malinowski,R.L., Roethe,L., Akagi,K., Waknitz,M., *et al.* (2009) Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping beauty transposon-based somatic mutagenesis screen. *Cancer Res.*, **69**, 4388–4397.

18. Henderson,D.J.P., Byrne,A., Dulla,K., Jenster,G., Hoffmann,R., Baillie,G.S. and Houslay,M.D. (2014) The cAMP phosphodiesterase-4D7 (PDE4D7) is downregulated in androgen-independent prostate cancer cells and mediates proliferation by compartmentalising cAMP at the plasma membrane of VCaP prostate cancer cells. *Br. J. Cancer*, **110**, 1278–87.

19. Gretarsdottir,S., Thorleifsson,G., Reynisdottir,S.T., Manolescu,A., Jonsdottir,S., Jonsdottir,T., Gudmundsdottir,T., Bjarnadottir,S.M., Einarsson,O.B., Gudjonsdottir,H.M., *et al.* (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.*, **35**, 131–138.

20. Hoffmann,R., Wilkinson,I.R., McCallum,J.F., Engels,P. and Houslay,M.D. (1998) cAMP-specific phosphodiesterase HSPDE4D3 mutants which mimic activation and changes in rolipram inhibition triggered by protein kinase A phosphorylation of Ser-54: generation of a molecular model. *Biochem. J.*, **333**, 139–49.

21. MacKenzie,K.F., Wallace,D.A., Hill,E. V, Anthony,D.F., Henderson,D.J.P., Houslay,D.M., Arthur,J.S.C., Baillie,G.S. and Houslay,M.D. (2011) Phosphorylation of cAMP-specific PDE4A5 (phosphodiesterase-4A5) by MK2 (MAPKAPK2) attenuates its activation through protein kinase A phosphorylation. *Biochem. J.*, **435**, 755–69.

22. Sheppard,C.L., Lee,L.C.Y., Hill,E. V, Henderson,D.J.P., Anthony,D.F., Houslay,D.M., Yalla,K.C., Cairns,L.S., Dunlop,A.J., Baillie,G.S., *et al.* (2014) Mitotic activation of the DISC1-inducible cyclic AMP phosphodiesterase-4D9 (PDE4D9), through multi-site phosphorylation, influences cell cycle progression. *Cell. Signal.*, **26**, 1958–74.

23. Wang,D., Deng,C., Bugaj-Gaweda,B., Kwan,M., Gunwaldsen,C., Leonard,C., Xin,X., Hu,Y., Unterbeck,A. and De Vivo,M. (2003) Cloning and characterization of novel PDE4D isoforms PDE4D6 and PDE4D7. *Cell. Signal.*, **15**, 883–91.

24. Hoffmann,R., Baillie,G.S., MacKenzie,S.J., Yarwood,S.J. and Houslay,M.D. (1999) The MAP kinase ERK2 inhibits the cyclic AMP-specific phosphodiesterase HSPDE4D3 by phosphorylating it at Ser579. *EMBO J.*, **18**, 893–903.

25. Baillie,G.S., MacKenzie,S.J., McPhee,I. and Houslay,M.D. (2000) Sub-family selective actions in the ability of Erk2 MAP kinase to phosphorylate and regulate the activity of PDE4 cyclic AMP-specific phosphodiesterases. *Br. J. Pharmacol.*, **131**, 811–9.

26. Byrne,A.M., Elliott,C., Hoffmann,R. and Baillie,G.S. (2015) The activity of cAMP-phosphodiesterase 4D7 (PDE4D7) is regulated by protein kinase A-dependent phosphorylation within its unique N-terminus. *FEBS Lett.*, **589**, 750–5.

27. Purdom,E., Simpson,K.M., Robinson,M.D., Conboy,J.G., Lapuk,A. V and Speed,T.P. (2008) FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, **24**, 1707–14.

28. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

29. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.-W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R., *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (80-.).*, **310**, 644–8.

30. Taylor,B.S., Schultz,N., Hieronymus,H., Gopalan,A., Xiao,Y., Carver,B.S., Arora,V.K., Kaushik,P., Cerami,E., Reva,B., *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell*, **18**, 11–22.

31. Brase,J.C., Johannes,M., Mannsperger,H., Fälth,M., Metzger,J., Kacprzyk,L.A., Andrasiuk,T., Gade,S., Meister,M., Sirma,H., *et al.* (2011) TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. *BMC Cancer*, **11**, 507.

32. Boormans,J.L., Korsten,H., Ziel-van der Made,A.J.C., van Leenders,G.J.L.H., de Vos,C. V, Jenster,G. and Trapman,J. (2013) Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. *Int. J. Cancer*, **133**, 335–45.

33. Böttcher,R., Hoogland,A.M., Dits,N., Verhoef,E.I., Kweldam,C., Waranecki,P., Bangma,C.H., van Leenders,G.J.L.H. and Jenster,G. (2015) Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. *Oncotarget*.

34. Erho,N., Crisan,A., Vergara,I.A., Mitra,A.P., Ghadessi,M., Buerki,C., Bergstralh,E.J., Kollmeyer,T., Fink,S., Haddad,Z., *et al.* (2013) Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One*, **8**, e66855.

35. Zack,T.I., Schumacher,S.E., Carter,S.L., Cherniack,A.D., Saksena,G., Tabak,B., Lawrence,M.S., Zhang,C.-Z., Wala,J., Mermel,C.H., *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.

36. Baca,S.C., Prandi,D., Lawrence,M.S., Mosquera,J.M., Romanel,A., Drier,Y., Park,K., Kitabayashi,N., MacDonald,T.Y., Ghandi,M., *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666–77.

37. Kumar-Sinha,C., Tomlins,S.A. and Chinnaiyan,A.M. (2008) Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, **8**, 497–511.

38. Petrovics,G., Liu,A., Shaheduzzaman,S., Furusato,B., Furasato,B., Sun,C., Chen,Y., Nau,M., Ravindranath,L., Chen,Y., *et al.* (2005) Frequent overexpression of ETS-related gene-1 (ERG1) in prostate cancer transcriptome. *Oncogene*, **24**, 3847–52.

39. Mosquera,J.-M., Perner,S., Demichelis,F., Kim,R., Hofer,M.D., Mertz,K.D., Paris,P.L., Simko,J., Collins,C., Bismar,T.A., *et al.* (2007) Morphological features of TMPRSS2-ERG gene fusion prostate cancer. *J. Pathol.*, **212**, 91–101.

40. Saramäki,O.R., Harjula,A.E., Martikainen,P.M., Vessella,R.L., Tammela,T.L.J. and Visakorpi,T. (2008) TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis. *Clin. Cancer Res.*, **14**, 3395–400.

41. Hermans,K.G., Boormans,J.L., Gasi,D., van Leenders,G.J.H.L., Jenster,G., Verhagen,P.C.M.S. and Trapman,J. (2009) Overexpression of prostate-specific TMPRSS2(exon 0)-ERG fusion transcripts corresponds with favorable prognosis of prostate cancer. *Clin. Cancer Res.*, **15**, 6398–403.

42. Bangma,C.H. and Roobol,M.J. (2012) Defining and predicting indolent and low risk prostate cancer. *Crit. Rev. Oncol. Hematol.*, **83**, 235–41.

43. Whittemore,A.S., Keller,J.B. and Betensky,R. (1991) Low-grade, latent prostate cancer volume: predictor of clinical cancer incidence? *J. Natl. Cancer Inst.*, **83**, 1231–5.

44. Sowalsky,A.G., Ye,H., Bubley,G.J. and Balk,S.P. (2013) Clonal progression of prostate cancers from Gleason grade 3 to grade 4. *Cancer Res.*, **73**, 1050–5.

Chapter 5

Human PDE4D isoform composition is deregulated in primary prostate cancer and indicative for disease progression and development of distant metastases

René Böttcher[1,2], Kalyan Dulla[3], Dianne van Strijp[3], Natasja Dits[1], Esther I. Verhoef[5], George S. Baillie[4], Geert J.L.H. van Leenders[5], Miles D. Houslay[6], Guido Jenster[1], Ralf Hoffmann[3,4]

1 Department of Urology, Erasmus Medical Center, Rotterdam, The Netherlands
2 Department of Bioinformatics, Technical University of Applied Sciences Wildau, Wildau, Germany
3 Department of Oncology Solutions and Precision Diagnostics, Philips Research Europe, Eindhoven, The Netherlands
4 Institute of Cardiovascular and Medical Science, University of Glasgow, Glasgow, Scotland, UK
5 Department of Pathology, Erasmus Medical Center, Rotterdam, The Netherlands
6 Institute of Pharmaceutical Science, King's College London, London, UK

Published in

Oncotarget (2016) [epub ahead of print]

Supplementary Material is available via

http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=rt&op=suppFiles&path[]=12204&path[]=0

Abstract

Phosphodiesterase 4D7 was recently shown to be specifically over-expressed in localized prostate cancer, raising the question as to which regulatory mechanisms are involved and whether other isoforms of this gene family (*PDE4D*) are affected under the same conditions. We investigated PDE4D isoform composition in prostatic tissues using a total of seven independent expression datasets and also included data on DNA methylation, copy number and AR and ERG binding in PDE4D promoters to gain insight into their effect on PDE4D transcription.

We show that expression of PDE4D isoforms is consistently altered in primary human prostate cancer compared to benign tissue, with PDE4D7 being up-regulated while PDE4D5 and PDE4D9 are down-regulated. Disease progression is marked by an overall down-regulation of long PDE4D isoforms, while short isoforms (PDE4D1/2) appear to be relatively unaffected. While these alterations seem to be independent of copy number alterations in the PDE4D locus and driven by AR and ERG binding, we also observed increased DNA methylation in the promoter region of PDE4D5, indicating a long lasting alteration of the isoform composition in prostate cancer tissues.

We propose two independent metrics that may serve as diagnostic and prognostic markers for prostate disease: $(PDE4D7 - PDE4D5)$ provides an effective means for distinguishing PCa from normal adjacent prostate, whereas $PDE4D1/2 - (PDE4D5 + PDE4D7 + PDE4D9)$ offers strong prognostic potential to detect aggressive forms of PCa and is associated with metastasis free survival. Overall, our findings highlight the relevance of PDE4D as prostate cancer biomarker and potential drug target.

Introduction

With an estimated 417,000 new cases in 2014 in Europe, prostate cancer (PCa) remains the most often diagnosed gender-specific carcinoma for men (1). The current routine of diagnosing PCa results in a significant number of unnecessary biopsies and treatments of non-cancerous, benign prostate conditions and non-aggressive cancers, leading to severe negative effects for both men and healthcare systems (2, 3).

Next to well-studied pathways such as androgen receptor (AR) and PI3K/AKT, cyclic AMP (cAMP) has been shown to play a role in the development and progression of PCa (4). The metabolism of cAMP in cells is complex and tailored by spatial and signalling cross-talk considerations involving both a large family of adenylyl cyclases responsible for its synthesis, and a large family of cyclic nucleotide phosphodiesterases (PDEs) responsible for its degradation (5). It is now well recognized that when particular cAMP degrading PDEs are recruited to specific signalling complexes they create and control cAMP gradients around them, allowing spatially compartmentalised and time-dependent regulation of localized cAMP signalling (6, 7). Protein domains involved in subcellular localization as well as independent regulatory mechanisms play a pivotal role in these processes, granting PDE isoforms the ability to fulfil functionally independent and unique roles in the cell (6, 8). Thus, changes in the expression of distinct PDE isoforms can be expected to reprogram downstream signalling pathways during disease development and progression, providing potential targets for novel markers and therapeutic interventions (6). Indeed, cAMP-degrading PDEs have been associated with several diseases in recent years, including stroke, acrodysostosis and COPD (9–14), and more recently, expression of a specific PDE4D isoform (PDE4D7) has been related to prostate cancer (15, 16).

The PDE4D7 transcript comprises the open reading frame for a long PDE4D isoform that contains both the UCR1 and UCR2 regulatory domains (17). These protein domains are common to all long PDE4D isoforms with UCR1 being phosphorylated by PKA (cAMP dependent protein kinase A), when cAMP levels within the cell are elevated, leading to enzyme activation (18, 19). Indeed, activation of long PDE4 isoforms, such as PDE4D7, by PKA provides a fundamental part of the cellular desensitization process to cAMP (6). Long PDE4 isoforms can also be dynamically regulated through phosphorylation by other key signalling system kinases, namely, by ERK (20), MK2 (21), Cdk5 (22) and AMPK (23). Additionally, PDE4D7 has been shown (15) to be specifically targeted to the sub-plasma membrane compartment in prostate cancer cells where it regulates local cAMP levels that are linked to cell proliferation (15).

We have previously shown that PDE4D7 is specifically overexpressed in both androgen sensitive PCa cells and in samples from patients with early androgen sensitive prostate disease (15, 16). However, in marked contrast to this, once PCa cells become androgen insensitive/independent (castration resistant), expression of PDE4D7 declines (15, 16).
Here, we show that PDE4D isoform composition is altered in localized prostate cancer and that it can be used both as a diagnostic as well as a prognostic biomarker. In conjunction with

our previous studies, we see that the long transcript isoform PDE4D7 is up-regulated in localized disease compared to normal adjacent prostate (NAP), while its expression diminishes with tumour progression. In contrast to PDE4D7, two other long isoforms, PDE4D5 and PDE4D9, do not undergo an initial up-regulation in primary PCa and instead are increasingly down-regulated during disease progression. Moreover, we suggest that this change in isoform composition may be influenced by the DNA methylation of specific regulatory elements of the PDE4D locus. These findings highlight the potential of using condition-specific mRNA isoforms of the PDE4D gene as biomarkers and potential novel therapy targets to restore benign conditions.

Results

The long isoforms PDE4D5 and PDE4D9 are significantly down-regulated in primary prostate cancer, independent of copy number alterations in the PDE4D gene locus

After previously identifying PDE4D7 as a novel biomarker candidate (16), we wanted to investigate the behaviour of other PDE4D transcript isoforms in PCa development and progression. Therefore, we focused on the nine major human PDE4D isoforms described in RefSeq and conducted a meta-analysis of six publicly available patient cohorts. Our analysis revealed that many PDE4D isoforms are seemingly expressed at stable levels when using Exon Arrays, whereas only PDE4D1/2, PDE4D5, PDE4D7, and PDE4D9 were detectable at higher levels in our independent qRT-PCR cohort of prostate tissues (see Figure 1 and Supplementary Figures 1-5). These findings were supported by the TCGA PRAD RNA-seq cohort, which mostly agreed with RT-PCR results, despite few outlier samples showing expression of other isoforms (Supplementary Figure 6). Based on these findings, we focused on the above mentioned PDE4D isoforms, as they showed consistent expression profiles in all used cohorts. Using these criteria, we found that both PDE4D5 and PDE4D9 are significantly down-regulated in primary localized PCa when compared to benign samples. Moreover, patient samples derived from castration-resistant prostate cancer (CRPC) showed further down-regulation of both isoforms, in line with our previous findings for PDE4D7 (16). Likewise, PCa metastasis samples followed this trend, but often displayed higher variance in PDE4D isoform expression, as can be expected given their very heterogeneous genomic background (24).

Since, partial or complete deletions of one or both alleles of the *PDE4D* gene have been reported previously in prostate cancer (25–27) we utilized TCGA SNP array data of matching patient samples to assess the potential impact of deletions occurring in *PDE4D* on isoform expression. Although we did observe a significant reduction in gene expression upon loss of genetic material, both isoforms were also expressed at significantly lower levels in PCa samples that did not harbour a deletion when comparing to matching normal samples (Figure 2).

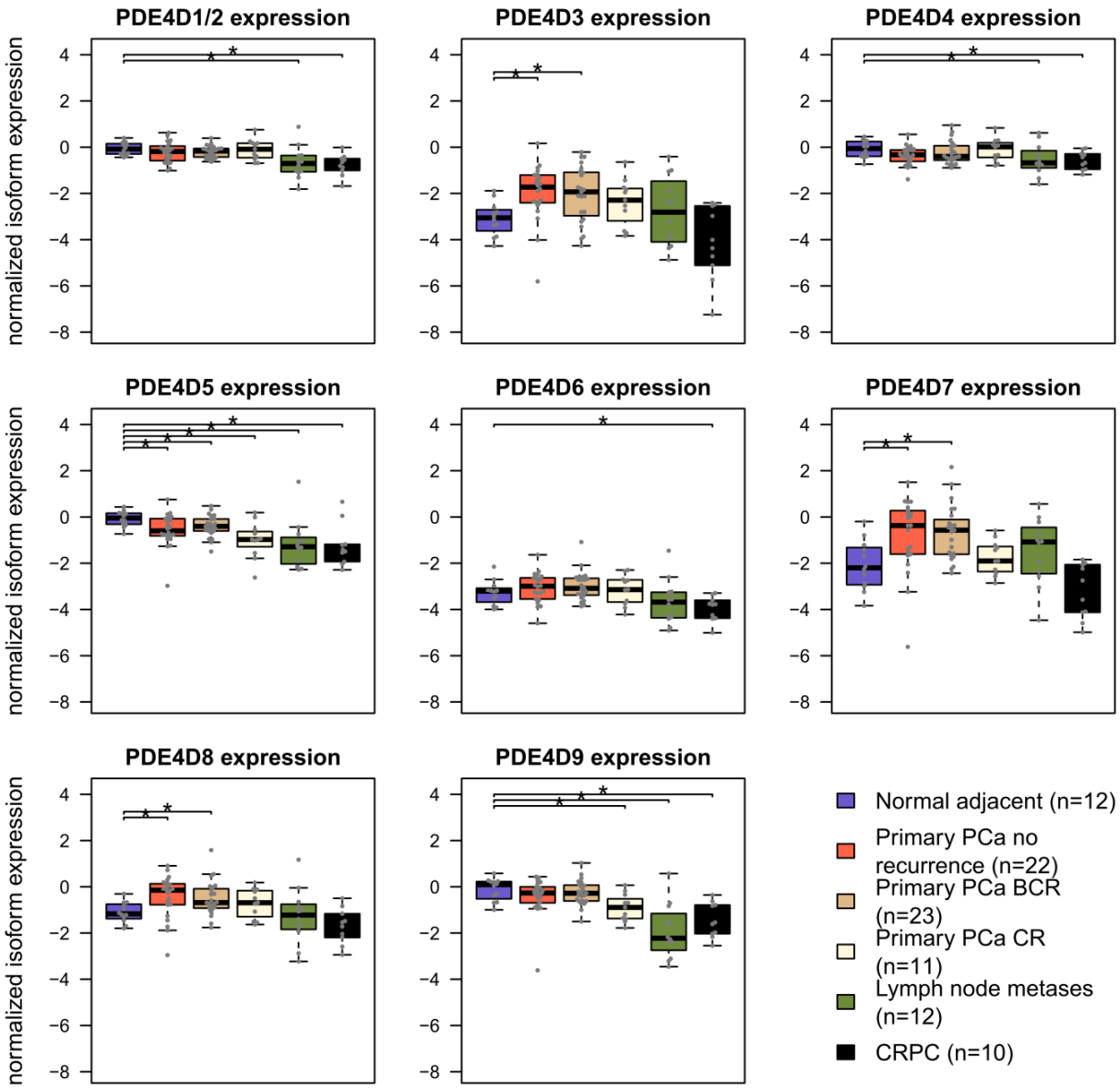


Figure 1: Overview of PDE4D isoform expression in prostatic tissues. Normalized PDE4D isoform expression in the EMC dataset across different prostate conditions. CR – clinical recurrence, BCR – biochemical recurrence, CPRC – castration resistant prostate cancer. Significant differences ($p < 0.05$, Wilcoxon-Mann-Whitney test) are indicated with *.

Androgen receptor and ERG are implicated in transcriptional regulation of PDE4D

Our previous work suggested an association between PDE4D7 expression and the presence of the TMPRSS2-ERG fusion gene (16). We therefore set out to investigate whether there was any comparable ERG involvement in the expression of PDE4D5, PDE4D7 and PDE4D9 in prostate disease. In order to do this, we assigned localized PCa samples to one of two groups based on an unsupervised clustering of ERG expression values by Partitioning Around Medoids and used available ERG IHC information of the EMC cohort to confirm the validity of this approach. Clustering based grouping showed good concordance with IHC results,

assigning four additional samples (10.2%) to the ERG positive group (Supplementary Figure 7).

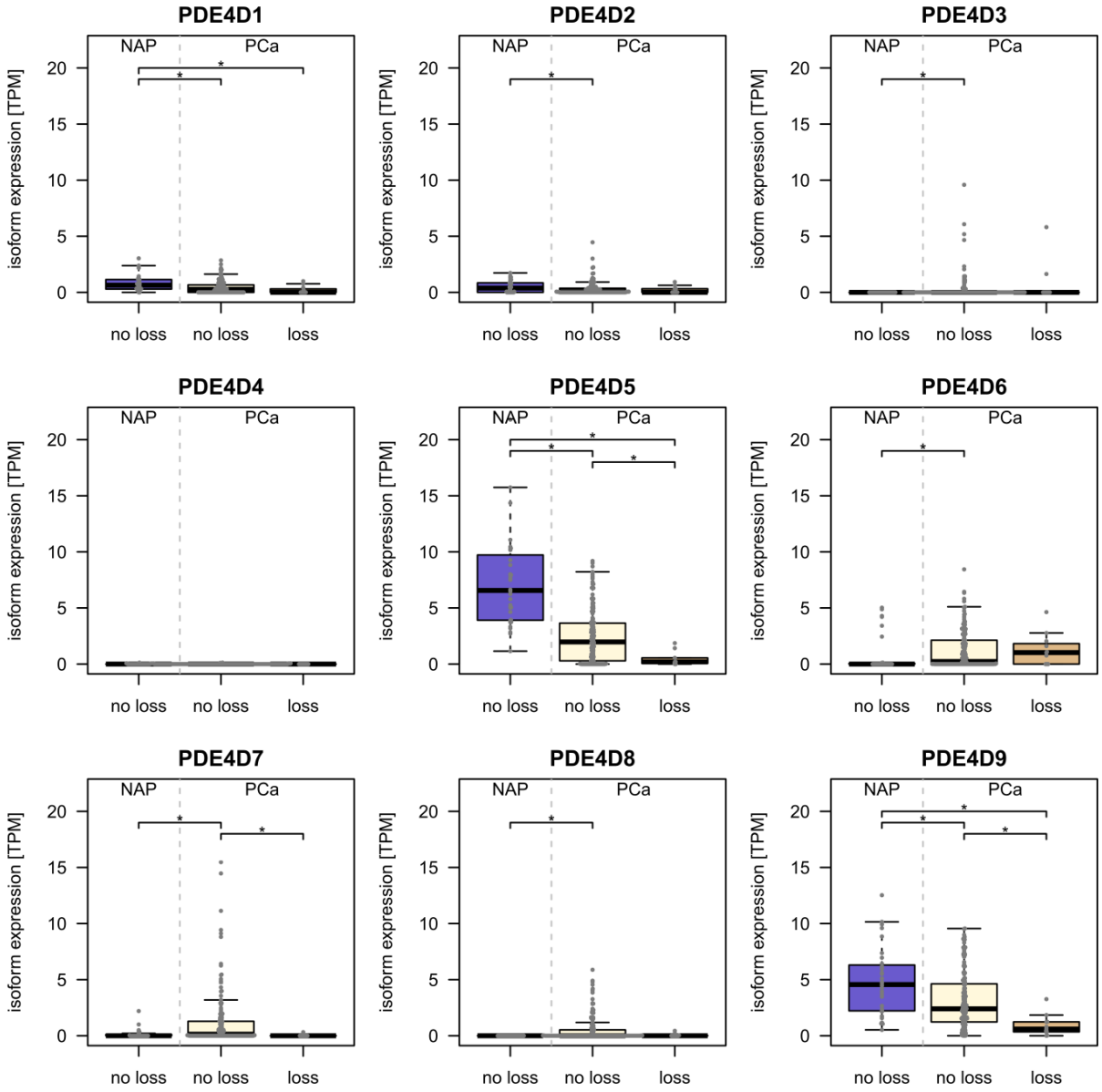


Figure 2: Relation of copy number events and PDE4D expression in the TCGA cohort. 32 normal adjacent prostate samples are compared to PCa samples with (n=12) and without (n=171) loss of genetic material in the PDE4D locus to investigate whether decreased expression occurs independently of PDE4D deletions. Significant differences in expression are denoted with * ($p < 0.05$, Wilcoxon-Mann-Whitney test).

Interestingly, while we were able to confirm PDE4D7 overexpression in ERG positive PCa samples, PDE4D1/2 and PDE4D9 seemed unaffected by ERG, whereas PDE4D5 expression was altered significantly in two out of five datasets, suggesting that any connection between PDE4D5 and ERG is weak at best (see Figure 3). Of note, the Erho dataset consistently showed significant changes for all isoforms, however, these likely do not reflect real events, as absolute log2 fold changes were small ($|log2FC| < 1$) except for PDE4D7 (data not shown). Therefore, ERG linkage discriminates between PDE4D7 and the grouping of PDE4D1/2,

PDE4D5 and PDE4D9, where we see differences between these two groups in the change of their expression in prostate disease.

