Title:

The genes controlling normal function of citrate and spermine secretion is lost in aggressive prostate cancer and prostate model systems

Morten Beck Rye[*,1,2,3,4]*, Sebastian Krossa*[5]*, Martina Hall*[6,7]*, Casper van Mourik*[1,8]*, Tone F. Bathen*[5]*, Finn Drabløs*[1]*, May-Britt Tessem*[2,5,9] *and Helena Bertilsson*[1,2,9]

[1] Department of Clinical and Molecular Medicine, NTNU – Norwegian University of Science and Technology, P.O. Box 8905, NO-7491, Trondheim, Norway

[2] Clinic of Surgery, St.Olavs Hospital, Trondheim University Hospital, NO-7030 Trondheim, Norway

[3] Clinic of Laboratory Medicine, St.Olavs Hospital, Trondheim University Hospital, NO-7030 Trondheim, Norway.

[4] BioCore - Bioinformatics Core Facility, NTNU – Norwegian University of Science and Technology, P.O. Box 8905, NO-7491, Trondheim, Norway

[5] Department of Circulation and Medical Imaging, NTNU - Norwegian University of Science and Technology, Trondheim.

[6] Department of Biotechnology and Food Science, NTNU – Norwegian University of Science and Technology, Trondheim.

[7] K. G. Jebsen Center for Genetic Epidemiology, NTNU – Norwegian University of Science and Technology, Trondheim.

[8] Institute for Life Science & Technology, Hanze University of Applied Sciences, 9747 AS Groningen, The Netherlands

[9] These authors contributed equally

*Corresponding author

1 **Abstract**

2 **Background:** Secretion of the metabolites citrate and spermine into prostate lumen is a unique
3 hallmark for normal prostate epithelial cells. However, the identity of the genes controlling citrate
4 and spermine secretion remains mostly unknown despite their obvious relevance for progression to
5 aggressive prostate cancer.
6 **Materials & Methods**: In this study, we have correlated simultaneous measurement of
7 citrate/spermine and transcriptomics data. We have refined these gene correlations in 12 prostate
8 cancer cohorts containing 2915 tissue samples to create a novel gene signature of 150 genes
9 connected with citrate and spermine secretion. We further explored the signature in public data,
10 interrogating over 18 000 samples from various tissues and model systems, including 3826 samples
11 from prostate and prostate cancer.
12 **Results:** In prostate cancer, the expression of this gene signature is gradually lost in tissue from
13 normal epithelial cells through PIN, low grade (Gleason <= 7), high grade cancer (Gleason >= 8) and
14 metastatic lesions. The accuracy of the signature is validated by its unique enrichment in prostate
15 compared to other tissues, and its strong enrichment in epithelial tissue compartments compared to
16 stroma. Several zinc-binding proteins that are not previously investigated in the prostate are present
17 in the gene signature, suggesting new mechanisms for controlling zinc homeostasis in
18 citrate/spermine secretion. However, the absence of the gene signature in all common prostate
19 normal and cancer cell-lines, as well as prostate organoids, underlines the challenge to study the
20 role of these genes during prostate cancer progression in model systems.
21 **Conclusions:** A large collection of transcriptomics data integrated with metabolomics identifies the
22 genes related to citrate and spermine secretion in the prostate, and show that the expression of
23 these genes gradually decreases on the path towards aggressive prostate cancer. In addition, the
24 study questions the relevance of currently available model systems to study metabolism in prostate
25 cancer development.

26

27

28

29

30

31

32

33

34

35

36

37

38

39

1   **Introduction**

2   A unique hallmark for normal prostate epithelial cells is their ability to secrete large amounts of the
3   metabolites citrate. While cells from other human organs convert citrate to isocitrate for oxidation
4   and energy production in the TCA cycle, citrate conversion in the prostate epithelium is partly blocked,
5   leading to accumulation of citrate [1]. It has been reported that citrate/isocitrate ratios in the prostate
6   are 40/1, compared to 10/1 in other normal organs [2]. In the prostate, excess citrate is secreted out
7   of the cell into the prostate lumen of the prostatic glands, where it is a major constituent of the
8   prostatic fluid and essential for normal prostate organ function.

9   Almost all types of prostate cancer originates from prostate epithelial cells, and nearly all prostate
10   cancer cells lose their ability to secrete citrate during the development of malignant phenotypes [1].
11   A reduced level of citrate has therefore been suggested as a biomarker, both for cancer detection and
12   for identification of cancers with more aggressive phenotypes [3-5]. Even with these demonstrations
13   of importance for normal prostate function and prostate cancer progression, the research in this field
14   has been markedly neglected, and thus the molecular mechanisms controlling these functions remain
15   mostly unknown. Nevertheless, a few dedicated research groups have managed to gain some insights:
16   First, high mitochondrial levels of zinc in the prostate block citrate conversion, and the regulation of
17   zinc transporter genes, in particular ZIP1 (*SLC39A1*), have been suggested to play a role in regulating
18   prostate zinc levels [6, 7]. Second, the metabolite aspartate is a possible source of increased citrate in
19   prostate cells [8, 9]. Aspartate is a precursor for oxaloacetate, which again is a precursor of citrate
20   synthesis, and works as a carrier of the carbon groups oxidized stepwise in the mitochondria during
21   the TCA cycle. Increased levels of aspartate could thus lead to increased levels of citrate when the
22   citrate/isocitrate conversion is blocked. Third, metabolites synthesized in the polyamine pathway are
23   regulated differently in the prostate [10-12]. Some interest has been directed to the polyamine
24   spermine, due to its extremely high correlation with citrate [13]. Spermine and citrate have opposite
25   polarity, and it has been speculated as to whether they may form a complex which is secreted into the
26   prostate lumen [13, 14].