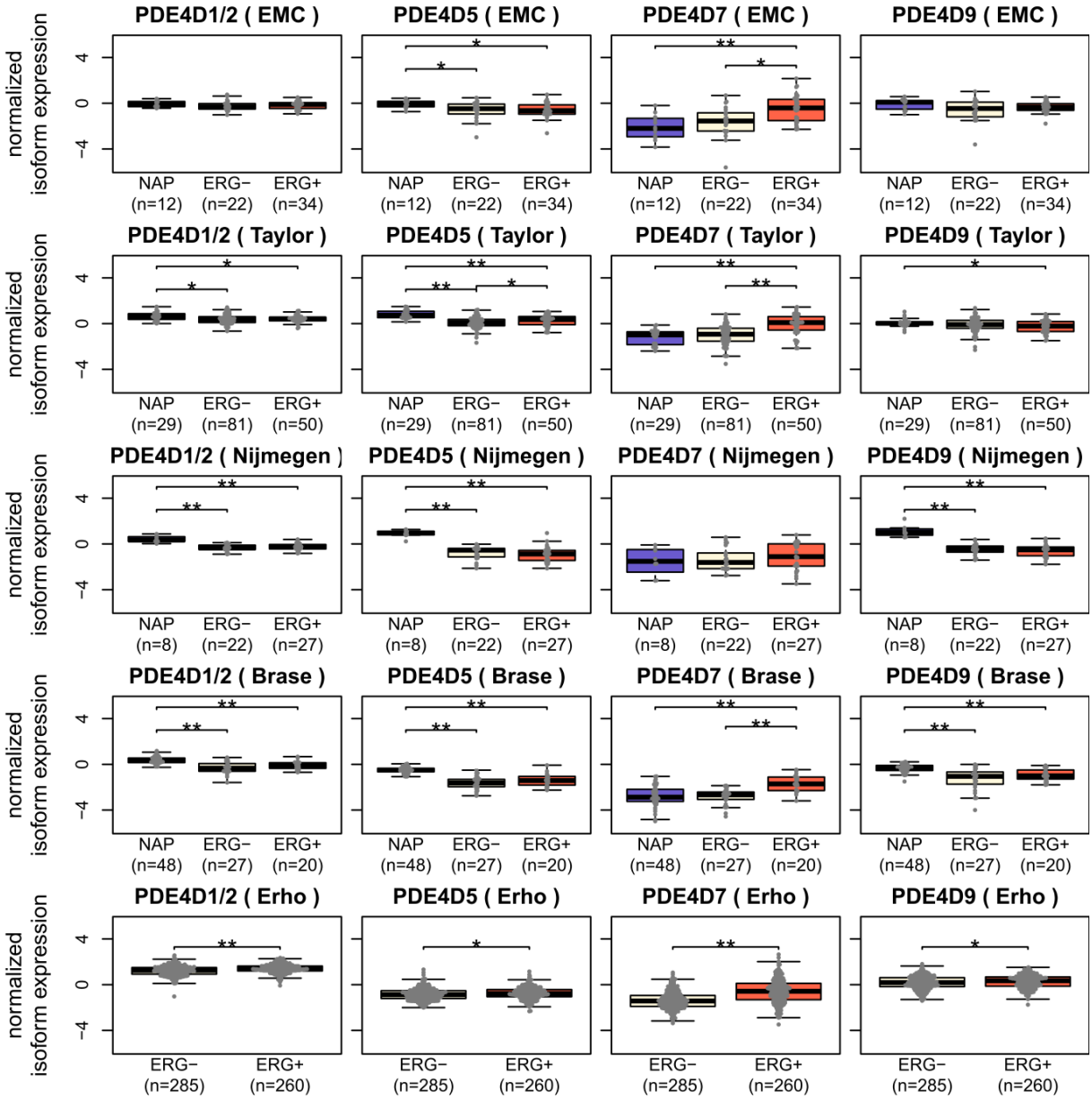


Figure 3: Investigating potential ERG regulation of PDE4D isoforms. Since only PDE4D7 has been previously reported as up-regulated in ERG positive PCa samples (16), expression of PDE4D1/2, PDE4D5, PDE4D7 and PDE4D9 was tested in ERG negative and ERG positive samples across five Exon Array datasets (* = $p < 0.05$, ** = $p < 0.001$, Wilcoxon-Mann-Whitney test).

To investigate androgen-dependence of PDE4D isoform expression, we incorporated a public dataset of LNCaP cells measured after being kept either in androgen stripped medium (using dextran-coated charcoal - DCC) or after addition of the synthetic androgen R1881 (28). While PDE4D9 expression was not altered after treatment, both PDE4D5 and PDE4D7 showed significant differences in expression after R1881 addition (Figure 4). Specifically, PDE4D5 expression appeared to be inhibited upon AR stimulation, while PDE4D7 was up-regulated in DCC by the synthetic androgen R1881 in LNCaP cells.

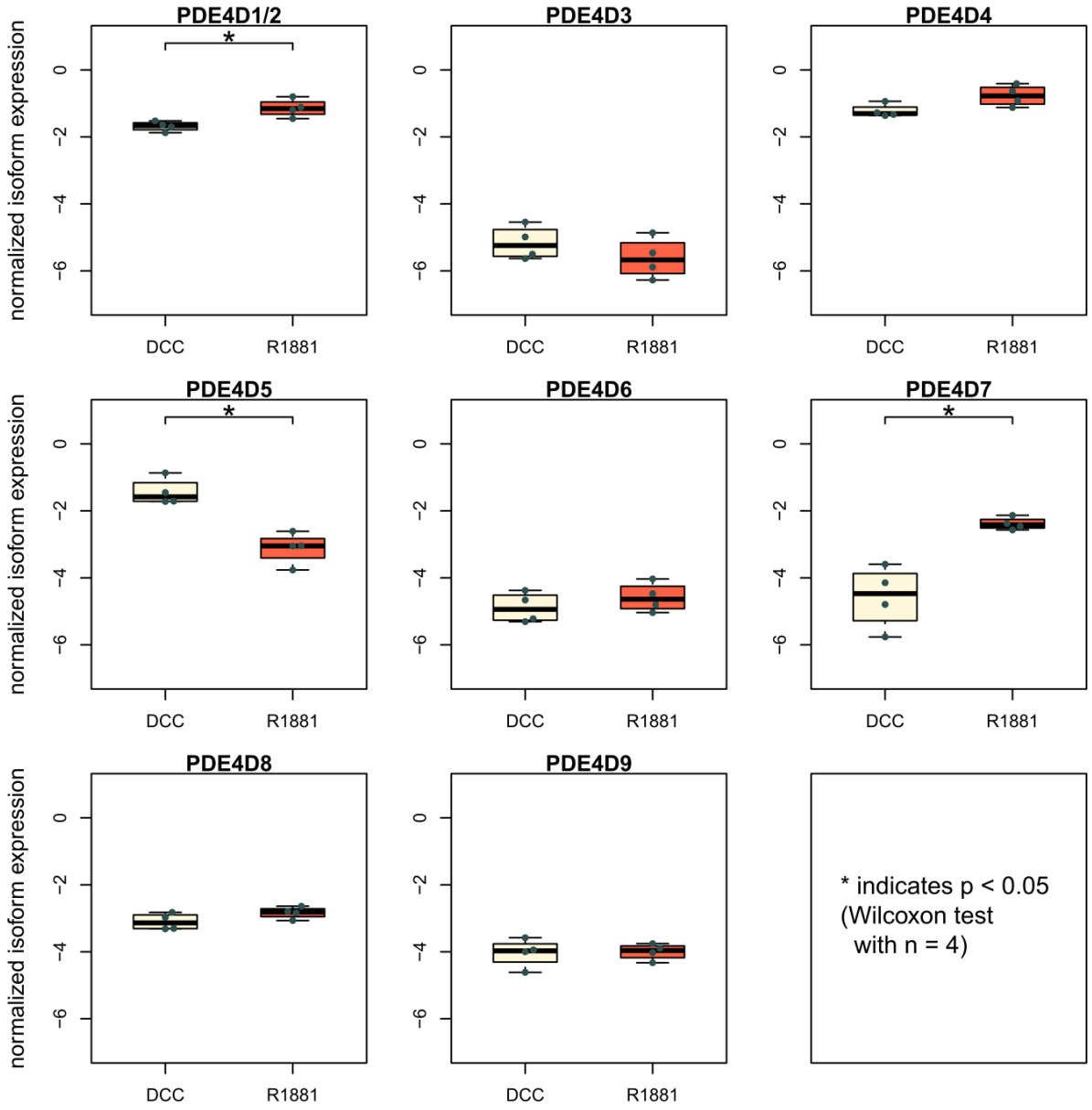


Figure 4: Investigation of androgen receptor involvement in PDE4D expression. Expression of PDE4D isoforms in LNCaP cells with or without addition of the synthetic androgen R1881 (28).

Next, we made use of public ChIP-seq data from the VCaP PCa cell line (29) treated with R1881 in order to gather further evidence of AR involvement in PDE4D expression. In ChIP-seq, DNA binding proteins and associated chromatin are cross-linked, followed by immunoprecipitation of a protein of interest and subsequent sequencing of the associated DNA fragments, allowing a genome-wide localisation of its DNA binding sites. Overall, we found 31 ChIP-seq peaks for AR in PDE4D, two of which were near the first exon of PDE4D7 (~2 kb and 3 kb upstream), while another was partially overlapping the first exon of PDE4D5 (see Supplementary Table 1). No peaks could be found in proximity to the PDE4D9 transcription start site (TSS), as the closest upstream and downstream peaks were found at an approximate distance of 85.5 kb and 44.2 kb, respectively. Since VCaP harbours the TMPRSS2-ERG gene fusion and ChIP-seq data for ERG was available from the same source,

we included it in our analysis and found 43 ERG peaks in the PDE4D gene locus, of which some were found to partially overlap the first exon of each of the long isoforms PDE4D5, PDE4D7 and PDE4D9 (see Figure 5 and Supplementary Table 2). Since the number of ChIP-seq peaks located in PDE4D appears to be rather high, we were wondering whether binding of AR and/or ERG within the gene locus occurs more often as compared to other regions. For this reason, we counted the number of AR and ERG peaks in 21,209 RefSeq gene loci and used these counts to construct empirical cumulative distribution functions (ECDFs) for both transcription factors. These ECDFs model the background distribution of the counts for both AR and ERG across all genes and enable us to calculate in which percentile the peak counts for AR and ERG in PDE4D are falling. Surprisingly, both AR and ERG were among the top 99.9% of all genes (99.953th and 99.995th percentiles, respectively), suggesting a very strong enrichment in AR and ERG binding within the PDE4D gene locus (see Supplementary Figure 8a). However, since PDE4D is a comparably large gene and spans approximately 1.5 Mb of genomic space, we repeated this analysis using more than three million randomly sampled genomic regions of 1.5 Mb size across all major chromosomes. Again, we found that PDE4D was highly enriched in AR and ERG binding peaks (95.151th and 87.624th percentiles, respectively) compared to random genomic stretches of comparable size (Supplementary Figure 8b). As a whole, these data support the observed expression profiles and suggest an involvement of both AR and ERG in overall PDE4D isoform regulation.

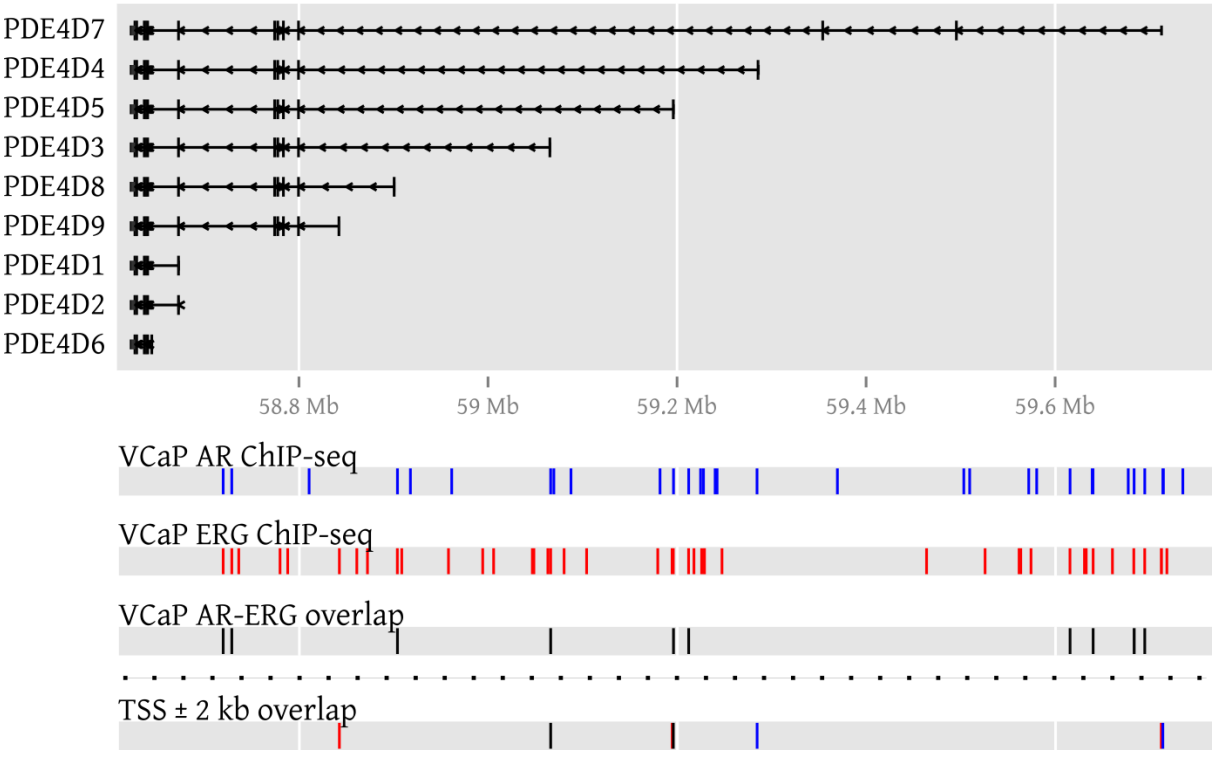


Figure 5: AR and ERG binding peaks in PDE4D in the VCaP cell line. To visualize AR and ERG binding in PDE4D, genomic locations of ChIP-seq peaks (GSE14092) denoting AR binding sites are coloured blue, while ERG peaks are coloured in red. If peaks of both transcription factors overlap, the affected genomic regions are coloured in black. A genomic region surrounding each transcription start site (TSS) is used to highlight binding events that could influence transcription.

DNA methylation of defined regions in PDE4D is altered in prostate cancer

To further study transcriptional regulation of the *PDE4D* locus, we obtained public data of DNA methylation in PCa patients from Gene Expression Omnibus (GEO) and TCGA and performed statistical analyses to identify hyper- and hypo-methylated regions in PCa as compared to normal adjacent prostate (NAP). The results of three different platforms determining DNA methylation patterns consistently detected hyper-methylated regions, indicating active silencing of several PDE4D promoters in PCa, involving the transcription start site (TSS) of a total of five PDE4D isoforms, namely the short PDE4D1/2 isoforms and the long PDE4D4, PDE4D5 and PDE4D8 isoforms (see Supplementary Figure 9).
To estimate the impact of these differentially methylated regions (DMRs) on isoform expression, we used Affymetrix Human Exon Array samples obtained from the same patients as the MeDIP-seq cohort (30, 31) and calculated Spearman's correlation coefficient for each of the differentially methylated regions (DMRs) and the associated PDE4D isoform. Of the five T regions involved, PDE4D5 showed the strongest negative association ($\rho = -0.571$, Supplementary Table 3), while the four DMRs near the PDE4D4 TSS showed varying agreement between methylation and expression measurements, ranging from $\rho = -0.215$ to $\rho = -0.394$. These results follow the expected behaviour, as increased DNA methylation impedes transcription (32). Since the PDE4D1 and PDE4D2 expression could not be independently measured with the Exon Arrays, a negative correlation ($\rho = -0.517$) was found for both. Lastly, PDE4D8 expression did not show any association with DNA methylation ($\rho = -0.233$), agreeing with our observation that this isoform is not consistently expressed in prostate tissues (see Supplementary Figure 1).

PDE4D isoforms can be used as diagnostic and prognostic signature for prostate cancer: application to prostate biopsies

Since PDE4D7 and PDE4D5 show opposing behaviours in prostatic tissues, we created a diagnostic signature based on the expression of PDE4D7 relative to that of PDE4D5 expression ($PDE4D7 - PDE4D5$). In order to evaluate its performance in distinguishing PCa and non-PCa samples, we carried out ROC analyses in all compatible datasets and compared the resulting AUCs with PCA3 (Supplementary Table 4). Overall, our diagnostic signature performed on par with PCA3, with AUCs ranging from 0.839 to 0.934 compared to 0.857 to 0.921.
In order to evaluate the value of PDE4D as a clinical biomarker, we used surgical resection materials of eighteen patients and subjected them to needle biopsies to obtain material from distinct areas, simulating both true positive and false negative biopsies (see Supplementary Table 5). In total, four biopsies with gradually increasing distance from the tumour were taken per patient (within tumour, edge of tumour, 5 mm from edge, and 10 mm from edge) and PDE4D5 as well as PDE4D7 expression were measured by qPCR. Ct values of both isoforms were normalized to several reference genes (see Methods) and adjusted to baseline expression in NAP tissue (10 mm from edge). Both expression profiles showed inverse correlation, with PDE4D5 expression decreasing in the vicinity of the tumour, while PDE4D7 expression as well as the diagnostic signature gradually increasing (see Figure 6), confirming our earlier

findings. Additionally, a transient change of PDE4D isoform expression at the tumour edge might suggest that nearby adjacent normal tissue is influenced by tumour presence through a 'field effect', but could also be due to averaging signals from normal and cancerous cells. Notably, expression of all long PDE4D isoforms including PDE4D5 and PDE4D7 appears to decrease during PCa progression (see Figure 1 and Supplementary Figures 2-3), while expression of the super-short PDE4D isoforms PDE4D1 and PDE4D2 seemed to be affected to a lesser extent. On this basis, we decided to create a prognostic signature based on the expression level of PDE4D1/2 relative to the sum of the expression levels of the long PDE4D5, PDE4D7 and PDE4D9 isoforms ($PDE4D1/2 - (PDE4D5 + PDE4D7 + PDE4D9)$). The performance of this signature was then evaluated in the Exon Array cohorts. Since, three datasets had appropriate follow-up available, we used clinical recurrence (CR) defined as development of metastases after RP as clinical endpoint. Overall, our signature performed well in distinguishing patients with CR from those without, yielding AUCs of 0.826, 0.794 and 0.614 for the EMC, Taylor and Erho cohort, respectively (Supplementary Table 4). Since the EMC dataset offered time to biochemical recurrence (BCR), metastases-free as well as overall survival time as follow-up information, we performed a Kaplan-Meier analysis for this dataset using our prognostic PDE4D signature. Two categories (signature high and low) were defined by Partitioning Around Medoids (PAM) and left-censoring was applied, resulting in well separated curves for both metastases-free and overall survival ($p < 0.05$, see Figure 6). Subsequently, we used Cox proportional hazards regression model to evaluate whether our PDE4D signature is an independent predictor for clinical metastasis, BCR and overall survival, taking into account the pre-operational PSA, Gleason score, pathological stage, surgical margins and patient age. For both metastases-free as well as overall survival, the prognostic PDE4D signature was found to be an independent predictor ($p < 0.1$), though confidence intervals were large due to low numbers of samples and events (Supplementary Table 6).

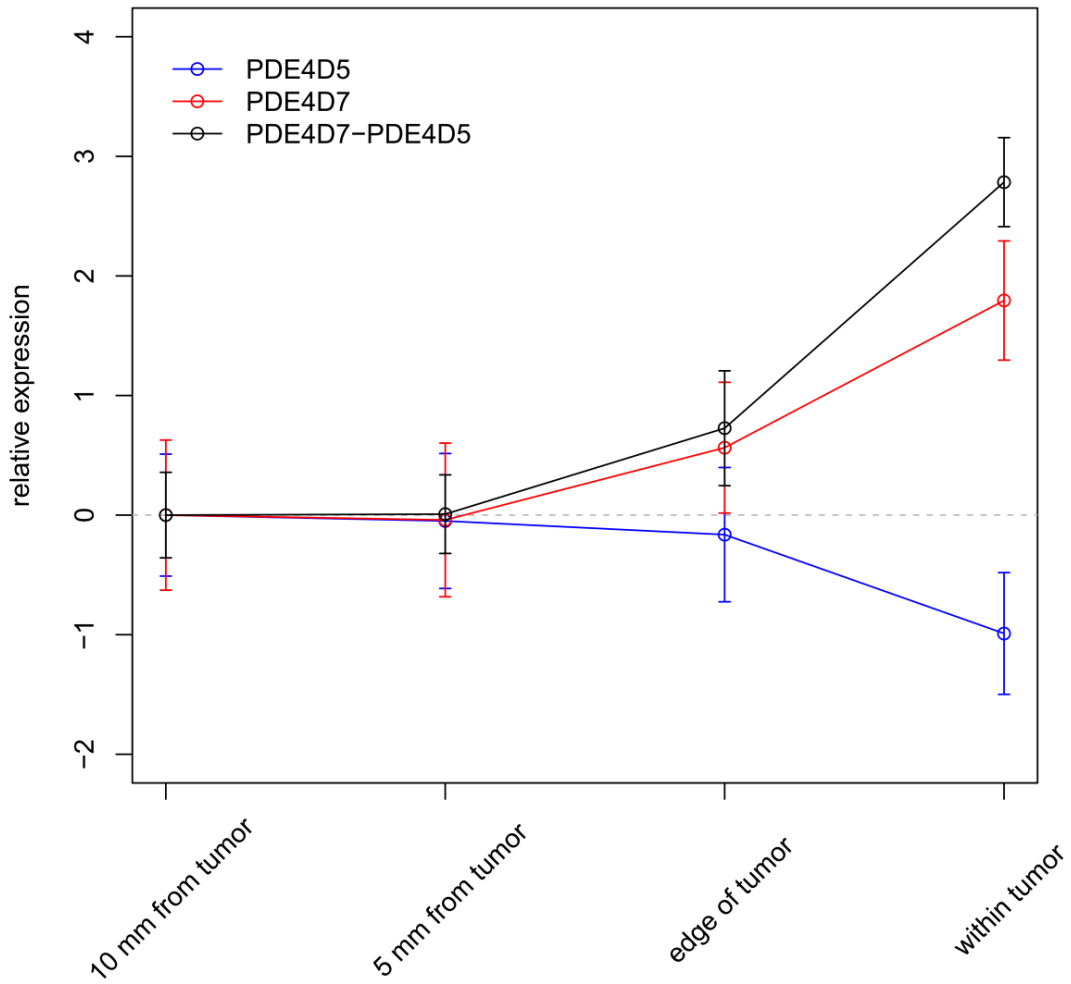


Figure 6: Applying the diagnostic PDE4D signature in needle biopsies. Expression of PDE4D5 and PDE4D7 in relation to distance to the tumour as measured by qRT-PCR in prostate tumour biopsies (n = 18). Error bars represent standard error of the mean.

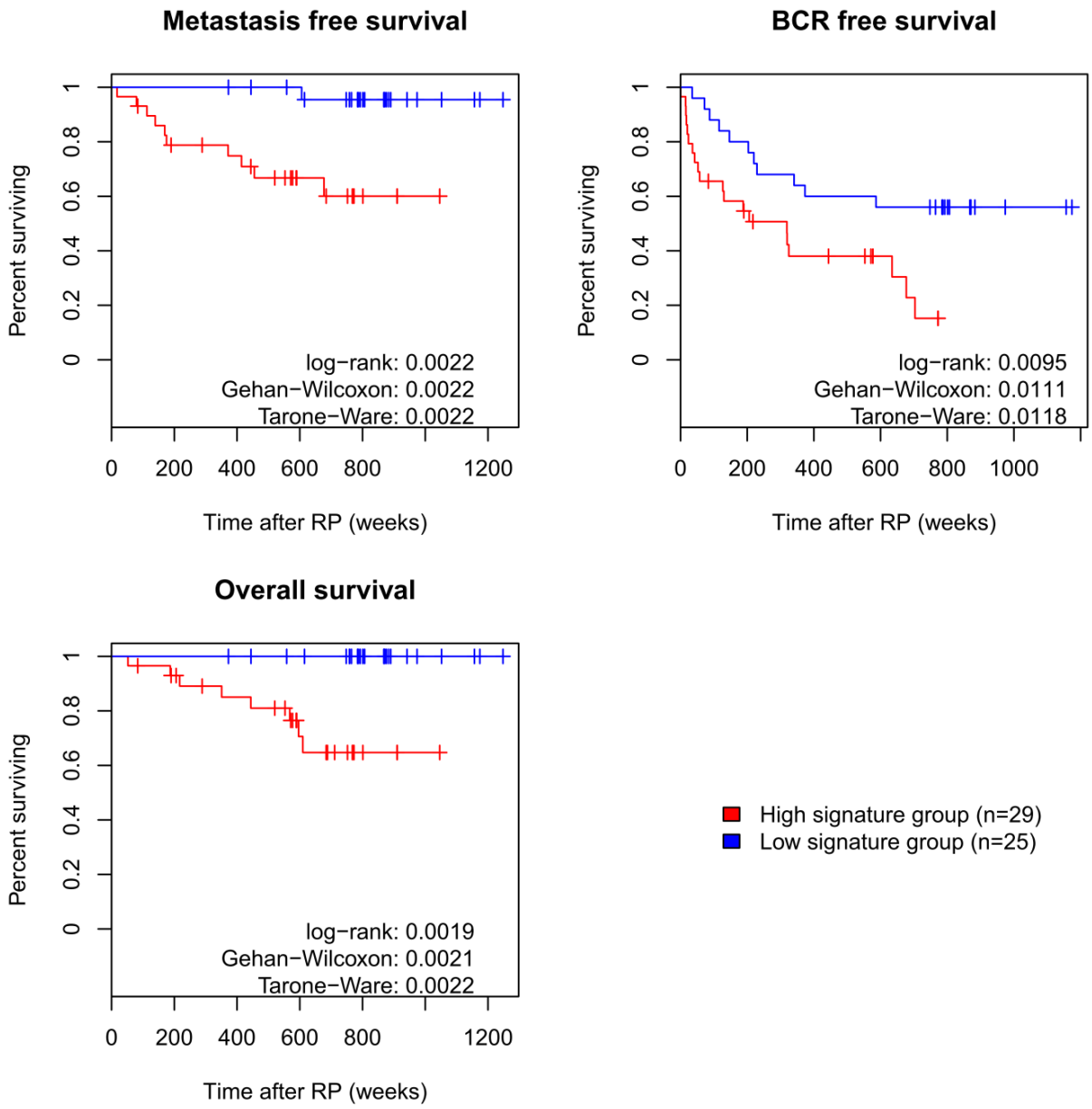


Figure 7: Survival analysis for prognostic PDE4D signature. Using the prognostic PDE4D signature to distinguish between outcomes, Kaplan Meier curves for three clinical endpoints were created based on the EMC dataset. Assignment of samples to the high and low signature group was performed by clustering of samples according to their signature values using Partitioning Around Medoids (PAM).

Discussion

Our investigation of the transcriptional dynamics of the *PDE4D* gene locus revealed a previously undescribed promoter switch involving the major contributors of PDE4D activity in normal prostate, namely the PDE4D5 and PDE4D9 long forms, as well as the prostate cancer-associated long isoform PDE4D7 (15, 16). Unique promoters for each PDE4D isoform, located upstream of the exon(s) encoding their unique N-terminal regions allow for the independent regulation of the different mRNA and corresponding protein expression (6, 9, 33). Here in this study we provide the first evidence of condition-specific PDE4D promoter switching in a cancer context.

Isoform switching in various genes (such as PKM, CXCR3 and FGR2 (34–36)) during cancer development has been described in several cancer types including prostate cancer (37–40), and likewise tumour-specific isoforms of known genes have been identified previously (41). Indeed, the androgen receptor variant 7 (AR-V7) provides a particularly important example of a PCa-specific isoform that is constitutively active and ligand-independent, contributing to castration resistance of prostate cancer cells (42, 43). Furthermore, alternative promoter usage of the androgen-regulated gene TMPRSS2 as part of the TMPRSS2-ERG fusion gene also has been associated with clinical outcome (44, 45).