27   These presented mechanisms, though intriguing, represent only small steps on the way to a widened
28   perspective of this unique prostate function which is not completely understood. This includes the
29   genes and proteins involved, as well as the connections between them. Most mechanisms suggested
30   so far have been studied in cell-lines and animal models, which are not able to fully capture the
31   complex interplay within human prostate tissue. However, a large amount of genome-wide –omics
32   data on prostate tissue, as well as their model systems, has accumulated since the introduction of the
33   genomic era more than 20 years ago. These data resources have yet to be fully exploited for research
34   on citrate and spermine secretion in the prostate. In this study we address this hallmark property of
35   the prostate by performing extensive bioinformatic analysis on data resources from the public
36   domain. We do this by first creating and a robust gene signature representing citrate and spermine
37   secretion in the prostate. We then explore and validate this gene signature in multiple datasets,
38   cohorts and public resources. In total we analyzed 32 datasets [15-42] with 18020 samples, of which
39   3826 were prostate samples including normal/normal-adjacent (epithelium and stroma) tissue,
40   high/low-grade cancer tissue, metastatic samples and model systems. Our study identifies genes
41   strongly associated with citrate and spermine secretion in the prostate and explore their behavior
42   during cancer progression and in model systems. Our results validate previously suggested
43   mechanisms, but also discover new central genes and mechanisms related to citrate and spermine
44   with potential importance for this unique and intriguing prostate hallmark.

45

1   **Results**

2   All datasets used in this study are listed in Supplementary File 1. It includes dataset ID and abbreviation
3   used in the main text, sample description, dataset accession and references.

4   **Metabolite concentrations between citrate and spermine are highly correlated across tissue**
5   **samples**

6   Our research group has previously generated a dataset with 129 normal and cancer tissue samples
7   from the prostate, where concentrations from 23 metabolites and the expression of 14161 unique
8   genes were measured on the exact same samples (*Bertilsson,* dataset ID 1). Similar to a previous report
9   [12], we observed a strong correlation (r=0.95) between concentrations of the metabolites citrate and
10   spermine across the samples (Figure 1A and B). To simplify calculations, we will use the average
11   concentration profile for citrate and spermine throughout the rest of this study. (Figure 1C). We use
12   the abbreviation CS (Citrate and Spermine) to denote all our results based on this average
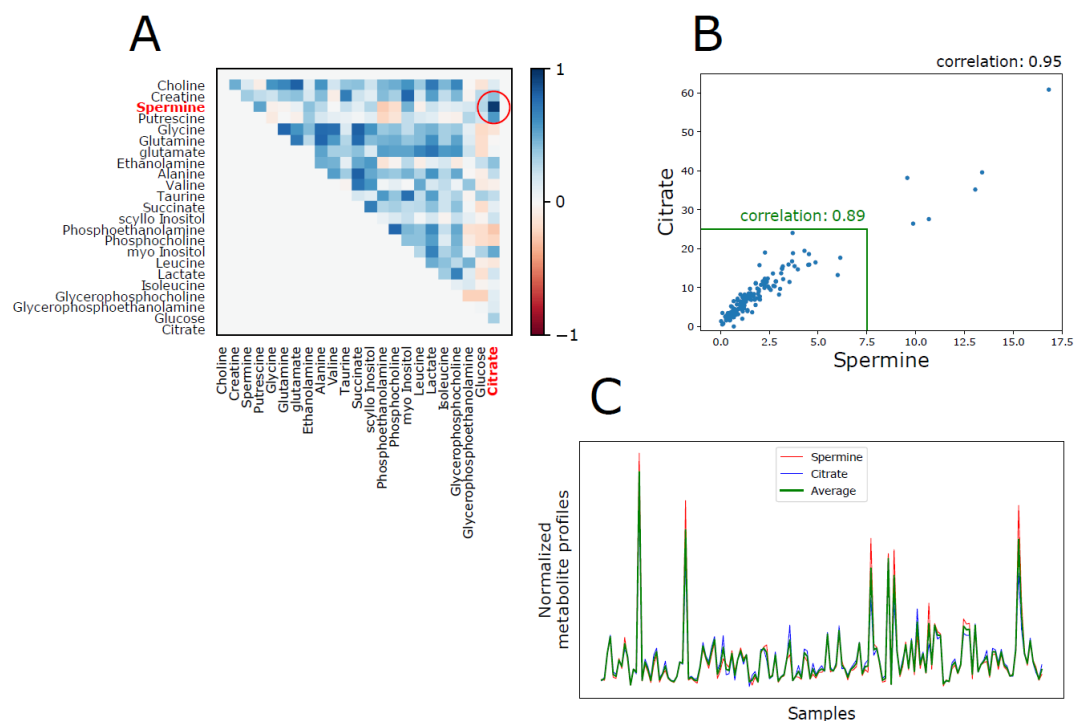13   concentration profile.

14   *Figure 1: Correlations between metabolites citrate and spermine. A) Correlations between all 23*
15   *metabolite measurements across 129 prostate normal and cancer samples from the Bertilsson*
16   *dataset (Dataset ID 1). The correlation is particularly strong for metabolites citrate and spermine. B)*
17   *Correlations between citrate and spermine across 129 prostate normal and cancer samples from*
18   *Bertilsson. The correlation is high also when extreme values are excluded (highlighted inside the*
19   *green boundary). C) Similarity between the citrate, spermine and the average citrate/spermine*
20   *metabolite profiles across 129 prostate normal and cancer samples from Bertilsson.*

21

22

23

24

4

A gene signature associated with the CS concentrations is validated in 12 datasets of prostate cancer and normal samples.