Interestingly, mounting evidence suggests crosstalk between AR and cAMP signalling pathways, with important cAMP downstream targets such as PKA and ERK interacting either with the AR or AR target genes (4, 46–48). PDEs, in providing the sole route for degrading cAMP are poised to play a key regulatory role, particularly so as the targeting of particular isoforms to distinct signalling complexes confers a spatial aspect that allows particular isoforms to have specific functional roles (6). Therefore, it is particularly intriguing to find that specific PDE4D isoforms expressed in prostatic tissues appear to be androgen regulated (PDE4D7 and PDE4D5), suggesting a complex network of interactions that links both pathways. We should, however, mention that studies of PDE4D7 expression in the VCaP prostate cancer cell line implied that it was not directly regulated by AR (15). However, VCaP harbours genomic rearrangements on chr5q that are characteristic of chromothripsis, and more importantly, PDE4D is reportedly involved in gene fusions with FAM172A and C5orf47 (49). With regards to the AR-induced up-regulation of PDE4D7 observed in LNCaP cells, these structural rearrangements in VCaP could be involved in a loss of AR-mediated regulation of PDE4D7 due to relocation or deletion of regulatory elements such as AR binding elements. An alternative explanation could be that PDE4D7 expression is indirectly linked to AR activity, as its promoter region overlaps PART1, a known AR target gene that showed clear association with androgen treatment in VCaP (15, 50, 51). In the Exon Array datasets that we analysed, both genes seem to be co-expressed in prostatic tissues (mean Spearman's rho = 0.7269). However, given the fact that PDE4D5 was significantly down-regulated in LNCaP upon AR stimulation as well, we believe that AR directly influences PDE4D isoform expression through interaction with proximal or distal regulatory elements (52, 53). This hypothesis is supported by the ChIP-seq data for AR, which identifies numerous binding peaks for AR in the PDE4D gene locus, including the PDE4D5 and PDE4D7/PART1 promoter regions.

Similarly, ERG seems to have a major contribution on PDE4D isoform expression, with us previously reporting that PDE4D7 is up-regulated in TMPRSS2-ERG positive PCa (16). Here, we provided further ChIP-seq support for ERG involvement in PDE4D expression. However, the Exon Array datasets analysed here do not provide conclusive evidence for a link of ERG overexpression with other isoforms, such as PDE4D5. It appears therefore plausible that ERG overexpression may be specifically linked to PDE4D7 expression, highlighting a connection of the latter to the AR pathway, as well as its potential oncogenic role (15).

To investigate whether DNA methylation could be involved in the promoter switch uncovered in this study, we analysed three independent datasets based on different technologies, whereupon we discovered consistent increases of DNA methylation near the PDE4D5 TSS in PCa samples. In conjunction with the observed AR-mediated down-regulation of PDE4D5, these results could well explain the profound down-regulation of PDE4D5 in localized and advanced PCa and could hint at a protective function in normal prostate that is inhibited by gene silencing in PCa. In addition, we found increased DNA methylation near the PDE4D1/2 TSS that could not be linked to significantly altered gene expression, while other isoforms showing differential methylation (PDE4D4, PDE4D8) do not seem to be consistently expressed in prostatic tissues. Indeed, it is even possible that the increased DNA methylation in the promoter regions of specific PDE4D isoforms might induce promoter switching to PDE4D7 by inhibiting expression of other PDE4D isoforms.

Unlike PDE4D5, PDE4D9 does not show signs of androgen regulation despite being down-regulated in PCa and we could not find evidence for DNA methylation-mediated regulation of PDE4D9 expression in PCa. Thus, its transcriptional regulation in PCa remains unclear at this point and solicits further study.

Taken together, the observed switch in isoform usage might imply that regulatory mechanisms of PDE4D-catalyzed cAMP degradation are subjected to AR signalling in PCa cells that, in turn, indicates that PDE4D7-specific protein domains are necessary to regulate cAMP signalling in an androgen-dependent manner, offering a potentially new drug target (15, 16, 18). Moreover, with the transition to an androgen-independent state, expression of long PDE4D isoforms seems to fade, reaching its minimum in castration-resistant conditions and distant metastases, while expression of the super-short isoforms PDE4D1 and PDE4D2 appears to remain rather stable. Importantly, these super-short isoforms contain the catalytic domain of PDE4D but lack the UCR1/UCR2 domains seen in long PDE4D isoforms, a module that confers regulation by various kinases and influences intracellular targeting (6, 18).

Hence, this effective loss of regulation of PDE4D activity can be expected to generate profound changes in compartmentalized cAMP signaling due to altered spatial localization and cross-talk governing cAMP degradation, and may thereby contribute to cancer aggressiveness similarly to mechanisms suggested for MAPKs (54) and AR in form of its splice variant AR-V7 (43).

PDE4D isoform composition appears to have merit in being used as a diagnostic signature following the expression of both PDE4D7 and PDE4D5, as well as serving as a prognostic signature following the difference between the expression of long and short PDE4D isoforms. Evaluating both signatures, we found that they exhibited good performance in distinguishing PCa from normal tissue and progressive from non-progressive samples, respectively.

Importantly, diagnostic performance was robust to differences in technology, data processing, as well as potential differences in composition and patient characteristics of the used cohorts, demonstrating a high cross-platform reproducibility of PDE4D isoforms as PCa biomarker and yielding results comparable to the established PCa-marker PCA3 in all tested cohorts. Hence, with further optimization to an appropriate test platform prior to clinical utilization, we could imagine that such signatures might provide a valuable addition to complement existing test procedures. When applying our diagnostic signature to prostate biopsies, PDE4D isoform expression appeared to return to its 'normal' state with increasing distance from the tumour, whereas the tumour edge showed an intermediate signal. This observation could hint at a 'field effect' of the tumour on and/or crosstalk of the tumour cells with the surrounding microenvironment (55–58). It would therefore be fascinating to further explore in the future whether such a 'field effect' indeed influences PDE4D isoform composition, effectively increasing the target area for biopsies, or whether our observations were caused by averaging signals from adjacent tumour and normal cells. If validated, an increased target area could boost accuracy of prostate biopsies, reducing the number of false negative tests. Furthermore, it would be highly interesting to see whether reversing the isoform composition to its normal state has an influence on prostate cell phenotype and behaviour.

While our study focused on PDE4D isoform expression in primary PCa samples, genomic alterations of the PDE4D locus such as microdeletions have been observed in other cancers (27). Moreover, a recent study found that mutations in other members of the PDE family could be related to PCa by affecting intracellular cAMP and/or cGMP levels (59). Considering the large number of PDE genes and isoforms as well as the tight regulation of cAMP signalling and its degradation, it is very well possible that PDEs such as PDE4D are key players in other conditions, as the broad panel of associated diseases underscores (10–14). Therefore, it is worthwhile to extend the presented study and screen the expression profiles of all known PDEs in various tissues and conditions to define basal expression levels and reveal potential alterations and novel targets for drug interventions.

Taken together, our findings highlight the potential of PDE4D isoforms to be promising new biomarkers and potential therapeutic targets for localized and advanced prostate cancer.

Materials and Methods

Analysis of PDE4D isoform expression in prostate tissues

Quantification of PDE4D isoforms in patient materials was performed by qRT-PCR as described in (16). In addition, six independent Exon Array datasets were used in this study and raw CEL files were obtained via Gene Expression Omnibus (GEO) or personal communication. The datasets comprised GSE21034 (25), GSE29079 (30), GSE46691 (60), GSE32875 (28) as well as patient samples from GSE41410 (61, 62) and samples published in (63). These datasets are referred to as 'Taylor', 'Brase', 'Erho', 'Rajan', 'EMC', and 'Nijmegen', respectively.
Of note, patients *PCA0041*, *PCA0042* and *PCA0119* of the Taylor dataset were marked as 'treated with salvage radical prostatectomy (RP)', meaning they previously failed radiotherapy treatment and were subsequently treated with RP. Therefore, Exon Array expression data for *PCA0119* were not used for survival analysis.
Raw data were processed and RMA normalized using the aroma.affymetrix R-package ((64), CDF used: HuEx-1_0-stv2,extendedR3,A20071112,EP.CDF, see http://www.aroma-project.org/). Expression of transcript isoforms was measured by using log2-transformed intensity values of isoform-specific probesets: PDE4D1/2 (2858166); PDE4D3 (2858290, 2858291); PDE4D4 (2858368, 2858369, 2858370); PDE4D5 (2858345, 2858346, 2858347); PDE4D6 (2858155, 2858156); PDE4D7 (2858406, 2858407, 2858408); PDE4D8 (2858257, 2858258); PDE4D9 (2858240, 2858241). These intensity values were normalized to a set of reference genes (HPRT1, PUM1, TBP, POLR2A, TUBA1B) by using the mean intensity of 'core' probesets of each gene's transcript cluster (3991698, 2404254, 2937984, 3453732, 3708704) to estimate gene expression and then using the average reference gene expression as normalization factor. This normalization factor was subtracted from the probeset intensity values, and normalized probeset expression was subsequently averaged per PDE4D isoform. In addition, expression of the PCa associated genes was normalized the same way as PDE4D, using 'core' and 'extended' probesets of transcript cluster 3175538 to measure PCA3 as well as 3931765 for ERG and 2811145 for PART1.
Lastly, level 3 processed RNA-seq expression values for PRAD samples were obtained from TCGA (https://tcga-data.nci.nih.gov/tcga/) via the TCGA-Assembler R-package (65). For each sample, the RSEM 'scaled estimate' values were used and multiplied by 10^6 to convert the values to transcripts per million (TPM). Error bars in plots represent standard deviation unless stated otherwise.

Analysis of deletions of PDE4D and impact on isoform expression

Gene-level copy number alterations were obtained from TCGA via the TCGA-Assembler R-package (65) and a cut-off of ±log2(1.5/2) was used to call gains and losses of genetic material, respectively. A Wilcoxon-Mann-Whitney test was used to identify significant changes in expression of PDE4D isoforms between samples with and without alterations.

Evaluation of AR and ERG expression / binding on PDE4D transcription

To determine the (TMPRSS2-)ERG status of patient samples in Exon Array cohorts, we used relative ERG expression values and applied Partitioning Around Medoids (PAM, R-package 'cluster', k = 2) to assign the patient samples to the ERG positive or negative group based on expression. Lastly, a Wilcoxon-Mann-Whitney-test was used to detect statistically significant differences ($p < 0.05$) between the ERG positive and ERG negative samples. Likewise, differences between R1881 treated and untreated LNCaP cells (28) were tested using a Wilcoxon-Mann-Whitney-test. To investigate transcription factor binding, public ChIP-seq peaks for AR and ERG were obtained from GEO (GSE14092) and overlapped with PDE4D TSS ± 2 kb regions using bedtools (66) after conversion to hg19 coordinates using the liftOver executable (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Distances of the nearest AR and ERG peaks to each PDE4D isoform TSS were calculated by 'bedtools closest' using the options '-k 5 and -d'. Data visualization was based on the ggBio R-package (67). Enrichment of AR and ERG peaks in the PDE4D gene locus was investigated by counting the number of ChIP-seq peaks of each transcription factor within 21,209 RefSeq gene loci (hg19) as well as randomly sampled genomic regions of 1.5 Mb. Unique gene loci were defined by the minimum and maximum chromosomal coordinates of RefSeq NM and NR transcripts belonging to the same gene identifier after associating them to HGNC gene symbols using biomaRt (68) and excluding minor chromosomes and haplotypes. For each chromosome, random regions were sampled according to: $number\ of\ regions = (\text{chromosome size in Mb} * 1000)$ and any regions overlapping the PDE4D gene locus were excluded. Counting was performed by bedtools (66) *annotate* using the option '-counts' and empirical cumulative distribution functions for both transcription factors were created by using the ecdf() function of R-package stats. Hexbinplots were generated using the BoutrosLab.plotting.general R-package (http://labs.oicr.on.ca/boutros-lab/software/bpg).

Investigation of PDE4D promoter methylation

Public methylation data were downloaded from GEO and TCGA data portal and comprised three different technologies. 1) Deduplicated and extended MeDIP-seq reads (200 nt) deposited under accession number GSE35342 (31) were downloaded from Gene Expression Omnibus (GEO) and processed via the MEDIPS R-package (69). Using genomic bins of 100 nt for chromosome 5, reads were counted for every sample and differential methylation status of each bin was tested using the following MEDIPS settings as suggested by the authors upon request: 'diff.method = "edgeR", prob.method = "poisson", MeDIP = F, CNV = F'. Bins covering the genomic region of PDE4D including 50 kb flanks and with a Bonferroni-adjusted p-value below 0.01 were selected and merged into larger regions of interest (ROIs) if they were directly adjacent. 2) Pre-processed public bisulfite sequencing (BiS-seq) data available from GEO (GSE41701, (70)) were downloaded, and measured positions found in the genomic region of PDE4D including 50 kb flanks were extracted. For each position, the percentage of reads indicating methylation was calculated by #base calls C / (#base calls C + #base calls T) based on the number of reads covering a particular base. Next, the limma R-

package (71, 72) was used to identify positions with significant differences in methylation between PCa vs. benign, as well as CRPC vs. PCa. Positions with FDR < 0.05 were selected and merged into larger regions if they were within 100 nts of each other. 3) TCGA level 3 data for Illumina Infinium HumanMethylation450 BeadChips were downloaded from TCGA data portal and only patients with available clinical information were used for further analysis. Pre-calculated beta values for chromosome 5 were imported into Minfi (73) and annotated using 'ilmn12.hg19'. Analysis of differential methylation was performed via bumphunter using 100 permutations and 'cutoff=0.15'. Lastly, any significant probes located within the genomic region of PDE4D including 50 kb flanks were extracted and methylation profiles were correlated to RNA expression via Spearman's correlation coefficient. Visualisation of methylated regions was performed using ggBio (67).

Analysis of signature performance, survival and independent predictor variable

We created a diagnostic signature based on PDE4D7 expression relative to PDE4D5 expression (PDE4D7-PDE4D5) as well as a prognostic signature for the Exon Array cohorts based on PDE4D1/2 relative to PDE4D5, PDE4D7 and PDE4D9 ((PDE4D1/2) - (PDE4D5+PDE4D7+PDE4D9)). Subsequently, the R-packages 'ROC' and 'survival' were used to carry out ROC analyses and perform a Cox regression as well as Kaplan-Meier analysis based on available survival data of the EMC dataset (61, 62).

Quantification of diagnostic PDE4D signature in prostate biopsies

Several biopsy punches (approximately 1 x 2 mm) were taken in a representative tumour area after surgical prostate resections in eighteen different men with prostate cancer. Experimental protocols were approved by the Erasmus MC Medical Ethics Committee following the Medical Research Involving Human Subjects Act. For each patient, these punches were performed within the tumour, at the edge of the tumour area, at 5 and at 10 mm distance to the tumour region. RNA was extracted and qRT-PCRs (quantitative real-time PCR) for PDE4D5 and PDE4D7 were performed as described in (16), using ACTB, HPRT1, TUBA1B, POLR2A, PUM1 and TBP as reference genes. The expression of PDE4D5 and PDE4D7 in each biopsy tissue was normalized as follows: mean(Ct(reference genes)) – Ct(PDE4DX). For each of the eighteen different patients, the normalized expression of PDE4D transcripts within the tumour was set to 1 and expression values for biopsies taken at various distances from the tumour were calculated relative to the expression in the tumour. Lastly, average relative expression and standard error of the mean of PDE4D transcript expression were plotted for each of the respective biopsy locations.

Acknowledgements

This study was supported by the framework of CTMM (The Netherlands), the Center for Translational Molecular Medicine, PCMM project (grant 03O-203) and NGS ProToCol (grant 03O-402). We would like to acknowledge support from the Biotechnology and Biological Sciences Research Council (UK) to GSB (BB/G01647X/1). We would like to thank Dr. Elena Martens-Uzunova (Erasmus MC, Rotterdam) for insightful scientific discussions and NovioGendix B.V (Nijmegen) for providing us with one of the used Exon Array datasets. The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Conflicts of interest

The authors R Böttcher, K Dulla, , G J L H van Leenders, G S Baillie, and G Jenster declare no financial interest and no conflict of interests in the presented work. D v Strijp and R Hoffmann are employees of Philips Research Eindhoven, which partly funded this work. The authors D v Strijp, M D Houslay, and R Hoffmann are inventors of patent applications related to the field of this work.

References

1. Ferlay,J., Steliarova-Foucher,E., Lortet-Tieulent,J., Rosso,S., Coebergh,J.W.W., Comber,H., Forman,D. and Bray,F. (2013) Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur. J. Cancer*, **49**, 1374–403.

2. Schröder,F.H., Hugosson,J., Roobol,M.J., Tammela,T.L.J., Ciatto,S., Nelen,V., Kwiatkowski,M., Lujan,M., Lilja,H., Zappa,M., *et al.* (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.

3. Andriole,G.L., Crawford,E.D., Grubb,R.L., Buys,S.S., Chia,D., Church,T.R., Fouad,M.N., Gelmann,E.P., Kvale,P.A., Reding,D.J., *et al.* (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.

4. Merkle,D. and Hoffmann,R. (2011) Roles of cAMP and cAMP-dependent protein kinase in the progression of prostate cancer: Cross-talk with the androgen receptor. *Cell. Signal.*, **23**, 507–515.

5. Conti,M. and Beavo,J. (2007) Biochemistry and physiology of cyclic nucleotide phosphodiesterases: essential components in cyclic nucleotide signaling. *Annu. Rev. Biochem.*, **76**, 481–511.

6. Houslay,M.D. (2010) Underpinning compartmentalised cAMP signalling through targeted cAMP breakdown. *Trends Biochem. Sci.*, **35**, 91–100.

7. Lugnier,C. (2006) Cyclic nucleotide phosphodiesterase (PDE) superfamily: A new target for the development of specific therapeutic agents. *Pharmacol. Ther.*, **109**, 366–398.

8. Francis,S.H., Houslay,M.D. and Conti,M. (2011) Phosphodiesterase inhibitors: Factors that influence potency, selectivity, and action. *Handb. Exp. Pharmacol.*, **204**, 47–84.

9. Rahrmann,E.P., Collier,L.S., Knutson,T.P., Doyal,M.E., Kuslak,S.L., Green,L.E., Malinowski,R.L., Roethe,L., Akagi,K., Waknitz,M., *et al.* (2009) Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping beauty transposon-based somatic mutagenesis screen. *Cancer Res.*, **69**, 4388–4397.

10. Kaname,T., Ki,C.-S., Niikawa,N., Baillie,G.S., Day,J.P., Yamamura,K.-I., Ohta,T., Nishimura,G., Mastuura,N., Kim,O.-H., *et al.* (2014) Heterozygous mutations in cyclic AMP phosphodiesterase-4D (PDE4D) and protein kinase A (PKA) provide new insights into the molecular pathology of acrodysostosis. *Cell. Signal.*, **26**, 2446–59.

11. Michot,C., Le Goff,C., Goldenberg,A., Abhyankar,A., Klein,C., Kinning,E., Guerrot,A.M., Flahaut,P., Duncombe,A., Baujat,G., *et al.* (2012) Exome sequencing identifies PDE4D mutations as another cause of acrodysostosis. *Am. J. Hum. Genet.*, **90**, 740–745.

12. Lee,H., Graham,J.M., Rimoin,D.L., Lachman,R.S., Krejci,P., Tompson,S.W., Nelson,S.F., Krakow,D. and Cohn,D.H. (2012) Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am. J. Hum. Genet.*, **90**, 746–751.

13. Yoon,H.-K., Hu,H.-J., Rhee,C.-K., Shin,S.-H., Oh,Y.-M., Lee,S.-D., Jung,S.-H., Yim,S.-H., Kim,T.-M. and Chung,Y.-J. (2014) Polymorphisms in PDE4D are associated with a risk of COPD in non-emphysematous Koreans. *COPD*, **11**, 652–8.

14. Gretarsdottir,S., Thorleifsson,G., Reynisdottir,S.T., Manolescu,A., Jonsdottir,S., Jonsdottir,T., Gudmundsdottir,T., Bjarnadottir,S.M., Einarsson,O.B., Gudjonsdottir,H.M., *et al.* (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.*, **35**, 131–138.

15. Henderson,D.J.P., Byrne,A., Dulla,K., Jenster,G., Hoffmann,R., Baillie,G.S. and Houslay,M.D. (2014) The cAMP phosphodiesterase-4D7 (PDE4D7) is downregulated in androgen-independent prostate cancer cells and mediates proliferation by compartmentalising cAMP at the plasma membrane of VCaP prostate cancer cells. *Br. J. Cancer*, **110**, 1278–87.

16. Böttcher,R., Henderson,D.J.P., Dulla,K., van Strijp,D., Waanders,L.F., Tevz,G., Lehman,M.L., Merkle,D., van Leenders,G.J.L.H., Baillie,G.S., *et al.* (2015) Human phosphodiesterase 4D7 (PDE4D7) expression is increased in TMPRSS2-ERG-positive primary prostate cancer and independently adds to a reduced risk of post-surgical disease progression. *Br. J. Cancer*, **113**, 1502–1511.

17. Wang,D., Deng,C., Bugaj-Gaweda,B., Kwan,M., Gunwaldsen,C., Leonard,C., Xin,X., Hu,Y., Unterbeck,A. and De Vivo,M. (2003) Cloning and characterization of novel PDE4D isoforms PDE4D6 and PDE4D7. *Cell. Signal.*, **15**, 883–91.

18. Byrne,A.M., Elliott,C., Hoffmann,R. and Baillie,G.S. (2015) The activity of cAMP-phosphodiesterase 4D7 (PDE4D7) is regulated by protein kinase A-dependent phosphorylation within its unique N-terminus. *FEBS Lett.*, **589**, 750–5.

19. Hoffmann,R., Wilkinson,I.R., McCallum,J.F., Engels,P. and Houslay,M.D. (1998) cAMP-specific phosphodiesterase HSPDE4D3 mutants which mimic activation and changes in rolipram inhibition triggered by protein kinase A phosphorylation of Ser-54: generation of a molecular model. *Biochem. J.*, **333**, 139–49.

20. Hoffmann,R., Baillie,G.S., MacKenzie,S.J., Yarwood,S.J. and Houslay,M.D. (1999) The MAP kinase ERK2 inhibits the cyclic AMP-specific phosphodiesterase HSPDE4D3 by phosphorylating it at Ser579. *EMBO J.*, **18**, 893–903.

21. MacKenzie,K.F., Wallace,D.A., Hill,E. V, Anthony,D.F., Henderson,D.J.P., Houslay,D.M., Arthur,J.S.C., Baillie,G.S. and Houslay,M.D. (2011) Phosphorylation of cAMP-specific PDE4A5 (phosphodiesterase-4A5) by MK2 (MAPKAPK2) attenuates its activation through protein kinase A phosphorylation. *Biochem. J.*, **435**, 755–69.

22. Plattner,F., Hayashi,K., Hernández,A., Benavides,D.R., Tassin,T.C., Tan,C., Day,J., Fina,M.W., Yuen,E.Y., Yan,Z., *et al.* (2015) The role of ventral striatal cAMP signaling in stress-induced behaviors. *Nat. Neurosci.*, **18**, 1094–100.

23. Sheppard,C.L., Lee,L.C.Y., Hill,E. V, Henderson,D.J.P., Anthony,D.F., Houslay,D.M., Yalla,K.C., Cairns,L.S., Dunlop,A.J., Baillie,G.S., *et al.* (2014) Mitotic activation of the DISC1-inducible cyclic AMP phosphodiesterase-4D9 (PDE4D9), through multi-site phosphorylation, influences cell cycle progression. *Cell. Signal.*, **26**, 1958–74.

24. Boutros,P.C., Fraser,M., Harding,N.J., de Borja,R., Trudel,D., Lalonde,E., Meng,A., Hennings-Yeomans,P.H., McPherson,A., Sabelnykova,V.Y., *et al.* (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.*, **47**, 736–45.

25. Taylor,B.S., Schultz,N., Hieronymus,H., Gopalan,A., Xiao,Y., Carver,B.S., Arora,V.K., Kaushik,P., Cerami,E., Reva,B., *et al.* (2010) Integrative genomic profiling of human prostate

cancer. *Cancer Cell*, **18**, 11–22.

26. Baca,S.C., Prandi,D., Lawrence,M.S., Mosquera,J.M., Romanel,A., Drier,Y., Park,K., Kitabayashi,N., MacDonald,T.Y., Ghandi,M., *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666–77.

27. Lin,D.-C., Xu,L., Ding,L.-W., Sharma,A., Liu,L.-Z., Yang,H., Tan,P., Vadgama,J., Karlan,B.Y., Lester,J., *et al.* (2013) Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. *Proc. Natl. Acad. Sci.*, **110**, 6109–6114.

28. Rajan,P., Dalgliesh,C., Carling,P.J., Buist,T., Zhang,C., Grellscheid,S.N., Armstrong,K., Stockley,J., Simillion,C., Gaughan,L., *et al.* (2011) Identification of novel androgen-regulated pathways and mRNA isoforms through genome-wide exon-specific profiling of the LNCaP transcriptome. *PLoS One*, **6**, e29088.

29. Yu,J., Yu,J., Mani,R.S., Cao,Q., Brenner,C.J., Cao,X., Wang,X., Wu,L., Li,J., Hu,M., *et al.* (2010) An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression. *Cancer Cell*, **17**, 443–454.

30. Brase,J.C., Johannes,M., Mannsperger,H., Fälth,M., Metzger,J., Kacprzyk,L.A., Andrasiuk,T., Gade,S., Meister,M., Sirma,H., *et al.* (2011) TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. *BMC Cancer*, **11**, 507.

31. Börno,S.T., Fischer,A., Kerick,M., Fälth,M., Laible,M., Brase,J.C., Kuner,R., Dahl,A., Grimm,C., Sayanjali,B., *et al.* (2012) Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov.*, **2**, 1024–35.

32. Baylin,S.B. (2005) DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.*, **2 Suppl 1**, S4–11.

33. Richter,W., Jin,S.-L.C. and Conti,M. (2005) Splice variants of the cyclic nucleotide phosphodiesterase PDE4D are differentially expressed and regulated in rat tissue. *Biochem. J.*, **388**, 803–11.

34. Desai,S., Ding,M., Wang,B., Lu,Z., Zhao,Q., Shaw,K., Yung,W.K.A., Weinstein,J.N., Tan,M. and Yao,J. (2014) Tissue-specific isoform switch and DNA hypomethylation of the pyruvate kinase PKM gene in human cancers. *Oncotarget*, **5**, 8202–10.

35. Wu,Q., Dhir,R. and Wells,A. (2012) Altered CXCR3 isoform expression regulates prostate cancer cell migration and invasion. *Mol. Cancer*, **11**, 3.

36. Zhao,Q., Caballero,O.L., Davis,I.D., Jonasch,E., Tamboli,P., Yung,W.K.A., Weinstein,J.N., Strausberg,R.L. and Yao,J. (2013) Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. *Clin. Cancer Res.*, **19**, 2460–72.

37. Eswaran,J., Horvath,A., Godbole,S., Reddy,S.D., Mudvari,P., Ohshiro,K., Cyanam,D., Nair,S., Fuqua,S.A.W., Polyak,K., *et al.* (2013) RNA sequencing of cancer reveals novel splicing alterations. *Sci. Rep.*, **3**, 1689.

38. Sebestyen,E., Zawisza,M. and Eyras,E. (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.*, **43**, 1345–1356.

39. Hamdollah Zadeh,M.A., Amin,E.M., Hoareau-Aveilla,C., Domingo,E., Symonds,K.E., Ye,X., Heesom,K.J., Salmon,A., D'Silva,O., Betteridge,K.B., *et al.* (2015) Alternative splicing of TIA-1 in human colon cancer regulates VEGF isoform expression, angiogenesis, tumour growth and bevacizumab resistance. *Mol. Oncol.*, **9**, 167–78.

40. Noordzij,M.A., van Steenbrugge,G.J., Verkaik,N.S., Schröder,F.H. and van der Kwast,T.H. (1997) The prognostic value of CD44 isoforms in prostate cancer patients treated by radical prostatectomy. *Clin. Cancer Res.*, **3**, 805–15.

41. Barrett,C.L., DeBoever,C., Jepsen,K., Saenz,C.C., Carson,D.A. and Frazer,K.A. (2015) Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E3050–7.

42. Hu,R., Dunn,T.A., Wei,S., Isharwal,S., Veltri,R.W., Humphreys,E., Han,M., Partin,A.W., Vessella,R.L., Isaacs,W.B., *et al.* (2009) Ligand-independent androgen receptor variants derived from splicing of cryptic exons signify hormone-refractory prostate cancer. *Cancer Res.*, **69**, 16–22.

43. Qu,Y., Dai,B., Ye,D., Kong,Y., Chang,K., Jia,Z., Yang,X., Zhang,H., Zhu,Y. and Shi,G. (2015) Constitutively active AR-V7 plays an essential role in the development and progression of castration-resistant prostate cancer. *Sci. Rep.*, **5**, 7654.

44. Hermans,K.G., Boormans,J.L., Gasi,D., van Leenders,G.J.H.L., Jenster,G., Verhagen,P.C.M.S. and Trapman,J. (2009) Overexpression of prostate-specific TMPRSS2(exon 0)-ERG fusion transcripts corresponds with favorable prognosis of prostate cancer. *Clin. Cancer Res.*, **15**, 6398–403.

45. Boormans,J.L., Porkka,K., Visakorpi,T. and Trapman,J. (2011) Confirmation of the association of TMPRSS2(exon 0):ERG expression and a favorable prognosis of primary prostate cancer. *Eur. Urol.*, **60**, 183–4.

46. Sarwar,M., Sandberg,S., Abrahamsson,P.-A. and Persson,J.L. (2014) Protein kinase A (PKA) pathway is functionally linked to androgen receptor (AR) in the progression of prostate cancer. *Urol. Oncol.*, **32**, 25.e1–12.

47. Carey,A.-M., Pramanik,R., Nicholson,L.J., Dew,T.K., Martin,F.L., Muir,G.H. and Morris,J.D.H. (2007) Ras-MEK-ERK signaling cascade regulates androgen receptor element-inducible gene transcription and DNA synthesis in prostate cancer cells. *Int. J. Cancer*, **121**, 520–7.

48. Peterziel,H., Mink,S., Schonert,A., Becker,M., Klocker,H. and Cato,A.C. (1999) Rapid signalling by androgen receptor in prostate cancer cells. *Oncogene*, **18**, 6322–9.

49. Teles Alves,I., Hiltemann,S., Hartjes,T., van der Spek,P., Stubbs,A., Trapman,J. and Jenster,G. (2013) Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. *Hum. Genet.*, 10.1007/s00439-013-1308-1.

50. Lin,B., White,J.T., Ferguson,C., Bumgarner,R., Friedman,C., Trask,B., Ellis,W., Lange,P., Hood,L. and Nelson,P.S. (2000) PART-1: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12. *Cancer Res.*, **60**, 858–63.

51. Sidiropoulos,M., Chang,A., Jung,K. and Diamandis,E.P. (2001) Expression and regulation of prostate androgen regulated transcript-1 (PART-1) and identification of differential expression in prostatic cancer. *Br. J. Cancer*, **85**, 393–7.

52. Makkonen,H., Kauhanen,M., Paakinaho,V., Jaaskelainen,T. and Palvimo,J.J. (2009) Long-range activation of FKBP51 transcription by the androgen receptor via distal intronic enhancers. *Nucleic Acids Res.*, **37**, 4135–4148.

53. Wu,D., Zhang,C., Shen,Y., Nephew,K.P. and Wang,Q. (2011) Androgen receptor-driven chromatin looping in prostate cancer. *Trends Endocrinol. Metab.*, **22**, 474–480.

54. Dhanasekaran,D.N. and Johnson,G.L. (2007) MAPKs: function, regulation, role in cancer and therapeutic targeting. *Oncogene*, **26**, 3097–9.

55. Chandran,U.R., Dhir,R., Ma,C., Michalopoulos,G., Becich,M. and Gilbertson,J. (2005) Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. *BMC Cancer*, **5**, 45.

56. Chen,F., Zhuang,X., Lin,L., Yu,P., Wang,Y., Shi,Y., Hu,G. and Sun,Y. (2015) New horizons in tumor microenvironment biology: challenges and opportunities. *BMC Med.*, **13**, 45.

57. Ma,X.-J., Dahiya,S., Richardson,E., Erlander,M. and Sgroi,D.C. (2009) Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.*, **11**, R7.

58. Jia,Z., Wang,Y., Sawyers,A., Yao,H., Rahmatpanah,F., Xia,X.-Q., Xu,Q., Pio,R., Turan,T., Koziol,J.A., *et al.* (2011) Diagnosis of prostate cancer using differentially expressed genes in stroma. *Cancer Res.*, **71**, 2476–87.

59. de Alexandre,R.B., Horvath,A.D., Szarek,E., Manning,A.D., Leal,L.F., Kardauke,F., Epstein,J.A., Carraro,D.M., Soares,F.A., Apanasovich,T. V, *et al.* (2015) Phosphodiesterase sequence variants may predispose to prostate cancer. *Endocr. Relat. Cancer*, **22**, 519–30.

60. Erho,N., Crisan,A., Vergara,I.A., Mitra,A.P., Ghadessi,M., Buerki,C., Bergstralh,E.J., Kollmeyer,T., Fink,S., Haddad,Z., *et al.* (2013) Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One*, **8**, e66855.

61. Boormans,J.L., Korsten,H., Ziel-van der Made,A.J.C., van Leenders,G.J.L.H., de Vos,C. V, Jenster,G. and Trapman,J. (2013) Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. *Int. J. Cancer*, **133**, 335–45.

62. Böttcher,R., Hoogland,A.M., Dits,N., Verhoef,E.I., Kweldam,C., Waranecki,P., Bangma,C.H., van Leenders,G.J.L.H. and Jenster,G. (2015) Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. *Oncotarget*.

63. Leyten,G.H.J.M., Hessels,D., Smit,F.P., Jannink,S.A., de Jong,H., Melchers,W.J.G., Cornel,E.B., de Reijke,T.M., Vergunst,H., Kil,P., *et al.* (2015) Identification of a Candidate Gene Panel for the Early Diagnosis of Prostate Cancer. *Clin. Cancer Res.*, **21**, 3061–70.

64. Purdom,E., Simpson,K.M., Robinson,M.D., Conboy,J.G., Lapuk,A. V and Speed,T.P. (2008) FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, **24**, 1707–14.

65. Zhu,Y., Qiu,P. and Ji,Y. (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods*, **11**, 599–600.

66. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.

67. Yin,T., Cook,D. and Lawrence,M. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.*, **13**, R77.

68. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–40.

69. Lienhard,M., Grimm,C., Morkel,M., Herwig,R. and Chavez,L. (2014) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, **30**, 284–6.

70. Lin,P.-C., Giannopoulou,E.G., Park,K., Mosquera,J.M., Sboner,A., Tewari,A.K., Garraway,L.A., Beltran,H., Rubin,M.A. and Elemento,O. (2013) Epigenomic alterations in localized and advanced prostate cancer. *Neoplasia*, **15**, 373–83.

71. Smyth,G.K. (2005) Limma : Linear Models for Microarray Data. *Bioinformatics*, **pages**, 397–420.

72. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 10.1093/nar/gkv007.

73. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–9.

Chapter 6

Using *a priori* knowledge to align sequencing reads to their exact genomic position

René Böttcher[1,2], Ronny Amberg[2], F. P. Ruzius[3], V. Guryev[3], Wim F. J. Verhaegh[1], Peter Beyerlein[1,2] and P. J. van der Zaag[1]

1 Philips Research Laboratories, High Tech Campus 11, 5656 AE Eindhoven, The Netherlands
2 University of Applied Sciences Wildau, Bahnhofstraße, 15475 Wildau, Germany
3 Hubrecht Institute and University Medical Center Utrecht, KNAW, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands

Published in

Nucleic Acids Research. 2012. 40(16):e125

Abstract

The use of *a priori* knowledge in the alignment of targeted sequencing data is investigated using computational experiments. Adapting a Needleman–Wunsch algorithm to incorporate the genomic position information from the targeted capture, we demonstrate that alignment can be done to just the target region of interest. When in addition use is made of direct string comparison, an improvement of up to a factor of 8 in alignment speed compared to the fastest conventional aligner (Bowtie) is obtained. This results in a total alignment time in targeted sequencing of around 7 min for aligning approximately 56 million captured reads. For conventional aligners such as Bowtie, BWA or MAQ, alignment to just the target region is not feasible as experiments show that this leads to an additional 88% SNP calls, the vast majority of which are false positives (~ 92%).

Introduction

Since the introduction of so-called next-generation sequencing in 2005, developments in the field of DNA sequencing proceed at a very rapid pace (1). Initially, in the newer sequencing technologies based on massively parallel sequencing (2), the time required to complete a sequencing study was around three weeks, equally divided among sample preparation, the actual sequencing and the bioinformatics analysis. New sequencing technologies are emerging, which promise to reduce the actual sequencing time from the present one week to much shorter. Ultimately, nanopore-based sequencing methods may reduce sequencing run time to matters of seconds (3). Hence, it would be desirable to speed up also the time required in the sample preparation as well as the bioinformatics analysis.

Sequence alignment is a challenge in biology since the first DNA sequences have been determined in the 1970s, with the earliest approaches utilizing dot plots to compute the optimal alignment of the sequences (4). Because of their complexity, dot plots were replaced by the dynamic programming (DP) approach developed by Bellman and Viterbi, first implemented for biological use by Needleman and Wunsch (5, 6). Since then, the Needleman–Wunsch algorithm has been modified several times to adapt it to other problems and to improve its performance (7, 8). Nevertheless, DP requires too much computation time and space to handle the increasing amount of sequencing data. Therefore, heuristic approaches for searching sequence databases such as BLAST and FASTA were developed to overcome this problem (9, 10). Though these programs and their successors are still commonly used, the upcoming of next-generation sequencing requires new software (11) to process the immense amount of short reads created, which lead to the development of hash table based aligners, as for example ELAND and MAQ (12, 13). Since then, considerable further effort has been made to reduce the alignment time. One of the most successful ones is the implementation of a Burrows– Wheeler transform to index the genome and speed up the alignment (14). Common examples of aligners utilizing the Burrows–Wheeler transform are Bowtie and BWA (15, 16).

In many branches of electronic data processing the use of *a priori* information is a proven method to improve data analysis. Thus far such an approach has not been adopted in the field of DNA sequencing, although it is conceivable that information arising from so-called targeted sequencing (17–19) could be used to this effect. Typically in targeted sequencing using on-array hybridization (17, 18), the fragments of the DNA sample are hybridized to a microarray with probes designed to capture the fragments of interest. After washing away any non-bound fragments, the DNA fragments of interest for the biological or clinical question at hand are eluted from the array and are further processed to be sequenced. In current practice the resulting eluate is a random mixture of the captured DNA fragments. Moreover, the subsequent alignment of the sequencing reads is done to the whole genome as, at the current specificity of the enrichment methods, aligning to just the target region introduces an unacceptably high error rate, as we will show. In targeted sequencing, one in principle can retain the capture probe information of the micro-genomic selection array, for instance by conducting the sequencing step directly on the capture spot (20) or by using labeled capture

beads. Specifically, the very recently proposed oligonucleotide-selective sequencing (OS-Seq) by Myllykangas et al. (20) enables this approach. In this method of targeted resequencing target-specific oligonucleotides are used to create 'primer-probes'. These primer-probes are immobilized on the surface of a flow cell and serve both as capture probes and sequencing primers i.e. after capturing the complementary targets from the library, these primer probes are extended. Subsequently, bridge PCR cluster formation is performed. These clusters can be sequenced twice to determine the captured target and subsequently the OS-Seq primer probe sequence (20). This enables the identification of the exact OS-Seq primer that mediated the targeting. Myllykangas et al. (20) have used this approach to facilitate the assessment of the performance of individual primer probes.

Here, we would like to investigate the potential benefit of this approach to improve the speed of sequence alignment. To do so we have performed computational experiments to investigate what benefit such an approach of using *a priori* information might bring to sequence alignment and to see whether this can reduce the still sizeable part of the time needed to perform DNA analysis. This investigation has been done by computer-generating a set of sequencing reads that contain the *a priori* known genomic position of their capture probes. These reads are then aligned with an implementation of the Needleman–Wunsch algorithm that uses the *a priori* information to map only to the corresponding sequence fragment. The required alignment time is compared to the time needed by a number of state-of-the-art aligners, which do not use this prior knowledge and which align to the whole genome. Although one could argue that conventional aligners would also be speeded up by aligning only to the target region, we will first show that this is not a viable option by analysis of real enrichment sequencing data, as this yields many false positive SNP calls.

Methods

Evaluation of the error introduced by alignment to just the target region by conventional aligners

In targeted sequencing, capture arrays are used to reduce the total amount of bases to be sequenced. This reduction is achieved by capturing only the sequences of interest, known as target region. Since enrichment methods do not have a specificity of 100% but typically of around 70% (17, 18, 21), a considerable amount of off-target reads are generated. Consequently, data from targeted sequencing are aligned to the whole genome, using aligners such as Bowtie, BWA or MAQ, and not just to

the target region. To evaluate the error introduced by aligning only to the target region, data (50 bp reads) from a previously published study (21) were used. The sequencing reads were aligned against the whole genome as well as to the target regions (including 100bp flanks) to evaluate the errors introduced. Subsequently, SNP calling was performed using filtering with the following criteria:

(1) Positions with lower than 20× and higher than 2000× coverage were excluded.
(2) Bases with quality below 10 were excluded from SNP calling.
(3) No more than five reads that have identical mapping position and strand were included.
(4) Each of the non-reference alleles has to be supported by reads mapping to the forward as well as by reads mapping to the reverse strand of the reference genome.
(5) The non-reference allele should be observed in 20% or more reads covering the polymorphic position.
(6) Sites with more than four alleles were excluded as representing positions with increased error rate.

Positions that passed this filtering were called as candidate SNPs (or small indels).

Including *a priori* knowledge in sequence alignment

As the capture probes of hybridization arrays are designed to catch specific sequences, their position on the genome must be known in advance. Therefore, if the location of a capture probe on the array as well as its position on the genome are known, the corresponding sequencing read of the captured fragment can be associated with the sequence of its expected mapping position within the target region, provided that this information is retained during the sequencing process. Hence, the read can be aligned against this associated 'reference sequence' instead of the whole genome.

To computer-generate reads containing information about the genomic position of their capture sequence and their associated reference sequence, first several different target sequences on the genome were selected to construct a target region of interest (Figure 1). For each of these target sequences, a number of capture probes is assumed that would be present on a hybridization array and act as primers for sequencing. Therefore, the genomic position of

a sequencing read as well as its associated reference sequence is located behind the capture probe. To cover the complete target sequence with sequencing reads, the capture probes need to be shifted along the genome, which results in the reference sequences being shifted as well to form a tiling of the target sequence with a constant offset (Figure 1A). Taking the reference sequences as templates, we next introduced errors, SNPs and Indels to simulate the sequencing reads (Figure 1B). The resulting reads were used as input for the computations to determine the speed performance of our approach compared with a regular alignment. The regular alignment against the whole genome was performed with Bowtie, BWA and MAQ (Figure 1C). For the alignment using the position information, different implementations of the Needleman–Wunsch algorithm were used (Figure 1D). These consist of a regular Needleman–Wunsch (NW) and a pruned version of the Needleman–Wunsch algorithm following the beam search paradigm (22). We refer to the latter implementation as 'banded' Needleman–Wunsch algorithm (NWB). Additionally, both algorithms were implemented using exact matching prior to the alignment to increase the computation speed (NWem and NWBem), as we describe further in the following section.

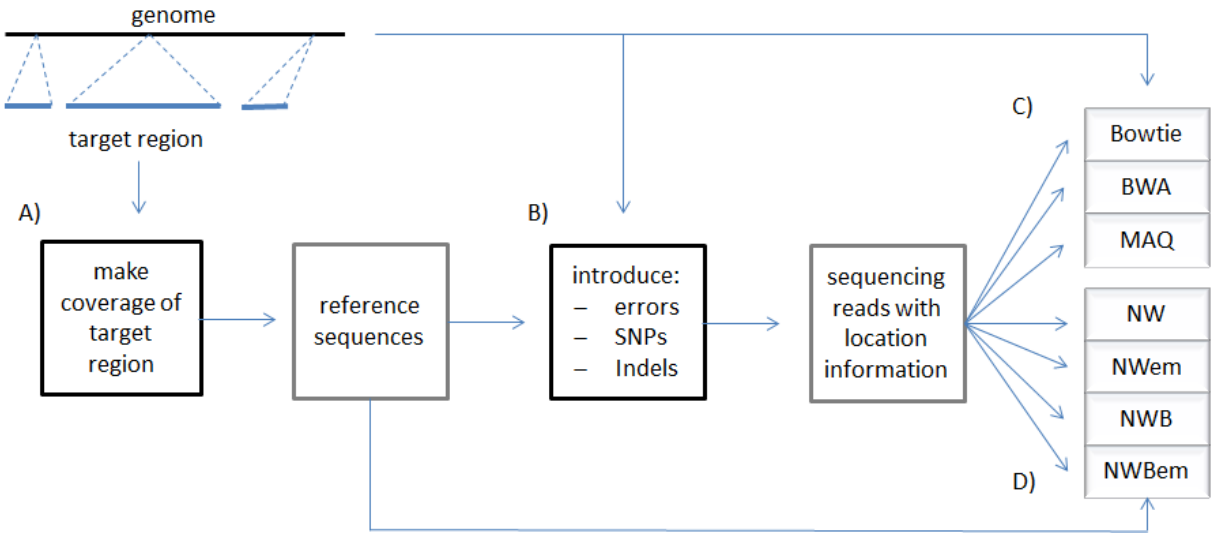


Figure 1: Overview of the workflow. (**A**) A target region was chosen from which the reference sequences were created. (**B**) Each reference sequence was then used to create the associated reads. To simulate realistic data, errors, SNPs and Indels were introduced. The resulting reads were then aligned to the whole genome (**C**) or to their associated reference sequence (**D**).

Different alignment approaches

The first implementation of alignment using position information was realized through a regular Needleman– Wunsch algorithm (NW), which aligns each read to its associated reference sequence. Since the reads are expected to be very similar to the reference sequence, we realized that a direct string comparison might be applicable to skip the alignment for exactly matching sequences. This insight led to a second implementation (NWem), which performs the alignment in two steps. First, the information included in the header of each read

is used to look up and identify the reference sequence associated to the read being processed, and subsequently the aligner checks whether the compared sequences match exactly. If so, the maximum alignment score is assigned; otherwise, a regular alignment is performed for the two sequences (as has been described in (10, 22); allowing up to two Indels for the beam search approach). Since the Needleman–Wunsch algorithm can be optimized for similar sequences, a banded version was also implemented (NWB, as described in the previous section) and exact matching was added (NWBem), which works similarly to NWem.

To compare the new approaches with established alignment methods, the reads were also aligned against the whole human reference genome using Bowtie (0.12.7), BWA (0.5.9-r16) and MAQ (0.7.1). Default settings were used for MAQ (map) and BWA (aln & samse). Bowtie was run using '-a -n 2 -q –solexa1.3-quals – quiet' settings. The calculations were executed on a grid of 1648 cores divided over 206 Dell PowerEdge M600 blade servers, each utilizing two Intel Xeon L5420 Quadcore CPUs @ 2.5Ghz with 16, 32 or 64 GB of random access memory (BiG Grid, see www.biggrid.nl).

Generation of sequencing data

The data necessary to determine the gain of the new alignment approach by comparison to the regular aligments was obtained from reference human genome GRCh37 and a recent gene annnotation (Ensembl database, release 62; http://www.ensembl.org) (23). In total, 7368 exons were chosen as the target region, representing approximately 3 million bases (Mb) based on previous microarray genomic enrichment experiments (21). Exons originating from the X and Y chromosomes as well as extrachromosomal DNA were excluded. A subset of the chosen exons was taken to create also a 300 kb target region (784 exons), while a 30Mb target region was also assembled to compare the performance for larger data sets (72 943 exons).

Figure 2 shows the principle of the data generation based on the captured sequences (dark green) which are complementary to the capture probes present for instance on a hybridization array. The capture probes would be designed in such a way that the reference sequences (light green) following the captured sequences form a tiling of the target sequence (continuous black). This target sequence is a part of the target region, and might be an exon of interest. To generate the sequencing data, each associated reference sequence was created by selecting a substring from the target sequence, while the starting base of the next reference sequence was shifted by an offset of 10 bases, covering the target sequence in the process (Figure 2A). This procedure was repeated until the remaining target sequence was too small to create a new reference sequence with the required length.

The associated reads were then created from their associated reference sequences, with a number of copies referred to as the read redundancy (Figure 2B).As indicated in red in Figure 2, sequencing errors and incorrectly captured reads were introduced into the data set. SNPs and Indels were additionally introduced to the sequencing data with probabilities corresponding to typical occurences mentioned in literature (24). After the read sequence was prepared, the assembly of the read was finished by including the genomic position information.

In the above approach, the length of the reference sequences influences the number of total reference sequences and associated reads, as with increasing length of the reference sequences, fewer complete sequences can be fitted into the target sequences, e.g. the exons chosen. As shown in Table 1, the number of reference sequences decreases for each step of 25 bases. To determine the number of sequencing reads for each combination of target region and read length, the number of reference sequences has to be multiplied by the read redundancy.

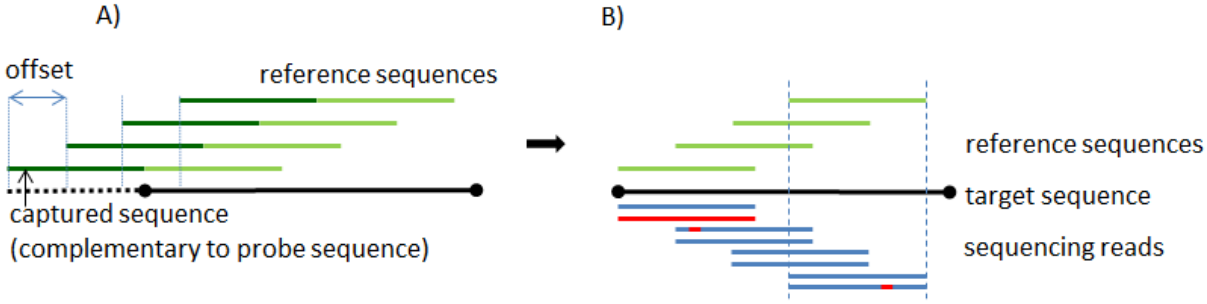


Figure 2. Principle of data generation. (**A**) Captured sequences (dark green) are complementary to the designed capture probes present on the array. These probes are designed in such a way that the following reference sequences (light green) form a tiling of the target sequence (continuous black) with a constant offset. Each reference sequence is therefore directly created from the target sequence. (**B**) For each reference sequence a number of associated reads (blue) is created, introducing different errors (red) in the process. The number of created reads per reference sequence is referred to as read redundancy (two in this example).

Table 1: Number of reference sequences for the different target regions, depending on the length of each reference sequence (as described in the section Generating of sequencing data). The decrease in number is due to the fact that fewer complete sequences cover the same target if the length of each generated sequence is increased.

Target region	25 base sequences	50 base sequences	75 base sequences	100 base sequences
0.3 Mb	28.163	26.218	24.243	22.298
3 Mb	283.042	264.616	246.202	227.776
30 Mb	2.857.844	2.676.092	2.493.129	2.311.377

Parameter space

To evaluate the influence of various parameters on the alignment time, we varied the values of five parameters:

- the size of the target region (0.3, 3 and 30 Mb),
- the length of the reads (25, 50, 75, 100 bases),
- the percentage of sequencing error per base (0.5%, 1%, 2%),
- the read redundancy (1, 2, 5, 10, 20) and
- the percentage of reads off-target but still captured and sequenced (0, 5, 10, 20, 40%).

Results and Discussion

Introduction of errors by aligning solely to the target region

As mentioned, the alignment speed of conventional aligners in targeted sequencing could perhaps be improved by aligning just to the target region instead of to the whole genome, which is the current practice (21), because this could seriously reduce the computational effort. To test whether this is a viable option, we first examined the effectiveness of sequence alignment to just the target region, using conventional aligners. Sequencing data from a previous experiment (21) was used for this study.

When using common enrichment methods, two classes of reads are generated, the first one consisting of all reads that originate inside the target region (referred to as ITR) and the second one comprising all reads that originate outside of the target region (referred to as OTR). When all these reads are aligned solely to the target region, two possible errors may occur that influence subsequent analysis (e.g. SNP calling). Firstly, OTRs that now align uniquely inside the target region are falsely classified as uniquely matching reads (UMRs) to the target, as they align at a position from which they do not originate (Type 1 error). Secondly, all reads (ITR and OTR) that align uniquely inside the target region, but would also align one or more times outside the target region [known as multiple matching reads (MMR)] and that would normally be excluded from analysis, are falsely classified as UMRs as well (Type 2 error).

We compared mapping strategies where reads were aligned to the full genome reference or only to the target. The previously published set (21) features 13.24 million mapped reads of which 8.36 million were uniquely mapped to the target region of genome reference NCBI36. Using the same analysis methods as described in (21), but mapping only against the target region, 8.48 million UMRs were obtained. From these, 0.78% were uniquely mapped to a different location (Type 1 error) and 0.83% were originally MMRs (Type 2 error) when the whole genome was used as a reference.

Subsequently, we evaluated the number of mismatches that were observed in reads that map consistently and in those that correspond to erroneous mappings. The result of this analysis is given in Figure 3. The data show that reads that erroneously map to the target region typically have several mismatches, while the vast majority of consistently mapped reads contains one or no mismatches with the target sequences. However, the distributions overlap and cannot be distinguished easily. For instance, accepting only reads with at most two mismatches to capture most of the consistently mapped reads, would still result in the inclusion of about half the erroneously mapped reads. Setting the threshold to 1 or 0 would on the other hand greatly reduce the information needed for SNP calling. Moreover, the use of a lower threshold to reduce type 1 and 2 errors is not feasible, since an analysis of the distance between SNPs (i.e. SNPs called when mapped against the full reference genome) showed that a third of all SNPs have neighboring SNPs not further than 50 bases apart (see Figure 3). Hence we conclude that allowing fewer than two mismatches per read would reduce the reliability of SNP calling for a substantial part of the exome.