We identified genes associated with the average CS levels in the prostate by calculating the Pearson correlations between the CS concentration profile and each of the 14161 unique genes measured in *Bertilsson* (dataset ID 1)*.* The correlation analyses were performed on the 95 cancer samples in the dataset. We selected the 150 genes with the highest positive correlation which we termed the initial gene module. We further assessed the integrity of this gene module by creating a Correlation Module Score (CMS, see methods) where a high CMS means that strong intra-correlation is present between the genes in the module, resulting in a high module integrity. On the contrary, a low CMS indicates that there is low correlation between the genes in the module, and that the module integrity is lost. Moreover, if the gene module has a significantly high CMS in a dataset, it indicates that the biological process(es) which this module represents is functionally active for the samples in this dataset.
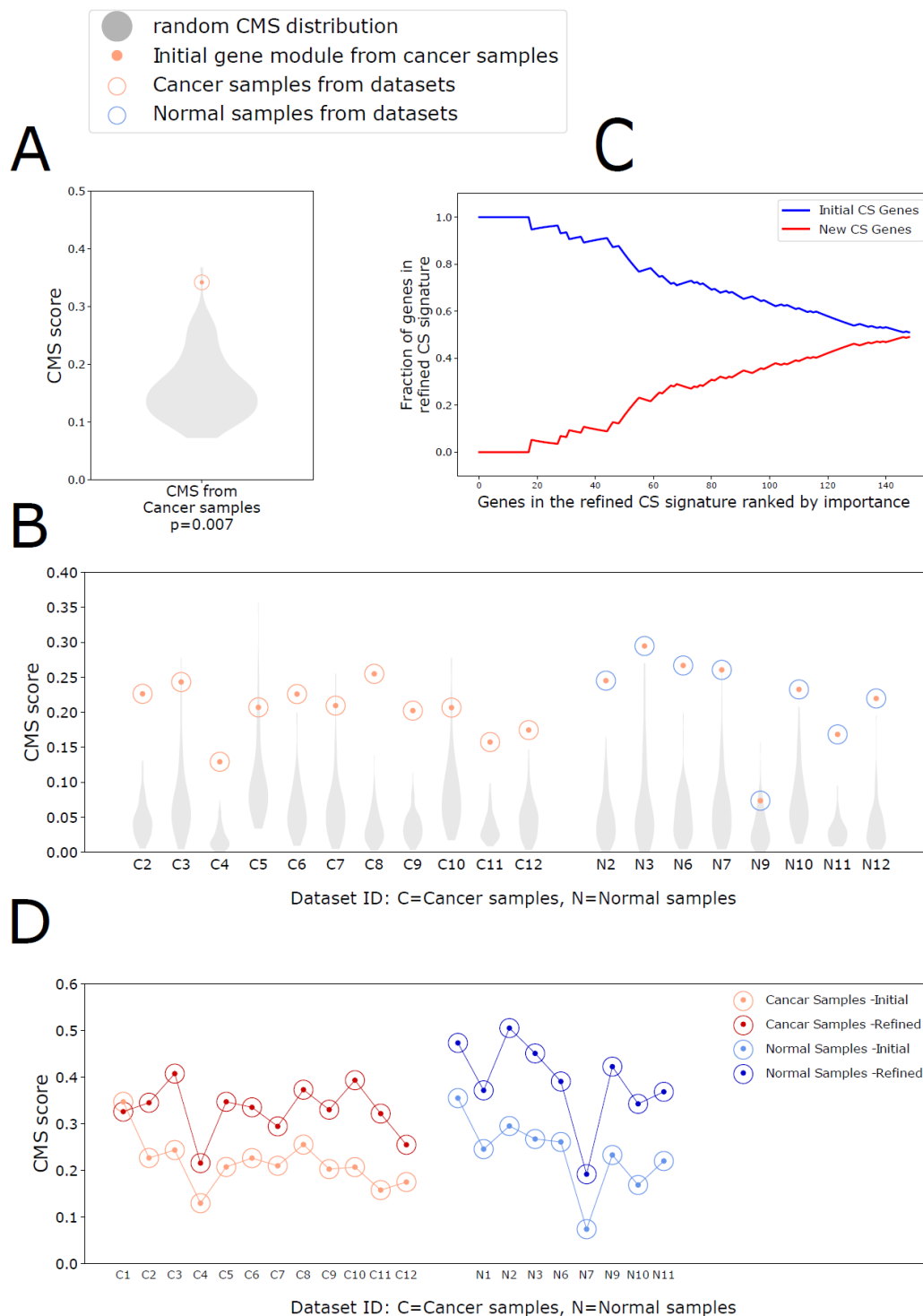
To test for significance, we generated 100 random CMSs by shuffling the CS metabolite levels between the respective samples in *Bertilsson* (dataset ID 1). The resulting CMS was statistically significant when compared to CMSs from random gene modules (CMS=0.34, p=0.007, lognormal distribution) (Figure 2A).

We then validated whether the integrity of the initial gene module was significantly preserved in 11 additional datasets (dataset ID 2-12), including a total of 2638 tissue samples from normal prostate and prostate cancer (ref dataset list). We calculated and compared the CMSs for the initial module, as well as the 100 random modules in all 11 datasets. The CMS from the initial gene module was validated in all 11 cancer datasets, and 7 out of 8 datasets with normal samples (lognormal test) (Figure 2B, Supplementary Table 1 and 2). We thus defined the initial module as our initial gene signature associated with CS concentrations in the prostate. We used this initial CS signature as a starting point for further refinement.

We also performed the above procedure using the 40 normal samples in *Bertilsson,* however the resulting gene module was not significant (CMS=0.37, p=0.14, lognormal distribution), and could not be validated in any of the additional datasets (Supplementary Figure 1, Supplementary Table 3 and 4). The reason for this is probably due to an insufficient number of normal samples in *Bertilsson* to perform a robust correlation analysis.

Refinement across 12 datasets produce a CS gene signature with improved integrity.

Corresponding gene expression and metabolite measurements are available only for *Bertilsson* (dataset ID 1). We thus implemented a bioinformatics strategy to identify a refined CS gene signature with improved integrity across all 12 datasets (2794 samples). We used the initial CS signature from *Bertilsson* as a starting point, and then used results from cancer samples across the other 11 datasets to nominate better gene candidates to replace genes from the initial CS signature (see Methods). When evaluating the new top 150 candidate genes, we found that 74 of the 150 genes in the initial CS signature had been replaced by new genes to improve integrity across all datasets (Figure 2C). Of note, the refined CS signature improved CMS for in both cancer and normal samples in all 11 additional datasets. This consistent improvement in normal samples (including normal samples from *Bertilsson*) affirms the relevance of the signature, since these samples were not used to generate or refine the CS signature. In addition, only a marginal reduction in CMS was observed for the cancer samples in *Bertilsson* (Figure 2D). We thus conclude that the refined CS signature (referred to as the CS signature from now) represents a more robust signature with improved integrity compared to the initial CS signature.

1

Figure 2: Integrity of initial and refined citrate – spermine (CS) gene signatures. *A) CMS (Correlation Module Scores) for the 150 genes most positively correlated with the CS metabolite profile for cancer samples in the Bertilsson dataset (Dataset ID 1). The CMS for was statistically significant when compared to CMSs using random metabolite profiles (lognorm test) B) CMS for the initial cancer CS gene module from Bertilsson (red dots) evaluated in prostate cancer (red circles) and normal (blue circles) samples from 11 additional datasets. The CMS was statistically significant in all datasets (lognorm test, Supplementary Table T1 and T2). C) Fraction of new genes in the refined CS gene*

1 *signature that replaced genes in the initial CS gene signature. The genes is sorted by importance*
2 *(rank) on the bottom axis. Of the 150 top ranked genes in the initial CS signature, 74 were replaced*
3 *in the refined CS signature. D) Increase in CMS after refinement of the initial CS signature for prostate*
4 *normal and cancer samples in 11 datasets. The CMS increased in the normal samples, even though*
5 *the normal samples were not used to create or refine the signature. Cohort ID 4, 5 and 8 did not*
6 *include normal samples.*

7

8 **The CS signature displays unique enrichment in prostate samples and negative correlation to**
9 **stromal tissue compartments.**

10 Having established a CS signature of 150 genes associated with citrate and spermine secretion in the
11 previous sections, we expanded our validation to additional datasets and sample types. For this
12 analysis we used single samples Gene Set Enrichment Analysis (ssGSEA) [43], which is a method used
13 to assess whether a specific gene signature is enriched or depleted in a single sample. Since citrate
14 and spermine secretion is a highly specific function associated with prostate epithelium, higher ssGSEA
15 scores in tissue samples from prostate compared to samples from other tissue types would confirm
16 its functional relevance. Further, since prostate cancer samples are a heterogeneous tissue mixture
17 of normal prostate epithelium, cancer and stroma, ssGSEA scores should be inversely proportional to
18 the content of stroma tissue in the samples, since stromal cells are unable to secrete citrate or
19 spermine.

20 We first compared ssGSEA scores for cancer and normal samples in the prostate specific *TCGA* dataset
21 (dataset ID 5) to all cancer and normal samples profiled in the *TCGA-complete* resource (11093
22 samples from 33 cancer types, dataset ID 21). We observed strongly elevated ssGSEA scores for both
23 prostate cancer and normal samples compared to cancer and normal tissues from other tissues (Figure
24 3A). We also tested the CS signature on average expression profiles for 53 normal tissues from the
25 *GTEx* portal (dataset ID 22) (Figure 3B) and 1829 CAGE expression profiles from cell-lines and tissues
26 in the *FANTOM* Consortium (dataset ID 23) (Figure 3C) where one tissue profile was from adult normal
27 prostate. In both datasets, normal prostate tissue showed the highest ssGSEA scores.

28 We next correlated the CS signature ssGSEA scores to a previously identified gene signature for stroma
29 content in prostate tissue [44] in datasets 1-12. We observed a strong reverse correlation between CS
30 and stroma signatures for normal samples in all datasets (Figure 3D) (with an exception for *Mortensen*
31 *(*dataset ID 9) which contained laser dissected epithelium/cancer with minimal amounts of stroma).
32 The correlation was weaker in cancer samples. This is expected, since prostate cancers vary in their
33 ability to secrete citrate and spermine. The negative association between CS signature and stroma was
34 also confirmed using spatial transcriptomics data from Berglund (dataset ID 20), where pixels with
35 high and low ssGSEA scores overlaid regions with epithelium/lumen and stroma respectively (Figure
36 3E, Supplementary Figure 2).

1   In summary, the CS signature shows strong associations with expected properties of the prostate
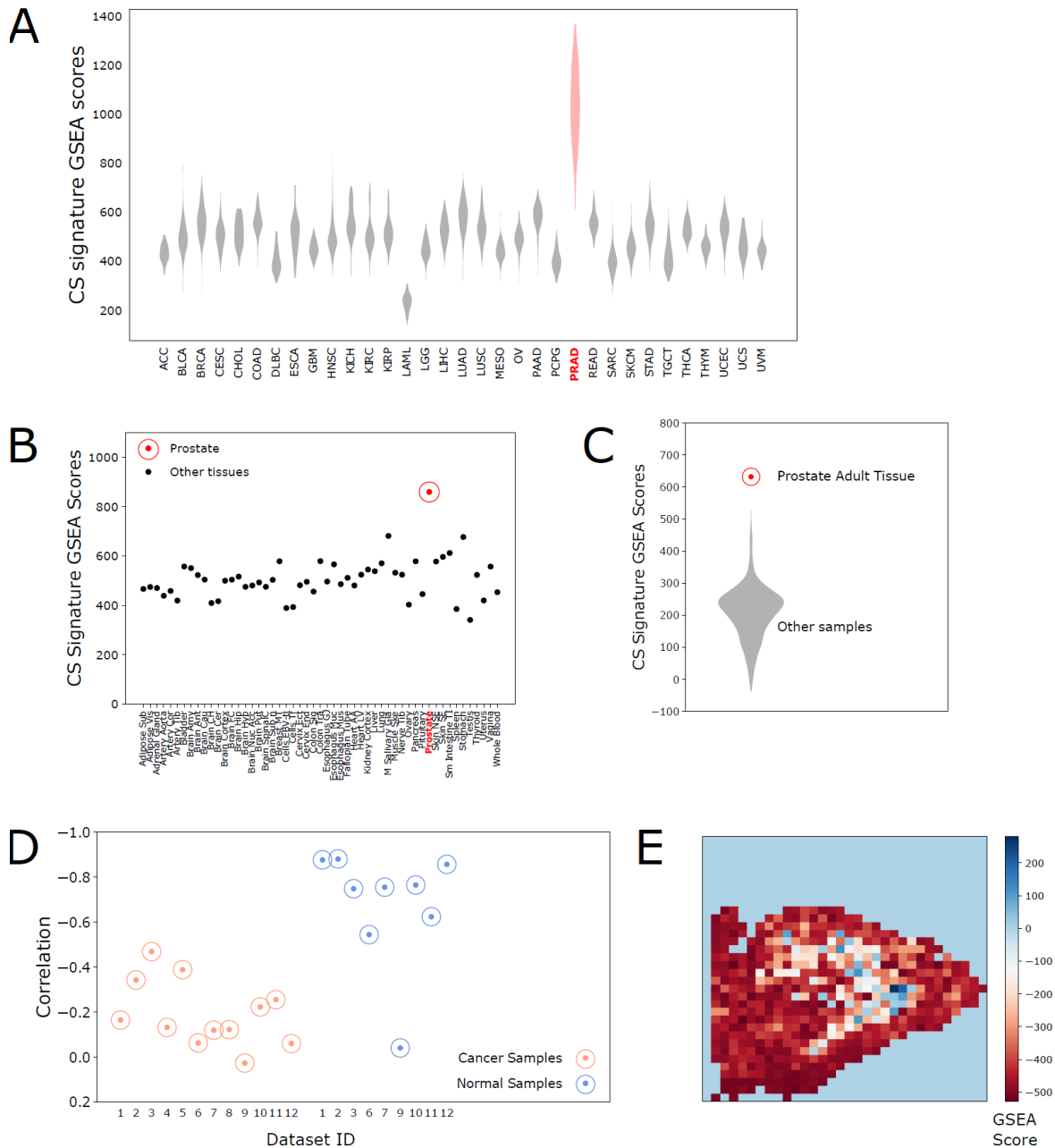2   epithelium, which strengthens the biologically validity of the genes in the signature.