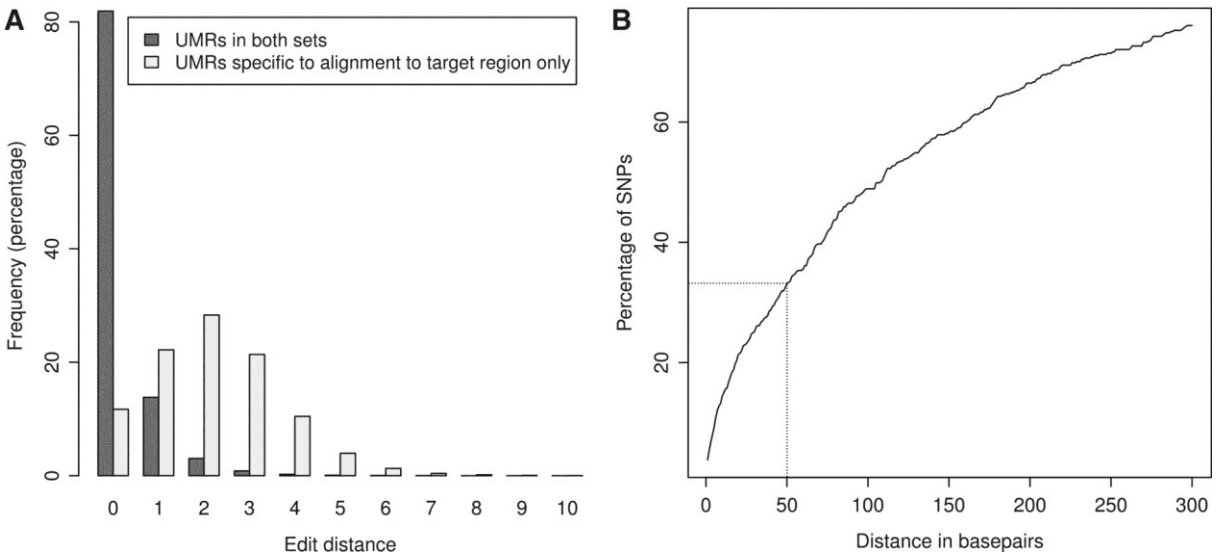


Figure 3. (**A**) Number of mismatches that were observed in reads that map consistently and in those that correspond to erroneous mappings. Reads which erroneously map to the target region typically have several mismatches, while the vast majority of consistently mapped reads have one or no mismatches with target sequences. (**B**) Distribution of distances between neighboring SNPs that map to the same target region of exome. Percentage of between-SNP ranges (Y-axis) that are below a certain distance (base pairs, X-axis) shows that one third of the between-SNP distances are 50 bp or less.

To test the effect of the additional 1.61% UMRs generated, supposedly uniquely mapping to the target region, on genomic analysis, SNP calling was performed [in the same way as done in (21)]. A direct comparison was made for sets mapped against the full genome reference and only to the target region. A total of 1886 SNPs were found in both sets, while an additional 1651 SNPs were specific to the set where mapping was done solely against the target region. Thus aligning to just the target region produces an additional 88% SNPs. The same analysis using 35 bp reads (20) yields similar results and a slightly higher overall false-positive rate (52 versus 47%), indicating that read length has an influence, but will unlikely solve the problem of mismapping. These two different SNP sets exhibit different overlap with a known SNP database: 78.8 and 8.4%, respectively (exact numbers: 1486 and 138, source Ensembl database v.54). The latter percentage implies that nearly 92% of these additionally found SNPs are false positives. In addition, both SNP sets have dissimilar distributions of percentage of non-reference calls, which are given in Figure 4. Figure 4A shows the histogram of the non-reference frequency for the overlapping SNPs in both data sets, while in Figure 4B this histogram is given for the SNPs that are unique to the mapping to the target only. The histogram in Figure 4A exhibits the expected profile with a peak at 100 (homozygous SNPs) and a secondary maximum a bit <50% expected for heterozygous SNPs. Interestingly the frequency spectrum in Figure 4B exhibits a 1/f trend with the frequency, f, which is indicative of noise (25) and suggests—in line with the low overlap with the SNPs known in Ensembl database—that nearly all of these SNPs are false positives. Therefore we conclude that, despite the small proportion of reads with ‘paralogous origin’ (1.61%) by

mapping just to the target region, they are more divergent from the target sequences and therefore can have a significant contribution to false positive SNP calls when detecting sequence variants, in an enrichment experiment when aligning just to the target region.

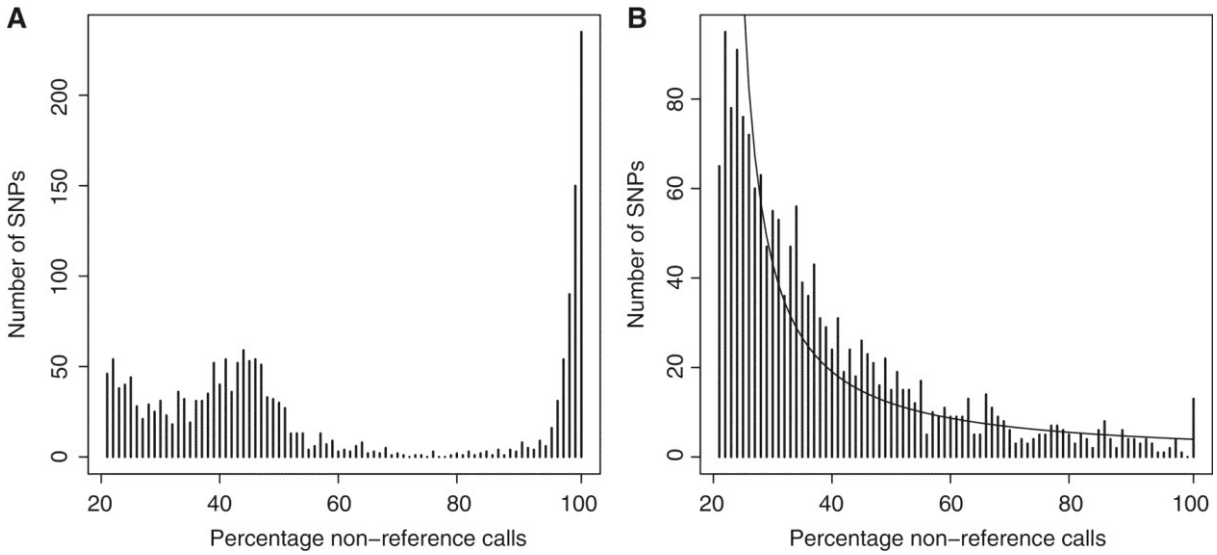


Figure 4: Distributions of percentage of non-reference calls for both SNP sets. (**A**) histogram of the non-reference call frequency for the overlapping SNPs in both data sets, (**B**) histogram for the SNPs specific to the set where read mapping was done to the target only.

Consequently, this validates the practice in targeted sequencing to perform whole genome alignment to avoid introducing additional errors during alignment. Thus, comparisons to determine the gain in alignment speed using *a priori* knowledge will be made by comparing the alignment speed of implementations of the Needleman–Wunsch algorithm, which align to just the target region, to the speed of conventional aligners (Bowtie, BWA, MAQ), which align to the whole genome.

Comparison of alignment speed

To evaluate the alignment speed of the new approach, the computation times required for aligning targeted sequencing experiments were compared to the performance of regular aligners (Bowtie, BWA and MAQ). These latter aligners do not use any *a priori* genome position information and align to the whole genome. Figure 5 shows the results of such a comparison for a 3 Mb target region, a read length of 75 bases, a sequencing error of 1% and with 10% reads off-target. These settings correspond to a total of 246,202 reference sequences. Four different implementations of the Needleman–Wunsch algorithm (NW, NWem, NWB and NWBem, see Section Different alignment approaches) were used.

As can be seen, MAQ (red) is the slowest of the aligners used in this comparison, with its computation time ranging from 8713 s up to 69,768 s depending on the read redundancy. The two Burrows–Wheeler transform-based aligners perform the same calculations much faster, requiring 661–9419 s (BWA, violet; ~6.86× faster than MAQ) and 159–2791 s (Bowtie, black; ~22.9x faster than MAQ) respectively. These results confirm previous observations

concerning the alignment speed of Burrow–Wheeler transform-based aligners (15, 16). Nevertheless, the Needleman–Wunsch algorithms using position information lead to considerably shorter alignment times.

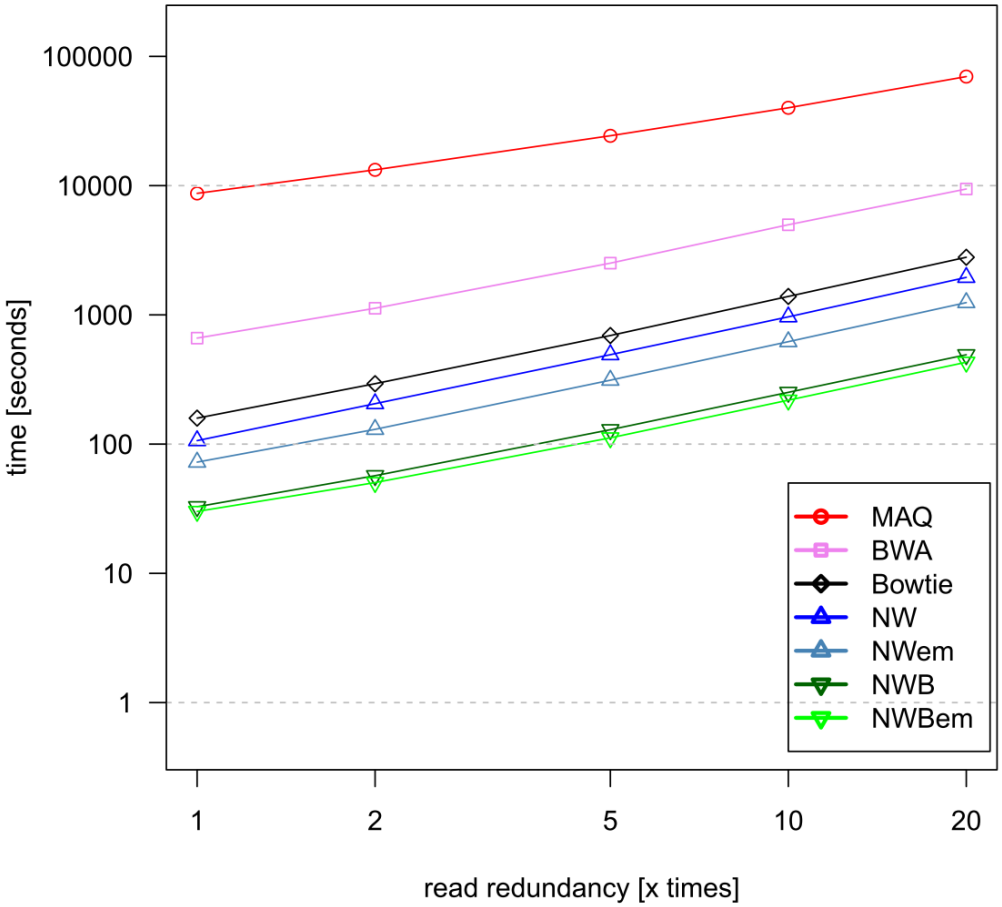


Figure 5: Comparison of the alignment speed of different aligners versus read redundancy. Bowtie, BWA and MAQ aligned against the whole genome; the Needleman–Wunsch implementations used the position information to align to the associated reference sequences. Settings: target size 30 Mb, read length 75 bases, 1% sequencing error, 10% reads off-target. Note that both axes are in logarithmic scale.

Compared to Bowtie, the computation time is decreased by a factor of ~1.4 for NW (blue; 106–1949 s), while NWem (light blue; 73–1244 s) even gains a factor of ~2.2. This gain increases further for NWB (dark green; 32–491 s or ~5.7× faster than Bowtie) and NWBem (green; 30–430 s or ~6.6× faster than Bowtie). Concluding, the total computation time for approximately 49.2 million reads of 75 bases length can be reduced from 46.5 to ~7 min when adapting a pruned Needleman–Wunsch algorithm to use the *a priori* information and comparing to the fastest regular aligner Bowtie.

Figures 6–8 show a more extensive comparison of computational experiments, regarding only two of the Needleman–Wunsch implementations (NW and NWBem) with a sequencing error of 1% per base in Figures 6 and 7, as well as 2% in Figure 8, respectively. Figure 5 is a

subplot of Figure 6 and can be found in the second row and the third column. When investigating over a broader range of conditions, Bowtie (black) shows to be the fastest of the tested common aligners, outperforming MAQ (red) and BWA (violet) in every tested parameter combination. Though the use of the position information still leads to a considerable reduction in alignment time, NW shows limitations for longer reads lengths (due to the time complexity of the regular Needleman–Wunsch algorithm being $O(max(n,m)^2)$), which are overcome by NWBem by pruning the alignment matrix.

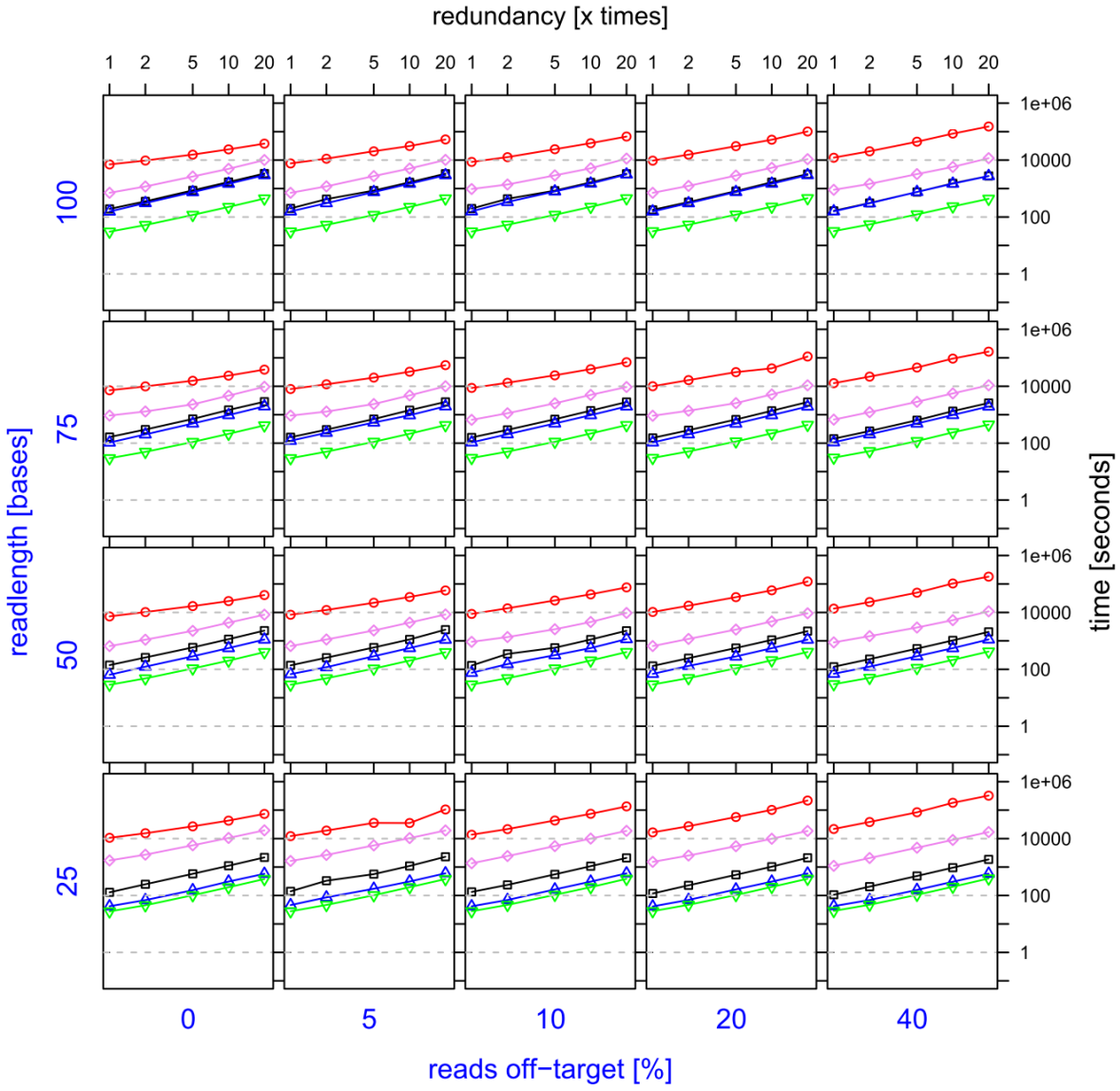


Figure 6: Comparison of different aligners for different read lengths, percentages of reads off-target and read redundancies. MAQ (red), BWA (violet) and Bowtie (black) aligned against the whole genome, NW (blue) and NWBem (green) used the position information to align to the associated reference sequence. Settings: target size 30 Mb, 1% sequencing error.

For example, in Figure 6, at a length of 100 bases and 40% reads off-target, Bowtie (164–2765 s) and NW (158–2750) compute at comparable speeds, while NWBem outperforms both (32–447 s). When considering shorter reads of 25 bases, both NW (42–583 s) and NWBem (29–396 s) are able to outperform Bowtie (106–1856 s). Concerning the amount of reads off-target, the exact matching shortcut of NWBem is skipped less often at 0% reads off-target and therefore fewer reads have to be aligned regularly (since NW performs no preselection, it is not influenced by this). Still the overall influence on computation time is only marginal, reducing alignment time to 32–445s.

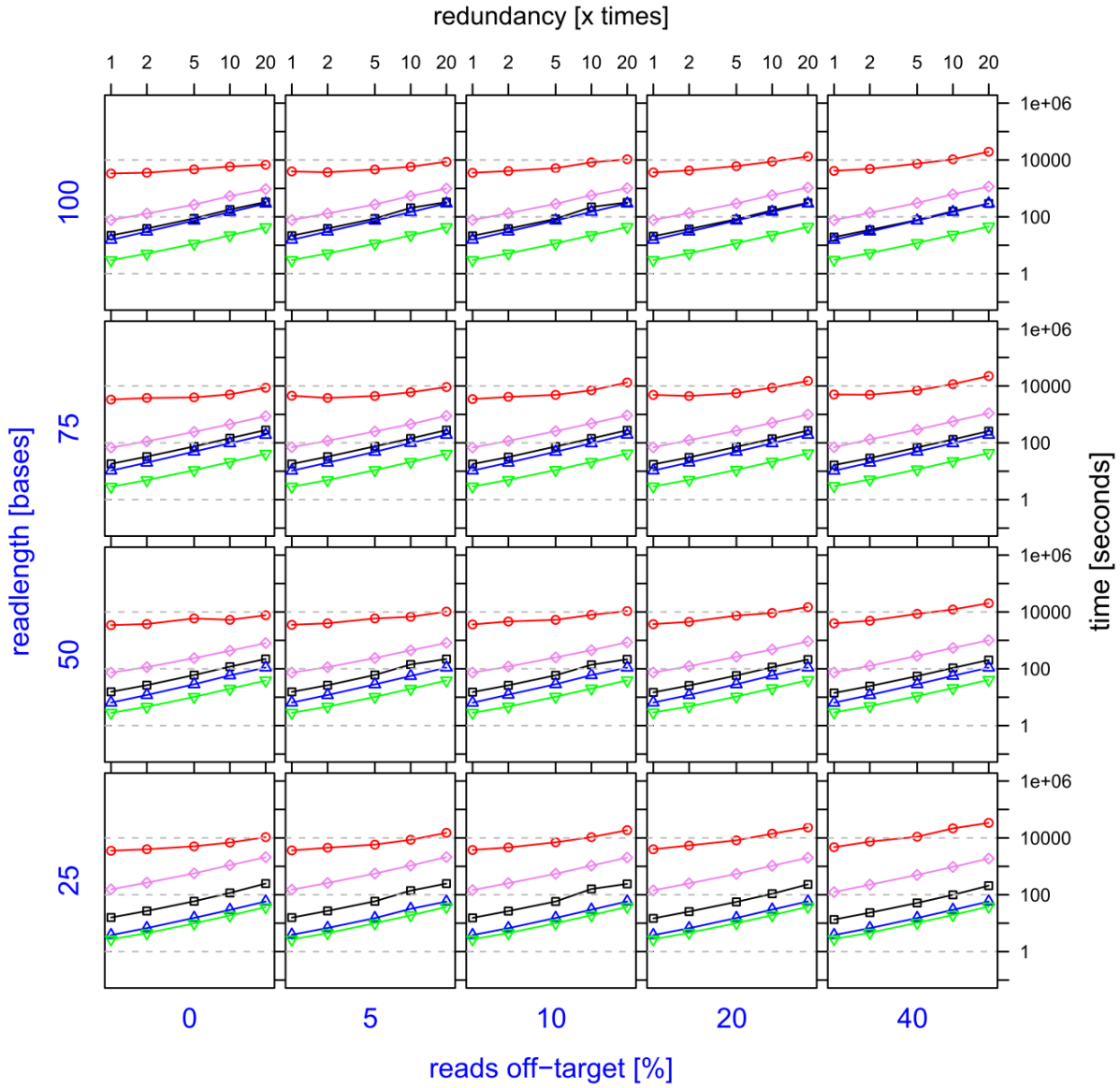


Figure 7: Comparison of different aligners for different read lengths, percentages of reads off-target and read redundancies. MAQ (red), BWA (violet) and Bowtie (black) aligned against the whole genome, NW (blue) and NWBem (green) used the position information to align to the associated reference sequence. Settings: target size 3 Mb, 1% sequencing error.

We also investigated the performance of the aligners for the 3 Mb target region (Figure 7) as well as the 300 kb target region (data not shown), which resulted in similar outcomes. In case of the 3 Mb target region, the performance gain varies between a factor of ~1.0 to ~4.3 for NW (average: 2.2±1.2) and a factor of ~5.0 to ~7.7 for NWBem (average: 6.8±0.8) when comparing to Bowtie. Similar results were observed for the 300 kb target region (NW: 2±0.9; NWBem: 6.5±1.1).

When investigating the influence of 2% sequencing error per base for the 30 Mb target region at a length of 100 bases and 40% reads off-target, the results are consistent to previous observations (Figure 8). Compared to 1% sequencing error (see Figure 6 and above), NW (158–2758 s) and NWBem (33–460 s) alignment times seem largely unchanged, while Bowtie (196–3311 s) requires ~20% more computation time. Hence, for 2% sequencing error and the 30 Mb target region, the average gain for NWBem increases to 7.8±0.8 compared to Bowtie, whereas for the 3 Mb target region it even reaches a factor of 8±0.8. Also compared to Bowtie, BWA exhibited a similar behaviour, while MAQ's performance remained stable.

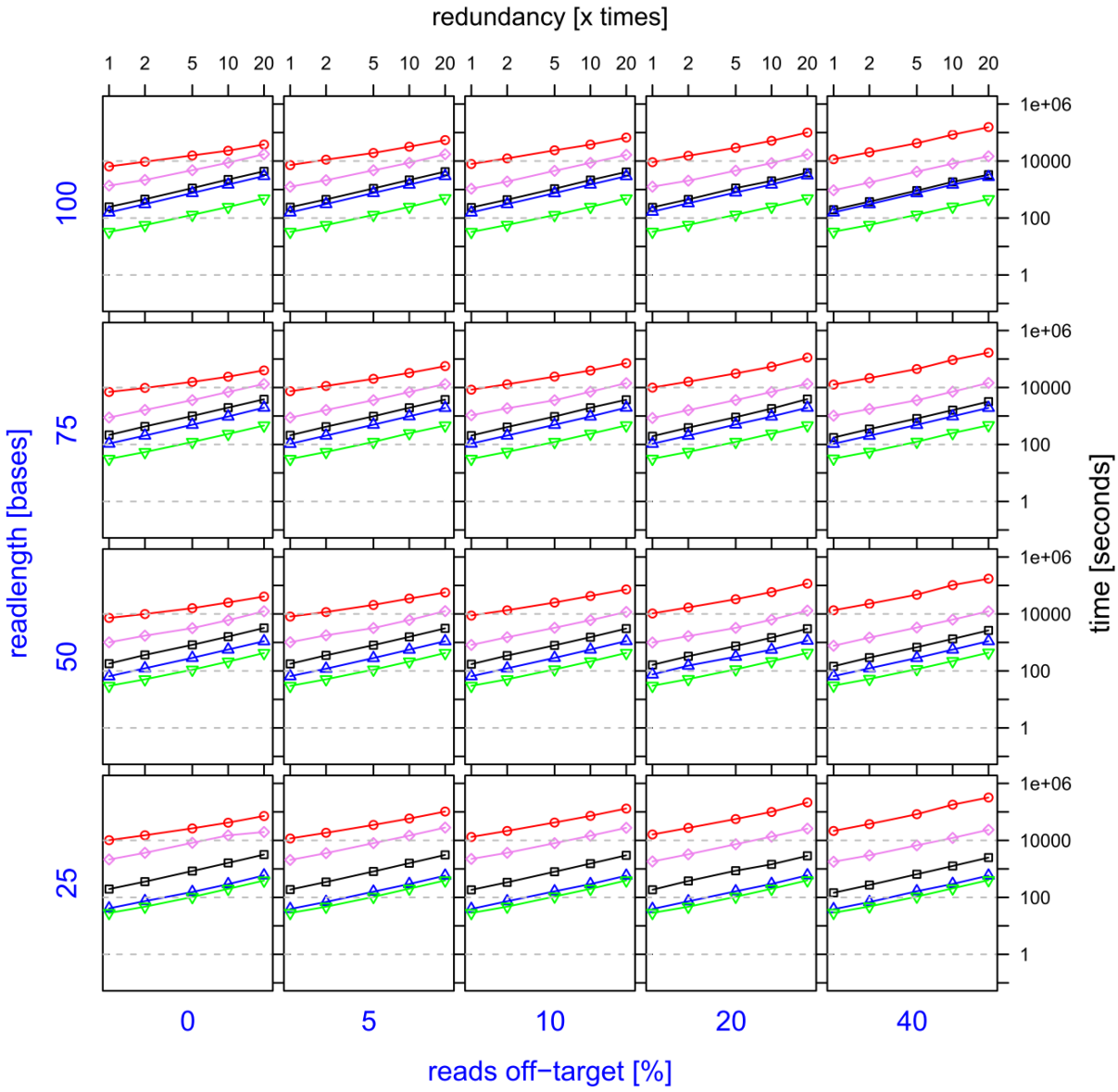


Figure 8: Comparison of different aligners for different read lengths, percentages of reads off-target and read redundancies. MAQ (red), BWA (violet) and Bowtie (black) aligned against the whole genome, NW (blue) and NWBem (green) used the position information to align to the associated reference sequence. Settings: target size 30 Mb, 2% sequencing error.

As expected, the amount of reads processed has the biggest impact on the computation time for all of the aligners, with our new approach showing a behavior similar to Bowtie and BWA. The percentage of sequencing error (in our tests up to 2%) influences the computation time of the common aligners (except for MAQ), while it has only a minor effect on the computation time of both NW and NWBem. Nevertheless, this gain in speed is sensitive to the similarity of the aligned sequences to the expected sequences, as it influences the number of exactly matching sequences. Therefore, both implementations using preselection by exact matching (NWem and NWBem) will benefit from a high specificity in enrichment and a low sequencing error.

Concerning the amount of reads off-target, Figure 6 shows that variations in the percentage influence the computation time of both implementations (NW and NWBem) only marginally, with NWBem having the performance of NWB as an upper limit for the computation time when all of the reads need to be aligned in case no exact matches are found (compare Figure 5). This can be understood as for NW, no preselection is performed and therefore all reads are aligned regardless of their origin, while for NWBem the biggest gain in computation time is achieved due to the use of the pruned Needleman–Wunsch algorithm.