3

Figure 3: Validation of citrate – spermine (CS) gene signature. *A) CS signature ssGSEA scores for 11093 cancer and normal samples in 33 tumor types from the TCGA-complete dataset. We have used standard TCGA cancer type abbreviations [https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations] where PRAD is the abbreviation for prostate cancer. B) CS signature ssGSEA scores for 53 averaged human normal tissue profiles from the GTex dataset. C) CS ssGSEA signature scores for 1829 human cell-line and tissue samples, including one prostate adult tissue sample, from the FANTOM dataset. D) Correlations between CS signature and stroma signature ssGSEA scores for cancer and normal samples in prostate cancer tissue datasets 1-12. The negative correlations in the normal samples are stronger due to less confounding from cancer tissue. Data*

8

1 *from Mortensen (dataset ID 9) contains laser dissected prostate normal epithelium and cancer, which*
2 *probably lack significant amounts of stroma. E) CS signature ssGSEA scores on spatial*
3 *transcriptomics data from tissue slice 3.2 in Berglund (dataset ID 20). The colorbar indicates*
4 *enrichment of the ssGSEA score in each pixel. The corresponding pathological tissue image can be*
5 *found for image 3.2 in Supplementary Information – Supplementary Figure 1b from Berglund et al.*
6 *[33], and comparison shows a strong overlap of high and low CS signature scores with*
7 *epithelial/lumen and stroma tissue compartments, respectively. Results from all 12 images in*
8 *Berglund are shown in Supplementary Figure 2.*

9

10 **The CS gene signature is gradually lost from normal samples through low grade, high grade and**
11 **metastatic lesions.**

12 Having established the relevance of the CS signature to prostate citrate and spermine secretion, we
13 next investigated how this signature associates with the various stages of prostate cancer
14 aggressiveness. Low citrate concentrations have previously been associated with high-grade prostate
15 cancer [4]. We identified nine datasets (1779 prostate cancer samples) where cancer samples had
16 been classified as either high-grade or low-grade, or assigned a Gleason score - the standard form of
17 prostate cancer grading. We found significant changes in CS signature ssGSEA scores between high-
18 and low-grade cancers in all eight datasets where samples where Gleason score 4+3 was classified as
19 high-grade (613 high and 737 low-grade samples), and six out of eight datasets where samples with
20 Gleason score 4+3 was classified as low-grade (578 high- and 1258 low-grade samples) (Figure 4A,
21 Table 1). The reduction in CS signature scores were not due to increased tumor fraction in high-grade
22 samples (shown in Table 1. These results support the previous findings and suggest that the changes
23 previously observed for citrate at metabolite level are accompanied by changes in gene expression. A
24 further significant reduction in CS signature scores were observed in metastatic compared to cancer
25 and normal samples in 8 datasets (157 metastatic samples) (Figure 4B, Table 2). Note that the lower
26 CS-signature scores in normal samples compared to cancer samples is due to the high level of stromal
27 tissue in normal prostate samples [8, 44] (Supplementary Figure 3). Overall, our results indicate that
28 the genes associated with citrate and spermine secretion is upregulated in the normal epithelium, and
29 then gradually downregulated with tumor progression from early low-grade lesions, through high-
30 grade cancers and finally metastatic tumors. This trajectory is nicely illustrated by the results from
31 *Tomlins* (dataset ID 18) (Figure 4C), where laser dissection has been used to purify stroma, normal
32 epithelium, Prostate Intreapeithelial Neoplasia (PIN, an early pre-cancerous prostate lesion) cancer
33 and metastatic tissues. In line with this trajectory, the PIN lesions show an average CS signature score
34 level between the levels of normal epithelium and cancer. Though the ability of cells to secret citrate
35 and spermine seems to be lost in metastatic samples, metastatic tissue from prostate cancer still
36 contains traces of the CS gene profile compared to metastatic tissue originating from other organs
37 (*Hsu,* dataset ID 19) (Figure 4D).

38

39

40

41

42

1 ***Table 1: P-values from differential CS signature ssGSEA score analysis between high grade (HG)***
2 ***and low grade (LG) cancer samples in 9 datasets.*** *Two comparisons are made: First HG1 vs LG1,*
3 *where HG1 is classified as samples with Gleason score 4+3 or higher, and LG1 as samples with*
4 *Gleason score 3+4 or lower. Second HG2 vs LG2, where HG2 is classified as samples with Gleason*
5 *score 8 or higher, and LG2 as samples with Gleason score 7 or lower. The correlation of CS signature*
6 *GSEA score to tumor content is also included when available, and show, in general, that decreased CS*
7 *scores in high-grade samples is not due to increased tumor content in these samples.*

Dataset ID	Dataset abbreviation	Number of HG1 samples	Number of LG1 samples	Number of HG2 samples	Number of LG2 samples	P-value HG1 vs LG1	p-value HG2 vs LG2	Correlation to tumor content
1	Bertilsson	56	60	36	80	4.7e-10	3.6e-5	-0.13
2	Chen	15	50	11	54	0.0040	0.0011	0.14
3	Taylor	36	94	15	115	0.015	0.010	NA
4	Sboner	119	162	81	200	3.7e-7	1.7e-4	-0.30
5	Erho	NA	NA	211	334	NA	0.060	NA
6	TCGA	296	200	200	296	2.5e-10	5.3e-13	-0.04
7	CMBR	30	82	9	103	0.014	0.09	0.11
8	STCK	34	57	15	76	0.021	0.007	NA
12	Kuner	27	32	NA	NA	0.021	NA	NA

8

9 ***Table 2: P-values from comparing CS signature ssGSEA scores from metastatic prostate cancer***
10 ***samples to cancer and normal samples and in 8 datasets.***

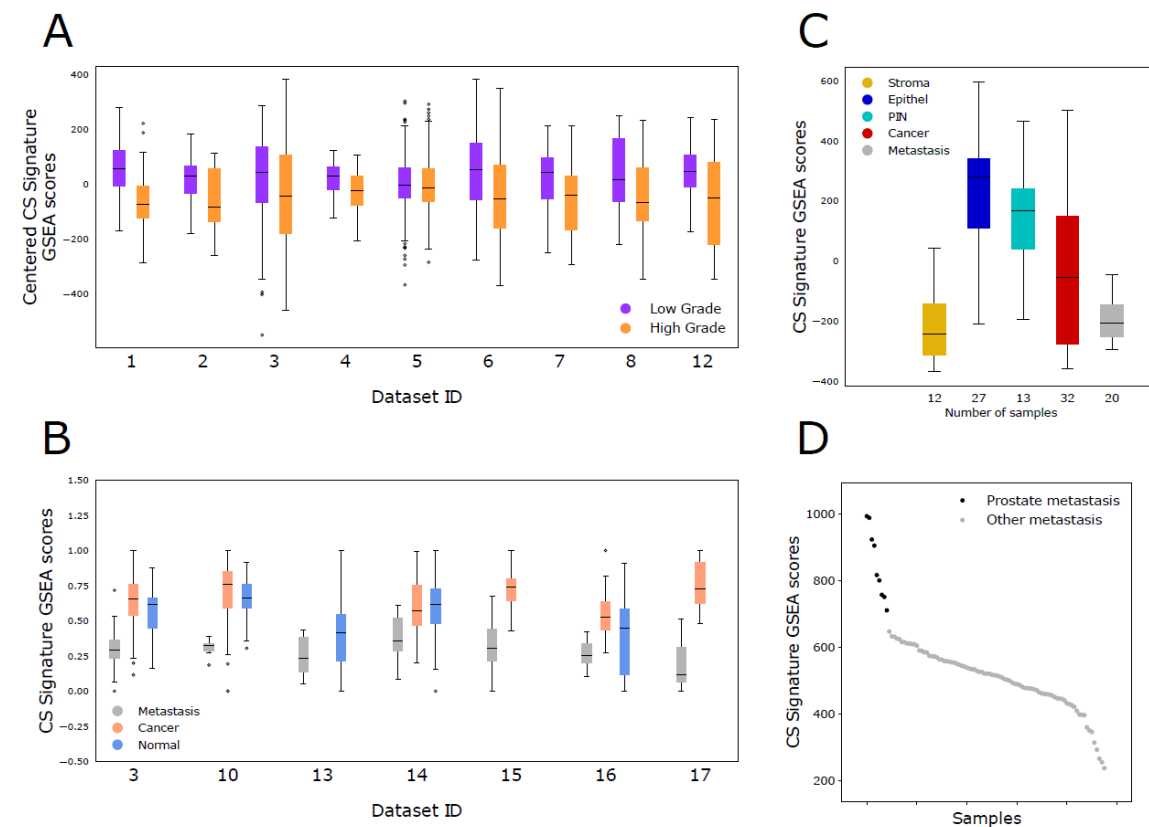| Dataset ID	Dataset abbreviation	Number of metastatic samples	Number of cancer samples	Number of normal samples	P-value metastatic vs cancer	P-value metastatic vs normal
3	Taylor	19	131	29	7.4e-15	2.8e-7
10	Prensner	12	78	38	3.4e-11	2.9e-10
13	Aryee	18	NA	21	NA	0.024
14	Chandran	25	65	81	8.8e-6	6.9e-6
15	Cai	29	22	NA	5.8e-13	NA
16	Poisson	13	12	16	1.5e-4	0.09
17	Monzon	21	10	NA	7.4e-11	NA
18	Tomlins	20	32	27	0.007	1.2e-10

11

12

13

Figure 4: Citrate – spermine (CS) signature, tumor grade and metastasis. A) CS signature ssGSEA scores for high-grade (Gleason higher or equal to 4+3) and low-grade (Gleason less than or equal to 3+4) cancers in 9 datasets. The samples compared corresponds to the HG1 and LG1 groups in Table 1. The scores were centered in each cohort before plotting to visualize similarities between datasets better. B) CS signature ssGSEA scores for metastatic, cancer and normal samples in 7 datasets. The scores were centered and normalized to range 0-1 before plotting to visualize similarities between datasets better. The low scores in normal samples is due to the higher content of stroma in normal samples (Supplementary Figure 3). C) CS signature ssGSEA scores from five laser dissected prostate and prostate cancer tissue types from Tomlins (datasets Id 18). The GSEA scores illustrates the loss in CS secretion from normal epithelium through PIN, Cancer and finally Metastasis. The CS signature scores from metastatic samples are comparable to the non-secreting stroma tissue. D) CS signature ssGSEA scores in 96 metastatic samples (9 from prostate and 87 from other organs) from the Hsu (dataset ID 19). Metastatic samples from prostate origin are more enriched for the CS signature than metastatic samples originating from other sites.