Implementation aspects

To investigate whether there is room to improve NW even further, the time consumption of different parts of the Needleman–Wunsch implementations were analyzed. As shown in Table 2, I/O makes up a major part of the total computation time, up to a fraction of 83.3%. Improvements should be possible by using a binary data format instead of the text format used in this study. In summary it can be said that our approach generally benefits from short reads with high quality, as the alignment time for dynamic programming implementations increases with the length of the reads. Furthermore, high-quality reads that match perfectly do not need to be aligned at all.

We next note that BWA and Bowtie benefit from using multiple computer cores, as they can perform their computations multithreaded. MAQ as well as the presented Needleman–Wunsch aligners are not implemented in a multithreaded form (yet) and therefore did not gain from multiple cores.

Table 2: Time consumption of alignment and input/output of the NW and NWBem aligners, for different read redundancies

Program part	1×, n(%)	2×, n(%)	5×, n(%)	10×, n(%)	20×, n(%)
NW—alignment	3.8	7.92	19.07	39.93	75.75
	(60.7)	(66.4)	(68.9)	(70.84)	(69.82)
NW—I/O	2.47	3.99	8.61	16.43	32.75
	(39.3)	(33.6)	(31.1)	(29.16)	(30.18)
NWBem—alignment	0.47	0.92	2.28	4.7	9.03
	(16.71)	(19.32)	(21.83)	(23.15)	(22.7)
NWBem—I/O	2.33	3.85	8.15	15.59	30.77
	(83.29)	(80.68)	(78.17)	(76.85)	(77.3)

Furthermore, the memory requirements for the different aligners vary, making great amounts of RAM advantageous or in case of MAQ necessary for the regular aligners when aligning large numbers of reads. As shown in Table 3, NW and NWBem require only a fraction (7.5–16.6%) of the memory necessary for the other aligners to perform the calculations when aligning approximately 5 million reads from a 3Mb target region. These low hardware requirements combined with the overall speed of the computations would allow one to include the alignment within the sequencing device, making this kind of post-processing of the sequencing data obsolete in clinical applications.

Table 3: RAM requirements (MB) of the different aligners when aligning approximately 5 million reads

Aligner/algorithm	NW	NWBem	Bowtie	BWA	MAQ
Virtual memory required	200	200	1202	2333	2666
Physical memory required	145	145	904	2322	2654

Outlook

Thus far our work has been focused on methods where the enrichment step and the sequencing are combined in what can be called embedded enrichment, such as in OS-Seq (20). However, our method for mapping targeted sequences could be exploited in studies that use other enrichment strategies such as long-range PCR or selector probes (26). One could envision that the high specificity that these methods offer could warrant confining the alignment just to the target region. However, this is not done in practice to avoid generating false SNPs, as even with 98–99% specificity, 1–2% of the amplicons may be misaligned to the target region, if alignment is restricted to this (M. Nilsson, personal communication). Furthermore, as has been shown in the first results section, the vast majority of any additional SNPs generated will be false positives. PCR- and selector-based methods do not necessarily retain a direct link between a probe and the corresponding sequence read through a positional dependence. However, for the selector approach to targeted resequencing (26) a link to the

capture probe can be made as the hybridization probes are somewhere in the captured fragment to be read. If these are read as well, the read alignment could proceed by combining this information (giving the expected genomic location) and the read. In the work done by Johansson et al. (26) this was not done and alignment was performed against the full genome reference (M. Nilsson personal communication). However, if in between the two selector hybridization probes a specific label is incorporated, which upon sequencing indicates that adjacent to this site both hybridization probes are to be found, then upon the random rolling circle amplification-based multiple displacement amplification the hybridization probes can be easily found in the sequence. Consequently, the genomic location of the fragments would be known and alignment can be done just to the target location in the manner described in this article. For PCR-based enrichment methods the oligonucleotide primers, designed to flank the amplicons, could in principle also be used in the read alignment as *a priori* information. However, in this case new methods would still have to be developed to ensure that the primer information is retained through the concatamerization and/or shearing process, typically applied in the resulting next-generation sequencing library preparation as the PCR-products are longer than the currently typical read length. Thus, as the hybridization probe information can more readily be retained in the selector approach (26), in the latter target enrichment technique our method for targeted alignment might be more readily adopted.

Conclusion

In this article we have investigated the use of *a priori* information in sequence alignment, based on a new implementation of current enrichment methods for targeted sequencing. For this purpose, sequencing reads were computer generated from the human genome while varying five parameters to evaluate their impact on alignment time. The presented alignment algorithms are based on straightforward dynamic programming and use *a priori* knowledge to map each read to the expected part of the genome. These implementations prove to be faster than Bowtie, BWA and MAQ. The latter three algorithms align against the whole human genome, since alignment solely to the target region using conventional aligners introduces falsely classified UMRs. We investigated this and found that 1.61% of a total of 8.48 million of the UMRs were incorrectly classified as UMR by aligning just to the target region. This seemingly small percentage of incorrectly classified UMR leads to a significant increase of around 88% more SNP calls, close to 92% of which are false positives.

The gain in computation speed was investigated for a total of 900 parameter variations and was observed to range from an average of 6.2±0.8 for a 30 Mb target region to an average of 8±0.8 for a 3Mb target region when comparing the fastest Needleman–Wunsch implementation (NWBem) to Bowtie. As the alignment itself consumes only a fraction of the total computation time, using a binary format to process the reads should give additional benefits. For example, speeding up the I/O by a factor of 3 would decrease the alignment time from ~40 s to ~20 s for the ~5 million reads of a 3 Mb target at 20× read redundancy, which is ~16× faster than Bowtie. Furthermore, since the alignment algorithm can be exchanged easily and the computations do not require sophisticated hardware, using *a priori* information proves from a bioinformatics point of view to be a flexible and efficient approach to minimize alignment efforts in targeted sequencing and to enable a clinical use of sequencing information without the necessity of large computing facilities. Finally, the alignment time of around 7 min or less for a targeted resequencing run of approximately 49 million reads would be very attractive for clinical use.

Acknowledgements

We would like to thank Peter van Hooft and Jurgen Rusch for their support concerning grid computing facilities and massive parallelized computations. Part of this work was performed using the WIOS pipeline (11). Therefore, we also thank the other WIOS team members, Steffen Pallarz and Anika Tillich for their support. Moreover we would like to thank Harma Feitsma for stimulating discussions and constructive feedback on the manuscript.

References

1. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.-J., Chen,Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–80.

2. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

3. Dekker,C. (2007) Solid-state nanopores. *Nat. Nanotechnol.*, **2**, 209–15.

4. Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.*, **16**, 1–11.

5. Bellman,R. (1957) Dynamic Programming Princeton University Press, Princeton, New Jersey, USA.

6. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–53.

7. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–7.

8. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–8.

9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

10. Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–41.

11. Hammer,P., Banck,M.S., Amberg,R., Wang,C., Petznick,G., Luo,S., Khrebtukova,I., Schroth,G.P., Beyerlein,P. and Beutler,A.S. (2010) mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res.*, **20**, 847–60.

12. Illumina (2009) Complete Secondary Analysis Workflow for the Genome Analyzer.

13. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–8.

14. Burrows,M. and Wheeler,D. (1994) A block sorting lossless data compression algorithm.

15. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

16. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

17. Albert,T.J., Molla,M.N., Muzny,D.M., Nazareth,L., Wheeler,D., Song,X., Richmond,T.A., Middle,C.M., Rodesch,M.J., Packard,C.J., *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–5.

18. Hodges,E., Xuan,Z., Balija,V., Kramer,M., Molla,M.N., Smith,S.W., Middle,C.M., Rodesch,M.J., Albert,T.J., Hannon,G.J., *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–7.

19. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C., *et al.* (2009) Solution hybrid selection with ultra-long

oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–9.

20. Myllykangas,S., Buenrostro,J.D., Natsoulis,G., Bell,J.M. and Ji,H.P. (2011) Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. *Nat. Biotechnol.*, **29**, 1024–7.

21. Mokry,M., Feitsma,H., Nijman,I.J., de Bruijn,E., van der Zaag,P.J., Guryev,V. and Cuppen,E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res.*, **38**, e116.

22. Tillmann,C. and Ney,H. (2003) Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Comput. Linguist.*, **29**, 97–133.

23. Ensembl Human (Homo sapiens) http://www.ensembl.org/Homo_sapiens/Info/Index.html (20 May 2011, date last accessed).

24. Abecasis,G.R., Altshuler,D., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–73.

25. Hooge,F.N., Kleinpenning,T.G.M. and Vandamme,L.K.J. (1981) Experimental studies on 1/f noise. *Reports Prog. Phys.*, **44**, 479–532.

26. Johansson,H., Isaksson,M., Sörqvist,E.F., Roos,F., Stenberg,J., Sjöblom,T., Botling,J., Micke,P., Edlund,K., Fredriksson,S., *et al.* (2011) Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res.*, **39**, e8.

Chapter 7

General Discussion

Background

Prostate cancer represents a heterogeneous disease with diverse outcomes that can range from long-term symptom free survival to aggressive metastatic disease. Because of this, clinicians and researchers face grave challenges in proper diagnosis and patient stratification to provide the best care possible. However, despite continuous research efforts and known flaws, the PSA-based serum test introduced in the 1980s remains the de-facto standard assay for the indication of presence of prostate cancer until this day. PSA testing is associated with overtreatment of insignificant disease cases, leading to unnecessary biopsies and surgical interventions (1–3). Moreover, PSA lacks prognostic value at the time of diagnosis and is not sufficient to predict disease progression or determine a treatment course (4). Many alternative markers have been proposed, but mainly due to lack of validation, the adoption rate for clinical use is low and the need for highly specific biomarkers that can predict outcome and allow disease monitoring persists.

In order to address these issues and uncover novel alternative biomarkers as well as gain further insight in the molecular characteristics of prostate cancer, we utilized data from numerous high throughput technologies for our genome-wide studies. One of the focuses of our research was to investigate disease-associated RNA isoforms of known genes, which included PCa-specific promoter switching as encountered with the PDE4D gene. In addition, we repurposed existing array data to identify novel prostate cancer-associated lncRNAs and evaluated their biomarkers potential. Lastly, the software we development for targeted re-sequencing proved this technology's potential for clinical applications.

1 A new generation of markers and profiles

Ideally, biomarkers for diagnostic and monitoring purposes should be absolutely disease-specific to prevent any misclassification (false positives) that could cause overtreatment of patients. For PCa, such disease-specific markers are known to exist since the discovery of the fusion gene TMPRSS2-ERG (5). However, the lack of sensitivity due to the limited number of fusion gene-positive samples encompasses a risk of missing significant cases (false negatives). This shortcoming can be overcome by adding complementing biomarkers and creating a gene panel for biomarker purposes. Therefore, our aim was to identify other PCa-specific RNA transcripts that would be suitable to use in such a biomarker panel. In order to ensure high specificity, we chose to pursue an outlier-based approach similar to the one used for discovering TMPRSS2-ERG (5). However, instead of fusion genes, we focused on the identification of novel genes in previously unannotated regions of the genome in an effort to discover PCa-associated lncRNAs, a class of RNAs known to exhibit highly tissue-specific expression patterns (see Chapter 2).

With this approach, we identified 334 candidate transcripts referred to as EPCATs (EMC prostate cancer-associated transcripts), of which 15 were subsequently validated by RT-PCR and 12 had working qPCR probes that could be used to validate their diagnostic performance in an independent patient cohort (AUC = 0.87). Moreover, two of the validated EPCATs showed association with patient outcome, making them interesting prognostic biomarker

candidates. We also used two selected EPCATs for *in situ* hybridization on a tissue microarray and successfully distinguished PCa from surrounding tissue in 39% of all cases with 100% specificity, underscoring their value for needle biopsy evaluation and staging. In comparison, the genes most commonly hit by point mutations in PCa, TP53, SPOP and PTEN, occur in less than 15% of patients (14%, 9% and 7%, respectively (6)), while MYC amplifications are present in 2%-20% and NKX3-1 deletions are found in 35%-86% of prostate tumors (7, 8). Thus, our EPCATs show promise as diagnostic and prognostic markers for tissue assessment, but require further study concerning their potential as urine markers for early disease detection and monitoring. Furthermore, the cause for the observed diverse outlier expression patterns of the EPCATs remains unknown and a number of possible mechanisms can be envisioned to be involved.

One possible explanation could be a transcriptional regulation by specific transcription factors (TFs), in which the TF itself follows an outlier pattern as exemplified by ERG and ETV1 when fused to TMPRSS2. However, few EPCATs revealed a clear coexpression with ERG or ETV1, leading us to conclude that other TFs or perhaps specific TF combinations could be required for their transcription. To investigate this possibility further, we conducted a preliminary transcriptome-wide follow-up study using the Weighted Gene Coexpression Network Analysis (WGCNA (9)) framework as well as a custom approach termed XDmapper (10–12). These analyses identified several coexpressed TFs for individual EPCATs, but did not reveal a common cause for EPCAT expression, suggesting the involvement of other regulatory mechanisms.

Among these mechanisms, epigenetic factors such as DNA methylation are interesting candidates for further investigation, which is why we conducted a preliminary study on the correlation of DNA methylation patterns and EPCAT expression using a public dataset (13–16). Since an increased DNA methylation has been correlated with gene silencing, our aim was to identify losses of methylation (hypomethylation) near the promoter regions of EPCATs that could indicate an activation of the gene. However, similar to the analysis of TFs we found correlations of hypomethylation and expression only for a few individual EPCATs, while a global mechanism was not detected.

Besides the mentioned processes, one could also envision that expression of some of the EPCATs is caused by reoccurring fusion events, although we could not find evidence of break points near EPCAT loci when looking at available DNA-seq data of several PCa cell lines (17). Other possible explanations could be alterations in copy number that disrupt chromatin organization and alter enhancer activity, or by elongated primary transcription and thereby related to read-through fusion transcripts (conjoined genes). These mechanisms have not yet been investigated further.

From these findings, we hypothesize that expression of the EPCATs is unlikely caused by a single mechanism, and since the mentioned mechanisms need not be exclusive, unraveling the transcriptional regulation of the EPCATs will be a formidable challenge for future research. Such efforts may also provide further insights into tumor biology, as the underlying mechanisms could mark cellular aberrations or features important for cancerous growth, such as specific enhancer elements or chromatin domain boundaries.

1.1 Functional aspects of long non-coding RNAs

Although our knowledge of lncRNAs is far from complete, a growing body of evidence has challenged the notion of lncRNAs as a curiosity of cellular transcription without functional role in the cell ("junk"). Nonetheless, the debate of what these genes actually are and if they should be considered as potentially functional despite a lack of evolutionary conservation, is still ongoing (18–20). Current genome annotations comprise more than 60,000 lncRNA genes (21), making them approximately 3-times more abundant than protein-coding genes (22), yet so far research has been unable to provide a conclusive explanation for the plethora of lncRNAs as well as their origin. A commonly cited explanation is that lncRNA genes represent either evolutionary left-overs or occur due to spontaneous formation of transcribable sequences, and that their expression is caused by spurious RNA Polymerase II activity ("leaky transcription") (23). While this hypothesis may explain the existence of a number of non-coding transcripts, is does not provide an explanation for the often observed tissue-specific expression of lncRNAs (24, 25), and can therefore not be generalized to all transcripts currently classified as lncRNA. Moreover, numerous lncRNAs have been described as cancer-associated (26–28), and while their expression could be caused by genomic alterations or a less tightly controlled transcription, the underlying DNA sequences would face negative evolutionary pressure if they would be solely cancer-promoting and without additional function.

Functional roles of lncRNAs in cancer are further supported by siRNA-mediated knockdown of PCa-associated transcripts in PCa cell lines, which revealed impaired growth and/or cell motility (preliminary data) for 5 of the 9 tested RNAs (see Chapter 3). Even though only few examples have been studied, a diverse panel of functions has been associated to this RNA class, ranging from miRNA decoys to protein scaffolding and transcriptional regulation (see Chapter 2). Furthermore, discoveries in RNA epigenetics revealed that specific RNA modifications such as methylation of adenosine at the N6 position (m6A) are involved in many cellular processes and can impact the structure of RNA transcripts and their interactions with other intracellular molecules such as RNA-binding proteins (29–33). This could imply that a second regulatory layer for RNA function besides intracellular concentration exists, and that some lncRNAs require presence or absence of modification to perform specific roles in the cell.

Recently, the CRISPR/Cas9 system has been receiving a lot of attention as a powerful tool for genome editing, allowing precise knock-out of target genes in a massively parallel matter (34). Since the function of most lncRNAs remains unknown, genome-wide CRISPR screens could provide a first step towards identifying functional lncRNAs as targets for further study in individual tissues or diseases (35). Additionally, it has often been speculated that lncRNAs can in fact encode small peptides and lately evidence for this property has been accumulating (36, 37). With emerging novel techniques such as ribosome profiling to study ongoing translation, the number of such discoveries is likely to increase in the coming years and it is therefore questionable whether the simplistic categorization of coding and non-coding RNAs should be continued in the future (see Chapter 2 and (38)).

Nevertheless, lncRNAs currently represent a large pool of uncharacterized transcripts with potential functional impact on defined cell types and conditions and need to be examined further in order to evaluate clinical relevance. Once appropriate targets have been identified, one can imagine that lncRNAs whose expression is associated with malignant disease could be therapy targets for knock-out via genome editing techniques and likewise, condition-specific peptides produced from lncRNAs would be of strong interest as biomarkers. Certainly, these prospects currently remain fictional, however, first human trials are now being conducted in China (39, 40) and have been approved in the USA (41), rapidly closing the gap between science and fiction.

1.2 Alterations in PDE4D isoform expression – A marker for prostate cancer and other malignancies

Phosphodiesterases (PDEs) have been subject to many studies across a wide panel of diseases and clinical conditions. The best known examples are PDE5 inhibitors such as sildenafil, which are used to treat erectile dysfunction. In addition, abnormalities in PDEs have been reported in acrodysostosis (42–44), stroke (45, 46), COPD (47) and cancer (48, 49), while PDEs have also been suggested as possible treatment targets for brain injuries (50) and Alzheimer's disease (51).

Since cancer initiation is linked to inflammatory reactions and members of the PDE4 family are predominating cAMP hydrolysis in inflammatory cells, with PDE4D making up approximately 80% of PDE activity (52–56), it seemed appropriate to study PDE4D expression in PCa and search for links between its expression patterns and cancer development. Importantly, while the individual PDE4D isoforms share a common catalytic domain for cAMP hydrolysis, their transcription is regulated by independent promoters and can be adapted to tightly control cAMP signaling (57). For this reason, our analyses focused on individual PDE4D isoforms and their diagnostic and prognostic biomarker potential, with PDE4D7 being the first promising candidate based on preliminary findings in PCa cell lines (58).

Indeed, we were able to confirm that the PDE4D7 mRNA isoform is consistently up-regulated in localized disease, while its expression declines during disease progression. This over-expression was especially pronounced in patient samples showing ERG expression, implicating presence of the TMPRSS2-ERG fusion gene (see Chapter 4). Interestingly, although the preliminary study in VCaP cells could not find evidence for androgen signaling being directly involved in PDE4D7 expression (58), the correlation with ERG expression does imply at least an indirect involvement, as ERG up-regulation in PCa is mostly linked to a fusion with the AR-regulated TMPRSS2 gene.

We therefore continued to investigate the expression profiles of other PDE4D isoforms and found that PDE4D5 and PDE4D9 were down-regulated when compared to normal adjacent prostate tissue, revealing a PCa-specific promoter switch leading to a change in isoform composition (see Chapter 5). This promoter switch could be mediated by multiple factors, as we found increased DNA methylation in several loci located within in the PDE4D gene, which also overlapped the PDE4D5 transcription start site. In addition, cell line expression as

well as transcription factor-profiling data suggested that both AR and ERG could be involved in PDE4D7 up-regulation, again implicating a link between androgen and cAMP signaling.

Our findings are supported by a recent study investigating differentially methylated genes between TMPRSS2-ERG positive and negative samples, which provided evidence for several loci located in PDE4D that were hypermethylated in TMPRSS2-ERG positive samples (59). Moreover, a study in mice found that tissue- and stage specific DNA methylation patterns were correlated to Pde4d transcription (60), indicating that PDE4D expression is indeed regulated by epigenetic mechanisms. Lastly, ERG presence and absence has been associated with distinct DNA methylation patterns (15) and transcriptional control of the Polycomb Group protein EZH2 (61, 62), which is involved in DNA methylation. It is therefore plausible to hypothesize that ERG over-expression, for instance via TMPRSS2-ERG, may be disrupting the EZH2-mediated methylation program of prostate cells and thereby participating in the observed promoter switch.

Another possible explanation for the observed down-regulation may be structural variants, deletions or amplifications occurring in or near the PDE4D gene body (49), which could prevent expression of certain isoforms. However, when examining the influence of copy number on the expression levels of all nine PDE4D isoforms, we found that down-regulation of both PDE4D5 and PDE4D9 was also present in samples without a PDE4D deletion (Chapter 5). For this reason, we concluded that an active process such as gene silencing is likely to be responsible for promoter switching.

With our current knowledge, a functional reason for the observed promoter switch can only be speculative. Since PDE4D isoforms not only utilize independent promoters but also differ in their N-terminal region (57), PDE4D7 expression could enable PCa cells to alter their cAMP signaling by re-targeting hydrolytic activity to other cellular compartments. Another possibility could be the availability of a PKA phosphorylation site in the PDE4D7 protein that allows to inhibits it activity (63) and creates an additional layer of fine-tuning cAMP degradation.

With these findings and the broad spectrum of other medical conditions associated with PDEs, it is plausible to assume that expression of PDE enzyme isoforms could be altered similarly in other diseases. This hypothesis had been suggested previously by investigators of the deCODE consortium and others, arguing that the relative expression of PDE4D isoforms may regulate its enzymatic activity, in line with the idea of a compartmentalized cAMP signaling (57, 64, 65).

Therefore, it might be worthwhile to extend the presented studies across all known human PDEs on RNA isoform level and perform a thorough statistical investigation of their transcriptional profiles in prostate cancer, as well as other tissues. In this way, we may be able to uncover further associations of individual PDE isoforms and development of disease as well as specific outcomes, and create a multifactorial model for cAMP degradation activity in different cancers. A promising source for the required information can be found in the publicly available TCGA cohort (66), which currently provides access to RNA and/or protein expression data for more than 30 types of cancers. With this information, previously

overlooked similarities between the different cancer types might be discovered that could highlight potentially shared treatment options.

1.3 Implementation of biomarkers in a clinical setting

Current biomarkers for PCa are utilizing single RNAs or proteins to detect and stage disease. However, it is very unlikely to find "the" cancer gene, as the commonly accepted multiple hit theory predicts cancer to be caused by an accumulation of (different) events (67, 68). Since numerous genes were found to be associated with PCa development and progression in recent years, multi-RNA/protein signatures have been proposed as biomarkers for PCa (see for instance Chapter 3 and 5) and implemented in commercial tests for diagnostic (MiPS, (66)) or prognostic evaluation (Oncotype DX, Genomic Health, Inc; Prolaris, Myriad Genetics, Inc; Decipher, GenomeDx Biosciences, Inc. and more). With various different tests to choose from, comprehensive benchmark studies are urgently needed to allow clinicians to select the optimal tool for patient assessment (69). Unfortunately such performance reviews do not yet exist for the mentioned tests, while available validation studies have been conducted retrospectively and are often limited by small cohort size and varying RNA quality (69, 70). Other factors limiting the use of multi-gene tests in the clinic are specific equipment requirements that may not be available for all institutes (microdissection equipment for Decipher), as well as interobserver variation when selecting appropriate tissue for extraction (Prolaris), which can decrease test reliability (70). Considering that increased price and/or labor cost of the mentioned tests need to be justifiable by their added clinical value over current PSA- or Gleason-based protocols, which can also be compiled in prostate cancer risk calculators (71), clinical implementation of these tests has been slow.

In addition to these limitations, most commercial panels rely on increases or decreases of gene expression and/or DNA methylation in cancer to create a diagnostic or prognostic score. However, most genes are not specifically expressed in one tissue or condition and gene expression changes are transient, which complicates binary classification approaches. For this reason, we suggested a signature consisting of multiple highly specific markers (EPCATs, see Chapter 3) that can be complemented with other biomarkers such as specific rearrangements, point mutations or copy number alterations to increase its sensitivity.

Moreover, individual RNA isoforms of genes are often summarized into overall gene expression based on the assumption that the resulting protein isoforms share common interaction partners and have similar functions. This assumption has recently been challenged by Yang *et al.* who provided evidence for vastly different interaction profiles between multiple isoforms (72). These findings underline our results for a PCa-specific promoter switch of PDE4D that can be harnessed to create PCa signatures based on RNA isoform composition (see Chapter 5) and demonstrate the advantages of isoform-level analyses.

Although the biomarker candidates proposed in this thesis (Chapter 3: EPCATs and Chapters 4&5: PDE4D-based signatures) are not yet readily usable for clinical applications, current efforts focus on an independent pre-clinical validation and evaluation of applicability in liquid biopsies. Once successfully validated, these markers can prove to be valuable additions to existing signatures, be it for solid or, as the case may be, for liquid tissue samples.

Ideally, a novel diagnostic signature should not only offer a significant performance increase over PSA testing, but also allow easy testing in body fluids without the need of invasive sampling to control costs and reduce treatment burden. For prostate cancer, this would include both blood as well as urine for sampling, and efforts to sequence urine-derived RNA (referred to as 'Urinome') or analyze contents of urine-derived exosomes are ongoing. These endeavors could ideally result in a transcriptome-wide classifier based on urine or blood composition that overcomes the reliance on a small number of genes and instead can be used for stratification of healthy and diseased men using topological comparisons. However, robust measurement of the often heavily degraded RNAs is still a challenge and sequencing cost needs to decrease substantially before clinical adaptation of such a protocol can be realized. Until that time, utilization and optimization of existing targeted PCR, hybridization or antibody-based assays to measure gene signatures will remain one of the main priorities of clinical biomarker-related research.

2 Technology development and the future of big data, informatics and personalized medicine

Decreasing cost and time for NGS in conjunction with higher base coverage and accuracy as well as longer reads will lead to an increasing usage of sequencing for various purposes in the coming years. With more genomic data being produced every day, current estimates project a total storage capacity in a range of 2-40 exabytes (2-40 million terabytes) needed by 2025 for human genomes alone (73). This massive amount of data will not only pose a major challenge for existing IT infrastructure, but also for scientists who will need adequate education in descriptive statistics and machine learning techniques to be able to interpret and utilize results of large scale analyses. Here, cloud-based solutions such as Galaxy, which allow sharing of tested and time-stamped analysis workflows will certainly be of great value to ensure ease of use and reproducibility (74). Moreover, improved algorithms for data processing and mining are necessary to control computational requirements. Here, utilizing expertise across different fields of natural sciences could prove valuable to address increasing computational complexity. For instance, algorithms specifically designed for the emerging field of quantum computing, referred to as quantum algorithms, may be able to overcome some of the major challenges when dealing with big data and extracting useful information (75). Likewise, advanced machine learning techniques such as neural networks and developments in artificial intelligence will be of great value for data assessment and classification problems that could for instance be based on large gene panels or transcriptome-wide measurements.

Ultimately, such data-driven categorizations could then be utilized in clinical decision making and for personalized medicine, as early trials with the IBM Watson system providing advice on patient treatment have demonstrated (76, 77). If such endeavors can be extended to genomic data profiling and incorporate the currently promoted concept of drug repurposing and mutation profile-based treatment of patients (78, 79), an early-stage implementation of precision medicine might be feasible rather soon. Of course, a solely genomics-based treatment approach would not be capable to account for other major external factors that can influence treatment response, including diet, lifestyle and environmental factors. Here, one might argue that the recent developments in wearable technology as well as increasing

popularity of the "Quantified Self" (QS) movement (80–82) might offer a prospective compensatory solution in risk assessment of and treatment suggestions for eligible individuals. Similarly, developments of wearable technology with purely health care-oriented intent are currently ongoing, as evidenced by first efforts in developing 'smart contact lenses' for continuous measurements of blood sugar levels in diabetic patients (83) and a wristband that allows non-invasive surveillance of circulating tumor cells (CTCs) in the bloodstream (84). Such technologies are only examples of a possibly continuous monitoring of an individual's health status that could technically allow detecting carcinogenesis through, for instance, measurements in bodily fluids. When available, this information could then also be utilized in the mentioned decision system to improve treatment recommendations.

However, it has to be noted that any QS-based solution would not only require a widespread adoption of a QS mentality, but also face severe ethical and data security concerns due a massive collection of personal information that can entice misuse and lead to unforeseen (and possibly unwanted) revelations about the patient (82). In addition, next to the associated costs of a continuous patient surveillance, the actual benefits of incorporating such technology in decision making are uncertain and can be controversial, as exemplified by the ongoing discussion concerning PSA screening (85–87). For these reasons, the actual impact of wearable technology and possibly QS on future healthcare is not yet reliably assessable beyond a marketing strategy perspective. Nevertheless, these limitations do not impair the theoretical strength of a decision system able to incorporate and balance a variety of available information to provide informed recommendations on the possible course of action. Lastly, as a distant extension of this personalized treatment concept one could also envision a combination of genome-sequencing and –editing of specific mutations to treat diseased cells or disorders, as has been showcased for retinitis pigmentosa, in which the CRISPR/Cas9 system was used to repair an RPGR point mutation (88).

Although these developments offer exciting opportunities and challenges for clinicians and scientists, ethical concerns of data usage and anonymization have to be addressed and discussed thoroughly in order to ensure a responsible data management and prevent issues such as genome-based discrimination.

2.1 Targeted re-sequencing as promising clinical method for disease classification and choice of treatment in precision medicine

The rapid cost decrease introduced by next generation sequencing (NGS) technologies has spawned many novel concepts for clinical sequencing applications including precision medicine (also known as personalized medicine), in which a patient is treated according to the molecular characteristics specifically identified in his / her disease. To implement such tailored treatments, a range of drugs targeting commonly found genetic alterations is required, which can be administered once an appropriate assay validates the presence of such alteration in the patient’s tumor. Current efforts for drug repurposing and combinations of drugs show promising results (89–91), however, until now the cost for genome-wide sequencing of patient samples still prevent a wide-spread application. Here, targeted re-sequencing of a panel of previously identified genomic regions such as susceptibility loci may be a valuable intermediate step to introduce the benefits of NGS to the clinic while reducing the associated

cost. As an example, recurrently mutated genes in PCa that are clinically actionable targets include PTEN (via PI3K inhibitors), ATM/BRCA1/BRCA2 (by PARP inhibitors) and BRAF/RAF1 (via RAF or MEK inhibitors) (92). In addition, next to reducing cost, the continuous advances in sequencing technology also decrease the time needed for an actual sequencing run, resulting in larger amounts of data being produced in shorter amounts of time. These data then need to be processed by post-sequencing bioinformatics analyses in a timely manner, effectively shifting the rate limiting step of mutation profiling to the downstream data analysis (93).

To explore how this emerging bottleneck posed by data processing could be circumvented, we tested whether a naive reduction of alignment space for genome-wide mapping to only the target region of interest could offer a sufficient solution. However, we found that this approach creates a considerable number of false positive mutation calls by reads being forced to map uniquely with high mapping scores due to the restricted search space, which could not be rescued by increasing quality thresholds. Hence, these reads were undistinguishable from reads truly originating from the target region although they would have not contributed to mutation calling at the positions of interest in a genome-wide setting. This finding highlights the necessity of genome-wide alignments for mutation calling where high alignment accuracy is desirable, but might also be relevant when aligning reads to the transcriptome only, as is often the case for gene quantification in RNA-seq experiments. However, although such forced alignments could increase the variance of gene expression estimates, their overall contribution on accuracy should be limited since expression estimates are based on the whole gene or transcript and therefore less dependent on contributions of individual reads or a single nucleotide resolution. For this reason, a recent generation of quantification tools that utilizes k-mer counting and/or "pseudoalignment" approaches is able to provide an accurate quantification without the need for a complete base-for-base alignment (94–97).

Nonetheless, it is apparent that specialized algorithms are required to address the informatics bottleneck for targeted sequencing and subsequent mutation calling. Consequently, we implemented a novel read alignment approach based on *a priori* information available from a capture platform now known as Haloplex™, demonstrating substantial improvements in both time and computational resources required for read mapping.

Since our approach utilized the specific capture design of the Haloplex technology, an adaptation to other targeted sequencing protocols does not appear trivial as *a priori* information on the likely origin of the reads is required for its function. However, this limitation could be resolved by a novel concept termed "quasi-mapping", in which the mapping algorithm does not perform a traditional base-for-base alignment and instead tries to narrow down the likely loci of origin using a combination of a suffix array and a hash table to index the region of interest perform search operations (98, 99). In this way, if a read can be mapped to the reference region, one or more genomic regions are identified as potential origins, which could then be supplied to our algorithm as *a priori* information to obtain an accurate sequence alignment including mismatches. Current software implementations utilizing this novel technique focus on rapid transcriptome quantification as previously mentioned (95, 97, 99), but adaptations for DNA-sequencing applications are planned for

future releases. With these, a time and resource costly genome-wide alignment for targeted re-sequencing could be implemented, enabling a rapid analysis in minutes and thereby eliminating the bioinformatics bottleneck in clinical applications. Subsequently, rapid profiling of diagnostic and prognostic point mutations or risk SNPs could be combined with RNA markers (see Chapter 3-5) to enable a faster and more accurate decision-making process (see Section 2).

3 The promise of big data and the computational revolution for molecular biology

Next generation sequencing as well as its derivative techniques has undoubtedly revolutionized the fields of molecular biology and cancer research by challenging existing concepts and providing detailed information on many different levels of cellular organization. Genome-wide studies of epigenetic regulation and chromatin organization revealed a much more complex system than originally anticipated and further discoveries are made continuously as demonstrated by the recent discoveries of tens-of-thousands of new lncRNAs and other epigenetic modifications of both DNA and RNA (30, 100, 101). However, due to the continuing improvements of NGS technologies and a growing number of experimental protocols, it can be expected that the discovery of novel intracellular molecules will soon reach a plateau. Functional characterization and mechanistic studies of molecules and their modifications will thus become essential to address the challenge to integrate all of these different concepts into a full working model of cellular architecture that recaptures the dynamic processes ongoing in every cell. Here, machine learning techniques will also be of great importance to identify patterns in existing data and predict functions and interaction partners for uncharacterized molecules. In this way, it can be envisioned that transcripts currently summarized under the generalized term lncRNAs will be reassigned to different functional classes and grouped with previously unrelated RNAs. This function-based classification in turn promotes concepts of multi-functional RNAs (38) and function as an emergent property which mainly depends on the availability of appropriate interaction partners. Using such concepts, modeling of the complex cellular system may grant us further insights into its organization and its responses to perturbations such as mutations and chromosomal aberrations.

Lastly, the sharing and repurposing of NGS data via repositories such as Gene Expression Omnibus and EGA allows anyone with sufficient biological and informatics knowledge to conduct their own studies and test hypotheses. As an example, in several of the studies presented in this thesis, usage of publicly available high-throughput data allowed a broader view on the study subject and more detailed analyses of regulatory mechanisms. Furthermore, platforms such as Galaxy allow cloud-based implementation and sharing of tested and time-stamped analysis workflows, improving reproducibility while simultaneously simplifying usage for scientists without bioinformatics training (74). Involvement of non-experts via open-access publishing as well as crowdsourcing may also be beneficial and help to raise public interest for research as demonstrated by the successful implementation of the Folding@home and Rosetta@home networks. As exemplified by these ongoing efforts, science has begun to transform into an open-source community that allows rapid exchange of scientific ideas and in doing so, shares many similarities with modern information technology.

Thus, to avoid false conclusions based on improper analyses, good data management as promoted by the Dutch FAIR Data Principles (Findable, Accessible, Interoperable and Reusable (102)) as well as the adoption of the scientific method including proper validation of results by all participating parties will be crucial mechanisms for future scientific efforts.

Addressing these numerous challenges and utilizing the opportunities provided by big data, will therefore be major stepping stones on the way to a better understanding of the molecular mechanisms involved in cancer formation and many other diseases.

References

1. Andriole,G.L., Crawford,E.D., Grubb,R.L., Buys,S.S., Chia,D., Church,T.R., Fouad,M.N., Gelmann,E.P., Kvale,P.A., Reding,D.J., *et al.* (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.

2. Roobol,M.J. and Carlsson,S. V (2013) Risk stratification in prostate cancer screening. *Nat. Rev. Urol.*, **10**, 38–48.

3. Schröder,F.H., Hugosson,J., Roobol,M.J., Tammela,T.L.J., Ciatto,S., Nelen,V., Kwiatkowski,M., Lujan,M., Lilja,H., Zappa,M., *et al.* (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.

4. Vicini,F.A., Vargas,C., Abner,A., Kestin,L., Horwitz,E. and Martinez,A. (2005) Limitations in the use of serum prostate specific antigen levels to monitor patients after treatment for prostate cancer. *J. Urol.*, **173**, 1456–62.

5. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.-W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R., *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–8.

6. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S., *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–11.

7. Khemlina,G., Ikeda,S. and Kurzrock,R. (2015) Molecular landscape of prostate cancer: Implications for current clinical trials. *Cancer Treat. Rev.*, **41**, 761–766.

8. Gurel,B., Ali,T.Z., Montgomery,E.A., Begum,S., Hicks,J., Goggins,M., Eberhart,C.G., Clark,D.P., Bieberich,C.J., Epstein,J.I., *et al.* (2010) NKX3.1 as a marker of prostatic origin in metastatic tumors. *Am. J. Surg. Pathol.*, **34**, 1097–105.

9. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

10. Schaeffer,F. (2012) Co-expression analysis of novel genes in prostate cancer. Hogeschool Leiden. B.Sc. thesis.

11. van der Lans,C. (2013) Co-expression analysis of potential prostate cancer biomarkers. Hogeschool Leiden. B.Sc. thesis.

12. Nieuwenhuijse,D. (2015) Investigation of the transcriptional regulatory network of novel prostate cancer biomarkers. Wageningen University and Research Centre. M.Sc. thesis.

13. Perez Vara,M. (2015) Novel long non coding RNAs as promising prostate cancer prognostic biomarkers. Erasmus Medical Center, M.Sc. thesis.

14. Plugge,W. (2016) In silico validation and analysis of epigenetic regulation of novel PCa-associated lncRNAs. Wageningen University and Research Centre. M.Sc. thesis.

15. Börno,S.T., Fischer,A., Kerick,M., Fälth,M., Laible,M., Brase,J.C., Kuner,R., Dahl,A., Grimm,C., Sayanjali,B., *et al.* (2012) Genome-wide DNA methylation events in TMPRSS2-ERG fusion-

negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov.*, **2**, 1024–35.

16. Brase,J.C., Johannes,M., Mannsperger,H., Fälth,M., Metzger,J., Kacprzyk,L.A., Andrasiuk,T., Gade,S., Meister,M., Sirma,H., *et al.* (2011) TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. *BMC Cancer*, **11**, 507.

17. Teles Alves,I., Hartjes,T., McClellan,E., Hiltemann,S., Böttcher,R., Dits,N., Temanni,M.R., Janssen,B., van Workum,W., van der Spek,P., *et al.* (2015) Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene*, **34**, 568–77.

18. Palazzo,A.F. and Lee,E.S. (2015) Non-coding RNA: what is functional and what is junk? *Front. Genet.*, **6**.

19. Ling,H., Vincent,K., Pichler,M., Fodde,R., Berindan-Neagoe,I., Slack,F.J. and Calin,G.A. (2015) Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*, **34**, 5003–11.

20. Graur,D., Zheng,Y. and Azevedo,R.B.R. (2015) An evolutionary classification of genomic function. *Genome Biol. Evol.*, **7**, 642–5.

21. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S., *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.

22. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760–1774.

23. Kung,J.T.Y., Colognori,D. and Lee,J.T. (2013) Long noncoding RNAs: past, present, and future. *Genetics*, **193**, 651–69.

24. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J., *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–5.

25. Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.

26. Böttcher,R., Hoogland,A.M., Dits,N., Verhoef,E.I., Kweldam,C., Waranecki,P., Bangma,C.H., van Leenders,G.J.L.H. and Jenster,G. (2015) Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. *Oncotarget*.

27. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S., Kominsky,H.D., *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.*, **29**, 742–9.

28. Bussemakers,M.J., van Bokhoven,A., Verhaegh,G.W., Smit,F.P., Karthaus,H.F., Schalken,J.A., Debruyne,F.M., Ru,N. and Isaacs,W.B. (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.*, **59**, 5975–9.

29. Wang,X., Zhao,B.S., Roundtree,I.A., Lu,Z., Han,D., Ma,H., Weng,X., Chen,K., Shi,H. and He,C.

(2015) N6-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell*, **161**, 1388–1399.

30. Sergiev,P. V, Golovina,A.Y., Osterman,I.A., Nesterchuk,M. V, Sergeeva,O. V, Chugunova,A.A., Evfratov,S.A., Andreianova,E.S., Pletnev,P.I., Laptev,I.G., *et al.* (2015) N6-methylated adenosine in RNA: From bacteria to humans. *J. Mol. Biol.*, 10.1016/j.jmb.2015.12.013.

31. Chandola,U., Das,R. and Panda,B. (2015) Role of the N6-methyladenosine RNA mark in gene regulation and its implications on development and disease. *Brief. Funct. Genomics*, **14**, 169–79.

32. Dominissini,D., Moshitch-Moshkovitz,S., Schwartz,S., Salmon-Divon,M., Ungar,L., Osenberg,S., Cesarkas,K., Jacob-Hirsch,J., Amariglio,N., Kupiec,M., *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–6.

33. Fu,Y., Dominissini,D., Rechavi,G. and He,C. (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, **15**, 293–306.

34. Sander,J.D. and Joung,J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, **32**, 347–55.

35. Goff,L.A. and Rinn,J.L. (2015) Linking RNA biology to lncRNAs. *Genome Res.*, **25**, 1456–1465.

36. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.

37. Anderson,D.M., Anderson,K.M., Chang,C.-L., Makarewich,C.A., Nelson,B.R., McAnally,J.R., Kasaragod,P., Shelton,J.M., Liou,J., Bassel-Duby,R., *et al.* (2015) A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell*, **160**, 595–606.

38. Karapetyan,A.R., Buiting,C., Kuiper,R.A. and Coolen,M.W. (2013) Regulatory Roles for Long ncRNA and mRNA. *Cancers (Basel).*, **5**, 462–90.

39. Liang,P., Xu,Y., Zhang,X., Ding,C., Huang,R., Zhang,Z., Lv,J., Xie,X., Chen,Y., Li,Y., *et al.* (2015) CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell*, **6**, 363–372.

40. Cyranoski,D. (2016) Chinese scientists to pioneer first human CRISPR trial. *Nature*, 10.1038/nature.2016.20302.

41. Reardon,S. (2016) First CRISPR clinical trial gets green light from US panel. *Nature*, 10.1038/nature.2016.20137.

42. Kaname,T., Ki,C.-S., Niikawa,N., Baillie,G.S., Day,J.P., Yamamura,K.-I., Ohta,T., Nishimura,G., Mastuura,N., Kim,O.-H., *et al.* (2014) Heterozygous mutations in cyclic AMP phosphodiesterase-4D (PDE4D) and protein kinase A (PKA) provide new insights into the molecular pathology of acrodysostosis. *Cell. Signal.*, **26**, 2446–59.

43. Michot,C., Le Goff,C., Goldenberg,A., Abhyankar,A., Klein,C., Kinning,E., Guerrot,A.M., Flahaut,P., Duncombe,A., Baujat,G., *et al.* (2012) Exome sequencing identifies PDE4D mutations as another cause of acrodysostosis. *Am. J. Hum. Genet.*, **90**, 740–745.

44. Lee,H., Graham,J.M., Rimoin,D.L., Lachman,R.S., Krejci,P., Tompson,S.W., Nelson,S.F.,

Krakow,D. and Cohn,D.H. (2012) Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am. J. Hum. Genet.*, **90**, 746–751.

45. Gretarsdottir,S., Thorleifsson,G., Reynisdottir,S.T., Manolescu,A., Jonsdottir,S., Jonsdottir,T., Gudmundsdottir,T., Bjarnadottir,S.M., Einarsson,O.B., Gudjonsdottir,H.M., *et al.* (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.*, **35**, 131–138.

46. Houslay,M.D. (2005) The long and short of vascular smooth muscle phosphodiesterase-4 as a putative therapeutic target. *Mol. Pharmacol.*, **68**, 563–7.

47. Yoon,H.-K., Hu,H.-J., Rhee,C.-K., Shin,S.-H., Oh,Y.-M., Lee,S.-D., Jung,S.-H., Yim,S.-H., Kim,T.-M. and Chung,Y.-J. (2014) Polymorphisms in PDE4D are associated with a risk of COPD in non-emphysematous Koreans. *COPD*, **11**, 652–8.

48. Rahrmann,E.P., Collier,L.S., Knutson,T.P., Doyal,M.E., Kuslak,S.L., Green,L.E., Malinowski,R.L., Roethe,L., Akagi,K., Waknitz,M., *et al.* (2009) Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping beauty transposon-based somatic mutagenesis screen. *Cancer Res.*, **69**, 4388–4397.

49. Lin,D.-C., Xu,L., Ding,L.-W., Sharma,A., Liu,L.-Z., Yang,H., Tan,P., Vadgama,J., Karlan,B.Y., Lester,J., *et al.* (2013) Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. *Proc. Natl. Acad. Sci.*, **110**, 6109–6114.

50. Titus,D.J., Oliva,A.A., Wilson,N.M. and Atkins,C.M. (2015) Phosphodiesterase inhibitors as therapeutics for traumatic brain injury. *Curr. Pharm. Des.*, **21**, 332–42.

51. Heckman,P.R.A., Wouters,C. and Prickaerts,J. (2015) Phosphodiesterase inhibitors as a target for cognition enhancement in aging and Alzheimer's disease: a translational overview. *Curr. Pharm. Des.*, **21**, 317–31.

52. Jacob,C., Martin-Chouly,C. and Lagente,V. (2002) Type 4 phosphodiesterase-dependent pathways: role in inflammatory processes. *Therapie*, **57**, 163–8.

53. Liu,H. and Maurice,D.H. (1999) Phosphorylation-mediated activation and translocation of the cyclic AMP-specific phosphodiesterase PDE4D3 by cyclic AMP-dependent protein kinase and mitogen-activated protein kinases. A potential mechanism allowing for the coordinated regulation of PDE4D . *J. Biol. Chem.*, **274**, 10557–65.

54. Coussens,L.M. and Werb,Z. (2002) Inflammation and cancer. *Nature*, **420**, 860–7.

55. Lu,H., Ouyang,W. and Huang,C. (2006) Inflammation, a key event in cancer development. *Mol. Cancer Res.*, **4**, 221–33.

56. Elinav,E., Nowarski,R., Thaiss,C.A., Hu,B., Jin,C. and Flavell,R.A. (2013) Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat. Rev. Cancer*, **13**, 759–71.

57. Houslay,M.D. (2010) Underpinning compartmentalised cAMP signalling through targeted cAMP breakdown. *Trends Biochem. Sci.*, **35**, 91–100.

58. Henderson,D.J.P., Byrne,A., Dulla,K., Jenster,G., Hoffmann,R., Baillie,G.S. and Houslay,M.D. (2014) The cAMP phosphodiesterase-4D7 (PDE4D7) is downregulated in androgen-independent prostate cancer cells and mediates proliferation by compartmentalising cAMP at the plasma

membrane of VCaP prostate cancer cells. *Br. J. Cancer*, **110**, 1278–87.

59. Geybels,M.S., Alumkal,J.J., Luedeke,M., Rinckleb,A., Zhao,S., Shui,I.M., Bibikova,M., Klotzle,B., van den Brandt,P.A., Ostrander,E.A., *et al.* (2015) Epigenomic profiling of prostate cancer identifies differentially methylated genes in TMPRSS2:ERG fusion-positive versus fusion-negative tumors. *Clin. Epigenetics*, **7**, 128.

60. Huang,Z., Han,Z., Cui,W., Zhang,F., He,H., Zeng,T., Sugimoto,K. and Wu,Q. (2013) Dynamic expression pattern of Pde4d and its relationship with CpG methylation in the promoter during mouse embryo development. *Biochem. Biophys. Res. Commun.*, **441**, 982–7.

61. Kunderfranco,P., Mello-Grand,M., Cangemi,R., Pellini,S., Mensah,A., Albertini,V., Malek,A., Chiorino,G., Catapano,C. V and Carbone,G.M. (2010) ETS transcription factors control transcription of EZH2 and epigenetic silencing of the tumor suppressor gene Nkx3.1 in prostate cancer. *PLoS One*, **5**, e10547.

62. Yu,J., Yu,J., Mani,R.S., Cao,Q., Brenner,C.J., Cao,X., Wang,X., Wu,L., Li,J., Hu,M., *et al.* (2010) An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression. *Cancer Cell*, **17**, 443–454.

63. Byrne,A.M., Elliott,C., Hoffmann,R. and Baillie,G.S. (2015) The activity of cAMP-phosphodiesterase 4D7 (PDE4D7) is regulated by protein kinase A-dependent phosphorylation within its unique N-terminus. *FEBS Lett.*, **589**, 750–5.

64. Wang,D., Deng,C., Bugaj-Gaweda,B., Kwan,M., Gunwaldsen,C., Leonard,C., Xin,X., Hu,Y., Unterbeck,A. and De Vivo,M. (2003) Cloning and characterization of novel PDE4D isoforms PDE4D6 and PDE4D7. *Cell. Signal.*, **15**, 883–91.

65. Gulcher,J.R., Gretarsdottir,S., Helgadottir,A. and Stefansson,K. (2005) Genes contributing to risk for common forms of stroke. *Trends Mol. Med.*, **11**, 217–24.

66. Cancer Genome Atlas Research Network. (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**, 1011–1025.

67. Nordling,C.O. (1953) A new theory on cancer-inducing mechanism. *Br. J. Cancer*, **7**, 68–72.

68. Knudson,A.G. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, **68**, 820–3.

69. Boström,P.J., Bjartell,A.S., Catto,J.W.F., Eggener,S.E., Lilja,H., Loeb,S., Schalken,J., Schlomm,T. and Cooperberg,M.R. (2015) Genomic Predictors of Outcome in Prostate Cancer. *Eur. Urol.*, **68**, 1033–44.

70. Na,R., Wu,Y., Ding,Q. and Xu,J. Clinically available RNA profiling tests of prostate tumors: utility and comparison. *Asian J. Androl.*, **18**, 575–9.

71. Roobol,M.J., Zhu,X., Schröder,F.H., van Leenders,G.J., van Schaik,R.H., Bangma,C.H. and Steyerberg,E.W. (2013) A Calculator for Prostate Cancer Risk 4 Years After an Initially Negative Screen: Findings from ERSPC Rotterdam. *Eur. Urol.*, **63**, 627–33.

72. Yang,X., Coulombe-Huntington,J., Kang,S., Sheynkman,G.M., Hao,T., Richardson,A., Sun,S., Yang,F., Shen,Y.A., Murray,R.R., *et al.* (2016) Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **164**, 805–817.

73. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big Data: Astronomical or Genomical? *PLOS Biol.*, **13**, e1002195.

74. Goecks,J., Nekrutenko,A., Taylor,J. and Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

75. Lloyd,S., Garnerone,S. and Zanardi,P. (2016) Quantum algorithms for topological and geometric analysis of data. *Nat. Commun.*, **7**, 10138.

76. Doyle-Lindrud,S. (2015) Watson will see you now: a supercomputer to help clinicians make informed treatment decisions. *Clin. J. Oncol. Nurs.*, **19**, 31–2.

77. Oncologists partner with Watson on genomics. (2015) *Cancer Discov.*, **5**, 788.

78. Kwak,E.L., Bang,Y.-J., Camidge,D.R., Shaw,A.T., Solomon,B., Maki,R.G., Ou,S.-H.I., Dezube,B.J., Jänne,P.A., Costa,D.B., *et al.* (2010) Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.*, **363**, 1693–703.

79. Oprea,T.I. and Mestres,J. (2012) Drug repurposing: far beyond new targets for old drugs. *AAPS J.*, **14**, 759–63.

80. Vesnic-Alujevic,L., Breitegger,M. and Guimarães Pereira,Â. (2016) 'Do-It-Yourself' Healthcare? Quality of Health and Healthcare Through Wearable Sensors. *Sci. Eng. Ethics*, 10.1007/s11948-016-9771-4.

81. Shull,P.B., Jirattigalachote,W., Hunt,M.A., Cutkosky,M.R. and Delp,S.L. (2014) Quantified self and human movement: a review on the clinical impact of wearable sensing and feedback for gait analysis and intervention. *Gait Posture*, **40**, 11–9.

82. Kostkova,P., Brewer,H., de Lusignan,S., Fottrell,E., Goldacre,B., Hart,G., Koczan,P., Knight,P., Marsolier,C., McKendry,R.A., *et al.* (2016) Who Owns the Data? Open Data for Healthcare. *Front. public Heal.*, **4**, 7.

83. Farandos,N.M., Yetisen,A.K., Monteiro,M.J., Lowe,C.R. and Yun,S.H. (2015) Contact lens sensors in ocular diagnostics. *Adv. Healthc. Mater.*, **4**, 792–810.

84. Conrad,A.J. (2015) Nanoparticle Phoresis.

85. Albertsen,P.C. (2015) Prostate-specific antigen testing: good or bad? *Oncologist*, **20**, 233–5.

86. Bailey,S.-J. V and Brewster,S.F. (2011) Prostate cancer: to screen or not to screen. *Arch. españoles Urol.*, **64**, 406–18.

87. Kim,E.H. and Andriole,G.L. (2015) Prostate-specific antigen-based screening: controversy and guidelines. *BMC Med.*, **13**, 61.