CS signature genes are enriched for genes with zinc-binding, mitochondrial and secretory functions.
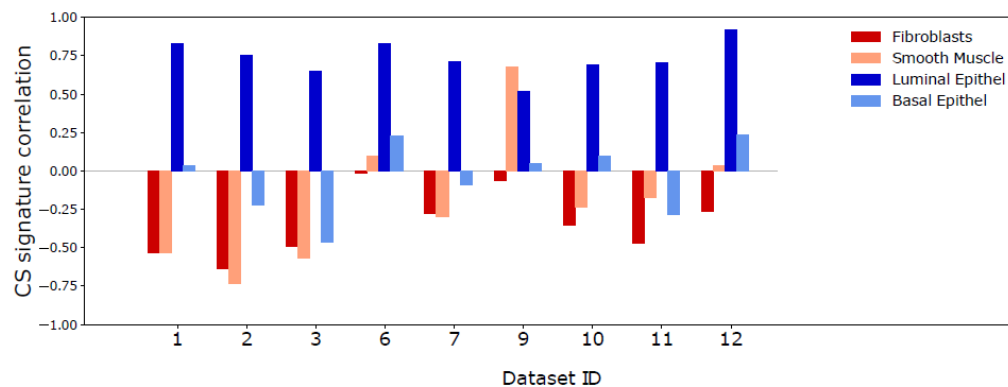
Gene Ontology analysis anchored the genes in the CS gene signature to the citrate secretory pathways from mitochondria towards the cell exterior via the ER and golgi. The analysis also highlighted several genes related to zinc-binding, in particular several metallothionines, and a role of branched chain fatty acid catabolism in upholding the normal prostate secretory function.

When examining gene functions on gene-by-gene basis using the NCBI-gene resource [ncbi.nlm.nih.gov/gene] (Supplementary File 3) we found that several genes in the CS signature were related to zinc-binding. Among these, six were Metallothioneins (*MT1G, MT1M, MT1F, MT1X, MT1E* and *MT2A*), one zinc-transporter (*SLC39A10*) and the two genes *AZGP1* (alpha-2-glycoprotein 1, zinc-binding) and *ANPEP* (alanyl aminopeptidase, membrane), the latter containing a consensus sequence known from zinc-binding metalloproteinases.

We then used DAVID [45, 46] and Enrichr [47, 48] to perform Gene Ontology (GO) analysis on the CS signature (Supplementary Figure 4). For both GO tools the most significant terms were related to *zinc/metal-ion binding* and *branched chain amino acid (Leucine, Isoleucine and Valine) degradation*. When inspecting the CS signature genes at NCBI, we noticed that many genes did not have any associated functional description, while at the same time having high prostate specific expression. A potential challenge for our GO analysis is that the prostate specific function of many of the genes in our CS signature is not known. We therefore applied GAPGOM, an alternative GO analysis tool where each gene is associated with functions based on consensus ontology terms for genes it is correlated with [49] (methods, Supplementary File 4). This analysis revealed strong gene ontology terms related to mitochondria, ER (endoplasmic reticulum), golgi, lysosome and exosome cellular components, which would fit well with a secretory path of citrate. When we performed Principal Component Analysis (PCA) to group genes according to their GO terms (see methods), the six metallothioneins formed a distinct cluster related strongly to zinc/metal-ion binding (Figure 5A, Supplementary File 4). In addition, the metallothioneins also showed strong associations with the mitochondrial respiratory complex and electron transport chain.

We also performed network analysis on the 150 genes in the CS signature (Supplementary Figure 5), where the genes *(ALOX15B, RAB27A, ENDOD1, SLC45A3, NCAPD3, EHHADH, ACAD8, AFF3, NANS* and *YIPF1*) were identified as the top 10 hubs in the network. These hub-genes were all ranked among the top 20 in the CS-signature and had significant GO terms associated with branched chain fatty acid catabolism/degradation. Finally, we compared CS signature ssGSEA scores with cell-types specific gene signatures from a single-cell RNA-Seq study on normal prostate tissue [50]. The CS signature scores showed high correlation with prostate luminal cells (the cells mainly responsible for secretion), but not basal cells, and a negative correlation to stromal fibroblasts and smooth muscle cells (Figure 5B). Moreover, the luminal gene signature also included several of the metallothionein genes, corroborating the potential importance of these genes.

1

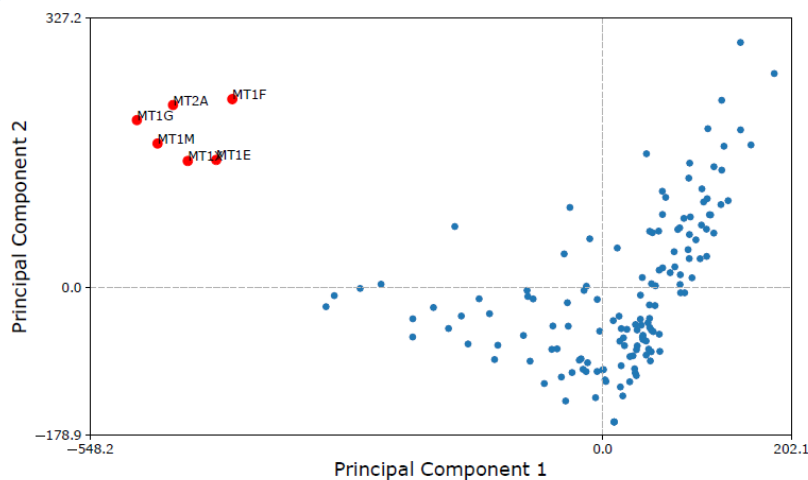

2 **Figure 5: Genes and ontologies.** *A) Principal Component Analysis (PCA) for Gene Ontology terms*
3 *associated with CS signature genes. The GO terms are based on a consensus analysis over datasets 1-*
4 *12 (see methods). The only observable cluster is formed by the six Metallothioneins from the CS*
5 *signature. B) Correlation of ssGSEA signature with four gene signatures from distinct cell types in a*
6 *single cell sequencing dataset (basal and luminal from the prostate epithelium, fibroblasts and*
7 *smooth muscle cells from stroma) [50] in 9 datasets with normal prostate samples. The CS gene*
8 *signature correlates with luminal epithelial cells (the cell-type mainly responsible for secreting citrate*
9 *and spermine), while the correlation to stroma tissue types is negative. Dataset 9 is from laser*
10 *dissected epithelium, which lack stroma tissue.*