88. Bassuk,A.G., Zheng,A., Li,Y., Tsang,S.H. and Mahajan,V.B. (2016) Precision Medicine: Genetic Repair of Retinitis Pigmentosa in Patient-Derived Stem Cells. *Sci. Rep.*, **6**, 19969.

89. Bertolini,F., Sukhatme,V.P. and Bouche,G. (2015) Drug repurposing in oncology-patient and health systems opportunities. *Nat. Rev. Clin. Oncol.*, **12**, 732–42.

90. Lavecchia,A. and Cerchia,C. (2015) In silico methods to address polypharmacology: Current status, applications and future perspectives. *Drug Discov. Today*, 10.1016/j.drudis.2015.12.007.

91. Marques,R.B., Aghai,A., de Ridder,C.M.A., Stuurman,D., Hoeben,S., Boer,A., Ellston,R.P., Barry,S.T., Davies,B.R., Trapman,J., *et al.* (2015) High Efficacy of Combination Therapy Using PI3K/AKT Inhibitors with Androgen Deprivation in Prostate Cancer Preclinical Models. *Eur. Urol.*, **67**, 1177–85.

92. Robinson,D., Van Allen,E.M., Wu,Y.-M., Schultz,N., Lonigro,R.J., Mosquera,J.-M., Montgomery,B., Taplin,M.-E., Pritchard,C.C., Attard,G., *et al.* (2015) Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*, **161**, 1215–1228.

93. Green,E.D. and Guyer,M.S. (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature*, **470**, 204–13.

94. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 10.1038/nbt.3519.

95. Patro,R., Mount,S.M. and Kingsford,C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–4.

96. Patro,R., Duggal,G. and Kingsford,C. (2015) Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment Cold Spring Harbor Labs Journals.

97. Zhang,Z. and Wang,W. (2014) RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics*, **30**, i283–i292.

98. Grabowski,S. and Raniszewski,M. (2014) Two simple full-text indexes based on the suffix array.

99. Srivastava,A., Sarkar,H., Gupta,N. and Patro,R. (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, **32**, i192–i200.

100. Plongthongkum,N., Diep,D.H. and Zhang,K. (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.*, **15**, 647–61.

101. Volders,P.-J., Helsens,K., Wang,X., Menten,B., Martens,L., Gevaert,K., Vandesompele,J. and Mestdagh,P. (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, **41**, D246–51.

102. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.

Summary

Prostate cancer (PCa) is a disease commonly found in western societies and has been associated with age as well as western lifestyle. Around 12,000 men are diagnosed with PCa in the Netherlands each year, and 3000 men die because of the disease. PCa is marked by divergent outcomes ranging from long-term symptom free survival to aggressive metastatic disease and until now it remains challenging to accurately predict how a tumor will behave. To be able to better distinguish indolent from malignant cases, a deeper understanding of the disease at the molecular level as well as better biomarkers are urgently needed. In this thesis, we focused on discovering novel biomarker candidates to aid the clinical need of better patient stratification. Moreover, we investigated a methodology for targeted next generation sequencing (NGS) that can be used for patient diagnosis and staging. In conjunction, our findings may provide valuable tools to improve patient care and reduce unnecessary treatments.

Chapter 1 represent a general introduction to PCa as well as the difficulties faced in detection and staging of the disease. We also describe the molecular characteristics of and cellular signaling pathways involved in PCa. Additionally, we outline technological and bioinformatics developments in recent years, specifically NGS that have had great impact on PCa research in this thesis and that show great promise for clinical applications.

In **chapter 2**, we provide an extensive review of long non-coding RNAs (lncRNAs) in urological malignancies and illustrate their potential for diagnosis and staging of PCa.

To study lncRNAs as potential PCa biomarkers, in **chapter 3**, we utilized the Affymetrix Human Exon Array platform which offers probe coverage of many genomic regions that do not have annotated genes in them. Our approach was founded on the cancer outlier profile analysis used for finding the PCa-specific TMPRSS2-ERG fusion gene and focused on probes without known gene association. With the additional samples provided by independent public datasets, we were able to identify 334 candidate regions in the genome showing signal in PCa samples only (referred to as EPCATs). We then set out to validate the top 20 EPCATs via RT-PCR and were able to confirm 15 novel PCa-associated transcripts. Our efforts in using a quantitative RT-PCR were successful for 12 of these RNAs, which we combined in a gene panel with very high diagnostic power. We also visualized two EPCATs in pathological tissue sections using *in situ* hybridization and confirmed their highly specific expression patterns. Moreover, we found that two EPCATs located on chromosome 2 were predictive of disease progression and development of metastasis. Lastly, we computationally evaluated the coding potential of our validated EPCATs and concluded that they most likely represent lncRNAs as initially proposed. In conclusion, we discovered and validated previously unannotated genes that can be used as highly specific biomarkers for PCa.

Previous research found that a specific isoform of the cAMP-specific 3',5'-cyclic phosphodiesterase 4D (PDE4D) was down-regulated in cell lines that represent more aggressive forms of PCa. We therefore investigated in **chapter 4**, whether this down-regulation of PDE4D7 could be verified in human tissue samples across a broad panel of

independent datasets. Our results confirmed the association of PDE4D7 expression with PCa stage, as samples of non-progressive primary PCa showed higher expression when compared to samples with progressive disease as well as normal control tissues. Moreover, we found that PDE4D7 expression was increased in samples harboring the TMPRSS2-ERG fusion gene compared to fusion-negative samples and normal tissue. Therefore, PDE4D7 is a potential PCa biomarker and represents a highly interesting therapy target.

Since mounting evidence points towards a cross-talk of the androgen receptor (AR) pathway and cAMP signaling, we investigated whether expression of other PDE4D isoforms besides PDE4D7, showed association with PCa progression in **chapter 5**. Utilizing several independent datasets, we found that both PDE4D5 and PDE4D9 are down-regulated in localized primary PCa, uncovering an isoform switch upon PCa development. To elucidate molecular mechanisms responsible, we checked whether chromosomal deletions, transcription factor binding or DNA methylation patterns were involved in PDE4D regulation. We found that down-regulation of PDE4D5 and PDE4D9 occurs independently of deletions in the gene locus, and that PDE4D5 expression was decreased in the LNCaP PCa cell line upon androgen signaling stimulation. Furthermore, ERG expression does not seem to affect PDE4D5 and PDE4D9, while conversely, PDE4D7 showed an increased expression when AR stimulus was supplied as well as in samples with high ERG expression. In addition, several loci with increased DNA methylation in PCa samples could be identified in the PDE4D gene, one of which located in proximity to the PDE4D5 promoter. Based on the gathered evidence, we created two signatures based on different isoforms and evaluated their diagnostic and prognostic performance across several datasets. Lastly, we also showed that our diagnostic signature can be used to improve needle biopsy staging. Our findings provide evidence for deregulation of PDE4D isoform composition in PCa and highlight the importance of this gene as PCa biomarker as well as target for therapeutic intervention.

While NGS is broadly used in research for discovery and validation purposes, its clinical implementation is still limited. To advance the utilization of cancer-associated genetic alterations in a clinical context, in **chapter 6**, we investigated informatics bottlenecks in the clinical application of targeted sequencing of patient genomes. Since standard software was designed for genome-wide alignment of NGS reads, a reduction of complexity by aligning only to the targeted region of interest causes many reads to be forcefully aligned with high mapping scores. This leads to severe false positives in mutation calling (88% additional SNP calls, 92% of which false positive), making a genome-wide alignment a necessity for conventional tools. We therefore implemented a novel alignment approach that utilized *a priori* information of a targeted sequencing technique to align NGS reads in a much shorter time and with lower memory requirements compared to existing methods. This methodology allows a highly efficient processing of NGS data that can provide valuable genetic information of a patient's tumor tissue for tailored treatment strategies and personalized care.

In conclusion, this thesis provides several new biomarkers for prostate cancer that can help to address the flaws of current protocols and discriminate indolent from aggressive cases. The EPCATs show great promise as highly specific lncRNAs, while PDE4D deregulation indicates broader alterations in cAMP signaling in PCa. Moreover, this thesis also advances

the clinical utilization of targeted re-sequencing by providing an efficient means to reduce computational burden during data analysis using *a priori* knowledge.

Samenvatting

Prostaatkanker is een vaak voorkomende ziekte in de Westerse wereld en geassocieerd met hoge leeftijd. In Nederland krijgen jaarlijks ongeveer 11.000 mannen te horen dat zij prostaatkanker hebben. De ziekte wordt gekenmerkt door een variabele patiënten uitkomst: van lange termijn symptoomvrije overleving tot agressieve uitgezaaide ziekte. Jaarlijks overlijden er ongeveer 2600 mannen in Nederland aan prostaatkanker. Tot op heden blijft het een uitdaging om nauwkeurig te voorspellen hoe de tumor zich zal gedragen. Een beter begrip van de ziekte op moleculair niveau en het gebruik van voorspellende biomarkers zouden het stratificeren tussen indolente en agressieve prostaatkankers kunnen verbeteren. In dit proefschrift hebben we de nadruk gelegd op het ontdekken van nieuwe kandidaat biomarkers ten behoeve van klinische risicostratificatie. Ook hebben we een methodologie voor doelgerichte 'next generation sequencing' (NGS) onderzocht dat gebruikt zou kunnen worden voor diagnose en stagering. Tezamen zouden onze bevindingen kunnen dienen als waardevolle middelen om patiëntenzorg te verbeteren en het aantal onnodige behandelingen te reduceren.

Hoofdstuk 1 omvat een algemene introductie over prostaatkanker en bijbehorende uitdagingen omtrent de detectie en stagering ervan. Ook beschrijven we de moleculaire karakteristieken en cellulaire signaaltransductiepaden die een rol spelen bij prostaatkanker. Voorts zetten we de technologische en bioinformatica ontwikkelingen van de laatste jaren uiteen, met de nadruk op NGS, een technologie die grote invloed op het prostaatkankeronderzoek heeft gehad en bovendien veelbelovend lijkt met betrekking tot klinische toepassingen. In **hoofdstuk 2** geven wij een uitgebreide literatuurstudie weer naar 'long non-coding RNAs' (lncRNAs) in urologische maligniteiten en hun potentie betreffende diagnose en stagering van prostaatkanker.

Om te onderzoeken of lncRNAs potentiële biomarkers in prostaatkanker zouden kunnen zijn, hebben we gebruik gemaakt van het 'Affymetrix Human Exon Array' platform dat de expressie van de bekende, maar ook onbekende genen weergeeft. In **hoofdstuk 3**, beschrijven we de ontdekking van een groot aantal nieuwe transcripten die in een deel van de prostaattumoren tot expressie komt, maar niet of zelden in de normale prostaat. Met behulp van aanvullende monsters uit publieke datasets waren we in staat 334 genomische kandidaatregio's, zogeheten 'EPCATs', te identificeren die alleen een signaal gaven in weefsel met prostaatkanker. Vervolgens hebben wij de top 20 EPCATs via RT-PCR expressie analyse getest en waren wij in staat 15 nieuwe prostaatkanker geassocieerde transcripten te bevestigen. Kwantitatieve RT-PCR validatie slaagde in 12 van deze transcripten, die we vervolgens combineerden in een genetisch panel met zeer sterke diagnostische potentie. Ook hebben we met behulp van *in situ* hybridisatie twee EPCATs in histologische coupes van patiëntenmateriaal gevisualiseerd en hun zeer specifieke expressiepatronen bevestigd. Wij hebben eveneens aangetoond dat twee EPCATS, beide gelokaliseerd op chromosoom 2, voorspellend waren voor progressie van de ziekte en ontwikkeling van uitzaaiingen. Als laatste hebben we coderende potentie van onze gevalideerde EPCATs onderzocht met behulp

van voorspellende programma's en geconcludeerd dat de EPCATs meest waarschijnlijk lncRNAs betreffen, zoals ook in eerste instantie werd voorgesteld.

In voorgaand onderzoek werd gesuggereerd dat expressie van de specifieke isovorm van cAMP-specifieke 3',5'-cyclische phosphodiësterase 4D (PDE4D) verlaagd is in cellijnen van agressieve prostaatkankers. Om die reden hebben wij vervolgens in **hoofdstuk 4** onderzocht of de expressie van PDE4D7 ook verlaagd was in humane weefsels van verschillende onafhankelijke datasets. Onze bevindingen bevestigen de associatie tussen PDE4D7 en tumorstadium, aangezien weefsels met niet-progressieve prostaatkankers een hogere expressie toonden dan de weefsels met progressieve ziekte en normaal controle weefsel. Wij vonden bovendien ook dat de PDE4D7 expressie verhoogd was in monsters met het TMPRSS2-ERG fusiegen vergeleken met de fusie-negatieve monsters en normaal controle weefsel. Daarom is PDE4D7 een potentiële biomarker in prostaatkanker en een mogelijk interessant therapeutisch doelwit.

Vanwege toenemend wetenschappelijk bewijs over de interactie tussen de androgeen receptor (AR) en de cAMP signaaltransductiepaden, hebben we in **hoofdstuk 5** onderzocht of de expressie van andere PDE4D isovormen geassocieerd was met prostaatkankerprogressie. Door gebruik te maken van verschillende onafhankelijke datasets vonden wij dat de expressie van PDE4D5 en PDE4D9 beide verlaagd waren in primaire prostaatkanker, waarbij we een isovormwisseling ontdekten bij de ontwikkeling van prostaatkanker. Om de verantwoordelijke moleculaire mechanismen op te helderen, hebben we gekeken of chromosomale deleties, transcriptiefactorbinding en DNA methyleringspatronen betrokken waren bij de regulatie van PDE4D. Wij vonden dat de expressieverlaging van PDE4D5 en PDE4D9 onafhankelijk van deleties in het genlocus gebeurde en dat PDE4D5 expressie verlaagd was in de LNCaP cellijn na androgeenstimulatie. ERG expressie lijkt geen effect te hebben op PDE4D5 en PDE4D9, terwijl PDE4D7 een toename van expressie liet zien in de weefsels met verhoogde ERG expressie en degene na androgeenstimulatie. Daarnaast konden we verscheidene loci identificeren met een toegenomen DNA methylering in prostaatkanker, waarvan één dichtbij de PDE4D5 promoter was gelokaliseerd. Op basis van het verkregen bewijs creëerden we twee genetische handtekeningen die gebaseerd waren op verschillende isovormen en evalueerden hun diagnostische en prognostische waarde in verscheidene onafhankelijke datasets. Als laatste lieten we zien dat onze diagnostische handtekening kan worden gebruikt om stagering van het naaldbiopt te verbeteren. Onze bevindingen laten eveneens zien dat PDE4D isovorm opbouw gedereguleerd is in prostaatkanker en benadrukken de relevantie van dit gen als biomarker en therapeutisch doelwit.

Alhoewel NGS extensief wordt gebruikt in het onderzoek ten behoeve van ontdekking en validatie, is diens klinische implementatie nog steeds beperkt. Om de vooruitgang van het gebruik van kanker-geassocieerde genetische afwijkingen in een klinische context te bevorderen, onderzochten wij in **hoofdstuk 6** de bioinformatica knelpunten van doelgerichte sequentieanalyse in genomen van patiënten. Aangezien standaard software voor het uitlijnen van NGS 'reads' ontworpen is voor genoom-wijde analyses, leidt een reductie van de complexiteit, door slechts alleen aan de doelgerichte regio van interesse uit te lijnen, tot een hoge 'mapping' score, doordat vele 'reads' gedwongen worden uitgelijnd. Hierdoor krijgt

men zeer veel vals positieve uitslagen in het afroepen van een mutatie (88% extra afgeroepte mutaties, waarvan 92% vals positief). Om die reden implementeerden wij een nieuwe benadering dat gebruik maakte van *a priori* informatie van een doelgerichte sequentie techniek om zo in een korter tijdsbestek 'reads' uit te lijnen en minder gebruik te hoeven maken van computergeheugen vergeleken met bestaande methoden. Deze methodologie laat een zeer efficiënte verwerking van NGS data toe dat waardevolle genetische informatie van een patiënt zijn tumor kan geven met betrekking tot op maat gemaakte behandelstrategieën en gepersonaliseerde zorg.

Concluderend voorziet dit proefschrift van enkele nieuwe biomarkers voor prostaatkanker die van waarde zouden kunnen zijn in de discriminatie tussen indolente en agressieve ziekte. De EPCATs tonen een veelbelovende rol als specifieke lncRNAs, terwijl PDE4D deregulatie een aanwijzing is voor bredere afwijkingen in de cAMP signaaltransductie bij prostaatkanker. Dit proefschrift bevordert eveneens de klinische applicatie van doelgerichte sequentie analyse door het voorzien van efficiënte middelen om de computationele last te reduceren bij het analyseren van data met *a priori* kennis.

Curriculum vitae

René Böttcher was born on the 22nd of October 1986 in Berlin-Steglitz, Germany. He completed his secondary education at Goethe Oberschule (Gymnasium) in Berlin-Steglitz in 2006, with majors in Biology and English. After high school graduation, René began his studies Biosystemstechnology / Bioinformatics at Technical University of Applied Sciences Wildau. He obtained his Bachelor's degree in 2009 after writing his Bachelor's thesis on Bacterial Nanocellulose as potential scaffold material in orthopedic applications at TransTissue Technologies GmbH Berlin. Having continued his education in Wildau, René decided to do an internship at Philips Research Eindhoven under supervision of Dr. Pieter Jan van der Zaag in January 2011. The project focused on the simulation of a novel targeted sequencing technology and development of appropriate alignment software. After finalizing the internship in summer 2011, René moved to Rotterdam to work on his Master's thesis under supervision of Prof. Dr. Guido Jenster at Erasmus Medical Center. René graduated his Master's with honors in 2012 and decided to continue his bioinformatics-driven research on prostate cancer biomarkers in Dr. Jenster's group to obtain his PhD. After graduation in 2016, René joined the groups of Prof. Dr. Francesc Posas (Cell signaling) and Prof. Dr. Juana Diez (Virology) at Universitat Pompeu Fabra in Barcelona, Spain. There, his expertise in next generation sequencing is applied to projects related to cellular stress signaling and emerging viruses.

List of publications

1 **Böttcher R**, Amberg R, Ruzius FP, Guryev V, Verhaegh WF, Beyerlein P, van der Zaag PJ. Using a priori knowledge to align sequencing reads to their exact genomic position. Nucleic Acids Res. 2012;40(16):e125. (this thesis)

2 Martens-Uzunova ES, **Böttcher R**, Croce CM, Jenster G, Visakorpi T, Calin GA. Long noncoding RNA in prostate, bladder, and kidney cancer. Eur Urol. 2014;65(6):1140-51. (this thesis)

3 Teles Alves I, Hartjes T, McClellan E, Hiltemann S, **Böttcher R**, Dits N, Temanni MR, Janssen B, van Workum W, van der Spek P, Stubbs A, de Klein A, Eussen B, Trapman J, Jenster G. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. Oncogene. 2015;34(5):568-77.

4 **Böttcher R**, Hoogland AM, Dits N, Verhoef EI, Kweldam C, Waranecki P, Bangma CH, van Leenders GJ, Jenster G. Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. Oncotarget. 2015;6(6):4036-50. (this thesis)

5 **Böttcher R**, Henderson DJ, Dulla K, van Strijp D, Waanders LF, Tevz G, Lehman ML, Merkle D, van Leenders GJ, Baillie GS, Jenster G, Houslay MD, Hoffmann R. Human phosphodiesterase 4D7 (PDE4D7) expression is increased in TMPRSS2-ERG-positive primary prostate cancer and independently adds to a reduced risk of post-surgical disease progression. Br J Cancer. 2015;113(10):1502-11. (this thesis)

6 Hoogstrate Y, **Böttcher R**, Hiltemann S, van der Spek PJ, Jenster G, Stubbs AP. FuMa: reporting overlap in RNA-seq detected fusion genes. Bioinformatics. 2016;32(8):1226-8.

7 Erdem-Eraslan L, van den Bent MJ, Hoogstrate Y, Naz-Khan H, Stubbs A, van der Spek P, **Böttcher R**, Gao Y, de Wit M, Taal W, Oosterkamp HM, Walenkamp A, Beerepoot LV, Hanse MC, Buter J, Honkoop AH, van der Holt B, Vernhout RM, Smitt PA, Kros JM, French PJ. Identification of Patients with Recurrent Glioblastoma Who May Benefit from Combined Bevacizumab and CCNU Therapy: A Report from the BELOB Trial. Cancer Res. 2016;76(3):525-34.

8 Schewe M, Franken PF, Sacchetti A, Schmitt M, Joosten R, **Böttcher R**, van Royen ME, Jeammet L, Payré C, Scott PM, Webb NR, Gelb M, Cormier RT, Lambeau G, Fodde R. Secreted phospholipases IIA and X are stem cell niche factors with context-dependent roles in intestinal homeostasis, inflammation and cancer. Cell Stem Cell. 2016;19(1):38-51.

9 de Morree E, **Böttcher R**, van Soest RJ, Aghai A, de Ridder CM, Gibson AA, Mathijssen RH, Burger H, Wiemer EA, Sparreboom A, de Wit R, van Weerden WM. Loss of SLCO1B3 drives taxane resistance in prostate cancer. Br J Cancer. 2016;115(6):674-81.

10 Hoogland AM, **Böttcher R**, Verhoef EI, Jenster G, van Leenders GJLH. Gene-expression analysis of gleason grade 3 tumor glands embedded in low- and high-risk prostate cancer. Oncotarget. 2016. doi: 10.18632/oncotarget.9344. [Epub ahead of print]

11 **Böttcher R**, Dulla K, van Strijp D, Dits N, Baillie GS, van Leenders GJLH, Houslay MD, Jenster G, Hoffmann R. Human PDE4D isoform composition is deregulated in primary prostate cancer and indicative for disease progression and development of distant metastases. Oncotarget. 2016. doi: 10.18632/oncotarget.12204. [Epub ahead of print]. (this thesis)

12 Studer RA, Rodriguez-Mias RA, Haas KM, Hsu JI, Viéitez C, Solé C, Swaney DL, Stanford LB, Ivan Liachko I, **Böttcher R**, Maitreya J. Dunham MJ, de Nadal E, Posas F, Beltrao P, Judit Villén J. Evolution of protein phosphorylation across 18 fungal species. Science. Accepted for publication.

13 Teles Alves I, Cano D, **Böttcher R**, van der Korput H, Dinjens W, Jenster G, Trapman J. A mononucleotide repeat in PRRT2 is an important, frequent target of mismatch repair deficiency in cancer. Submitted for publication

14 Teles Alves I, **Böttcher R**, van Royen ME, Dits N, Trapman J, Jenster G. The GPS2-MPP2 gene fusion promotes growth and decreases apoptosis in the LNCaP cell line. Submitted for publication

PhD Portfolio

Name PhD student	René Böttcher
Erasmus MC Department	Urology
Research School	Molecular Medicine
PhD Period	May 2012 to October 2016
Promotor	Prof. Dr. Guido W. Jenster
Copromotor	Prof. Dr. Peter Beyerlein

PhD training

General and specific courses (ECTS)

2011	Next Generation Sequencing Training: CLC Bio (0.5)
2012	Workshop Browsing Genes and Genomes with UCSC Browsing Genes and Genomes with UCSC: Advanced Workshop (0.8)
2012	Workshop on Photoshop and Illustrator CS5 for PhD-students and other researchers (0.3)
2013	Biomedical English Writing Course for MSc and PhD-students (2.0)
2014	The Ensembl Workshop X (0.6)

Seminars and scientific meetings

2011 – 2014	Annual CTMM-PCMM meeting
2012 – 2015	Journal Club Urology
2012 – 2015	Monthly MolMed Bridge Meetings
2012 – 2015	JNI Scientific Meetings

Teaching and supervision

2014	4th RNA-seq course – Leiden UMC (NLD) Fusion gene detection in cancer
2012	Fedde Schaeffer (Hogeschool Leiden – Bachelor thesis)
2013	Nikolas Strepis (Wageningen University – Master thesis)
2013	Chris van der Lans (Hogeschool Leiden – Bachelor thesis)
2015	Mónica Vara Perez (Erasmus MC – Master thesis)
2015	Wendy Plugge (Wageningen University – Master thesis)
2015	David Nieuwenhuijse (Wageningen University – Master thesis)

Presentations

2012	7th Benelux Bioinformatics Conference (poster presentation)
2013	8th Netherlands Bioinformatics Conference (poster presentation)
2013	20th MGC PhD Workshop Luxembourg (poster presentation)
2013	2nd Prostate Cancer Translational Research in Europe meeting (poster presentation)
2013	3rd Galaxy Community Conference (poster presentation)
2013	11th Heinrich-Warner Foundation Symposium Hamburg (oral presentation)

2013 21st Meeting of EAU Section of Urological Research (oral presentation)
2014 18th Molecular Medicine Day (poster presentation)
2014 5th Beyond the Genome: Cancer genomics (poster presentation)

International conferences

2012 7th Benelux Bioinformatics Conference – Nijmegen (NLD)
2013 8th Netherlands Bioinformatics Conference – Lunteren (NLD)
2013 2nd Prostate Cancer Translational Research in Europe – Malmö (SWE)
2013 3rd Galaxy Community Conference – Oslo (NOR)
2013 21st EAU Section of Urological Research – Dresden (GER)
2014 5th Beyond the Genome: Cancer genomics – Boston (USA)

Awards

2013 ESUR Travel Award
2014 Winner of Beyond the Genome: Cancer genomics Informatics Challenge

Acknowledgements

Firstly, I would like to thank my advisor Prof. Dr. Guido Jenster, whose continued support and guidance helped me during the final internship of my Master's as well as all the work for this thesis. Guido, I am truly grateful for all the advice you have provided me with over the years and I really enjoyed the time spend Rotterdam, so as we say in German, it is "with a smile in one eye and a tear in the other" that I am setting sail for new adventures.

Next, I would like to thank my co-promoter Prof. Dr. Peter Beyerlein as well as the rest of my thesis committee: Prof. Dr. Peter van der Spek, Prof. Dr. Reuven Agami, and Dr. Gabri van der Pluijm, for their insightful comments and questions.

My sincere thanks also go to Dr. Ralf Hoffmann, whom I closely collaborated with during the past years and whose determination and support led us to the publication of two terrific research articles.

I thank Dr. Arno van Leenders, whose expertise in pathology enlightened me in many of our shared projects, and Dr. Paul C. Boutros, who gave me the opportunity to join his team during a collaborative project and provided excellent statistics and bioinformatics advice.

I also would like to thank the members of the Urology Department, especially Elena Martens and Natasja Dits, for all the help, stimulating discussions, cakes and beers, as well as the various students I supervised who, over the years, drilled me with questions and taught me the value of patience and persistence.

As members of other departments, I would like to thank Saskia Hiltemann and Andrew Stubbs for the good times in bioinformatics, Charlotte Kweldam and Marije Hoogland for the good times mixing pathology and bioinformatics, as well as Youri, Thomas, Matthias, Tommaso, Caterina, Lucie, David and Antoine for their friendship and all the fun we have had in the last four years.

Last but not the least, I am blessed that my girlfriend, my family and friends at home for supported me spiritually throughout writing this thesis and continue to do so after. Thank you all.