11

12

13

14

15

16

17

1   **The CS signature is depleted in prostate model systems**
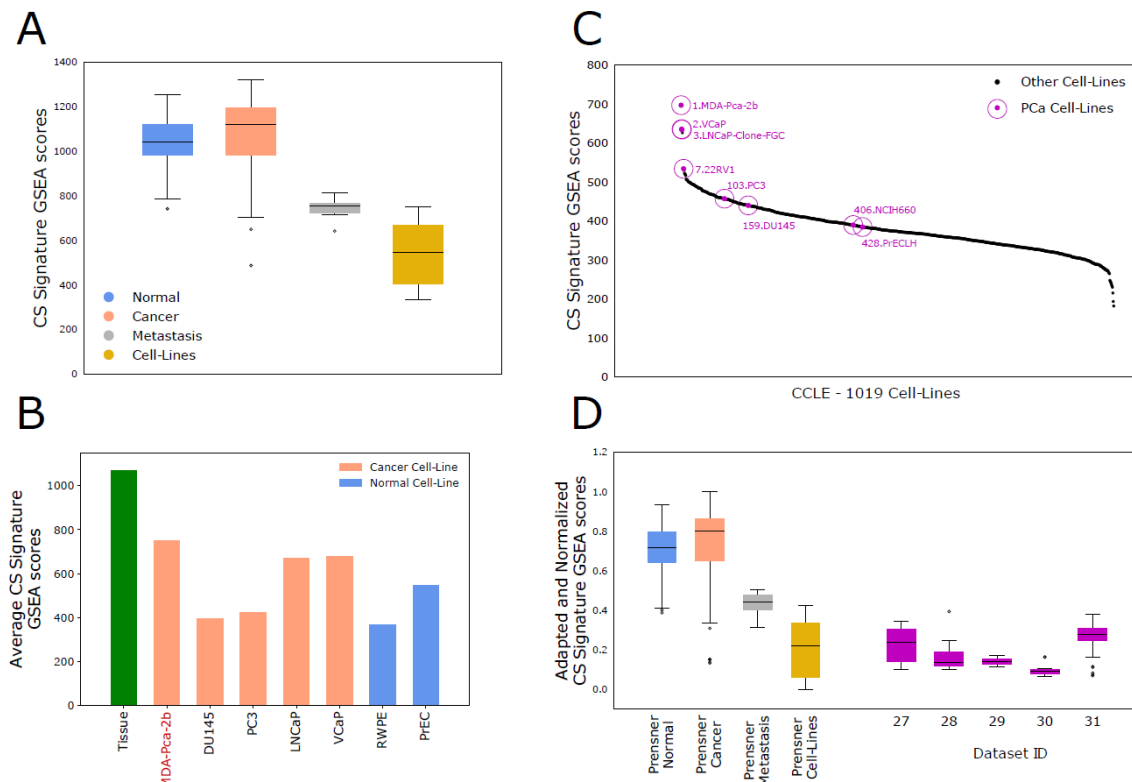
2   To reveal detailed function of the genes highlighted in the previous section will need further wetlab
3   experiments. We therefore wanted to identify suitable model-systems where the possible role of
4   these genes and their mechanisms could be further studied. Unfortunately, we found the
5   identification of model systems to be a challenge.

6   We tested the CS signature on publicly available gene expression datasets from various model systems
7   for prostate cancer using ssGSEA. In the *Prensner* (dataset ID 10), 58 expression profiles from cell-lines
8   were analyzed together with tissue and metastatic samples, including the four most common prostate
9   cancer cell-lines PC3, DU145, LNCaP, VCaP, and the normal prostate cell lines RWPE and PrEC. This
10 unique collection of sample types makes the *Prensner* dataset an excellent reference for comparisons.

11 CS signature ssGSEA scores from the *Prensner* cell-lines were consistently at the low end, compared
12 to both tissue and metastatic samples (Figure 6A), indicating that cell-lines derived from prostate do
13 not secrete citrate and spermine. This observation was confirmed in *Taylor* (dataset ID 3), which also
14 contained four cell-line samples (Supplementary Figure 6). When comparing different cell-lines,
15 androgen responsive cell-lines (LNCaP and VCaP) generally have higher CS signature ssGSEA scores
16 then androgen resistant (DU145 and PC3) cell-lines (Figure 6B). The overall highest CS scoring cell-line
17 is MDA-Pca-2b (an androgen responsive cancer cell-line). Interestingly, normal prostate cell-lines
18 (RWPE and PrEC) do not score higher than the cancer cell-lines. The relative differences in CS ssGSEA
19 scores between the different cell-lines were highly stable, which was confirmed when assessing
20 additional cell-line samples in *CCLE* (Cancer Cell Line Encyclopedia, dataset ID 24) (Figure 6C), *Søgaard*
21 (Dataset ID 26) and *E-MTAB-2706* (dataset ID 25). (Supplementary Figure 7 and 8, respectively).
22 Additionally, results from cohort *CCLE* and *E-MTAB-2706* show that androgen resistant prostate cancer
23 cell-lines (DU145 and PC3) do not separate from cell-lines of other cancer origin in terms of CS
24 signature GSEA scores (Figure 6C, Supplementary Figure 8).

25 To enable the comparison of ssGSEA scores between datasets, we implemented a method to adapt
26 ssGSEA scores from different datasets to the *Prensner* dataset (see Methods). To demonstrate the
27 utility of this implementation, we first adapted the CS ssGSEA scores from *Taylor* to *Prensner*, since
28 both datasets included tissue samples (normal and cancer), metastatic samples as well as cell-lines.
29 The CS ssGSEA scores where highly concordant down to the level of different cell-types between the
30 two cohorts after adaptation (Supplementary Figure 9). We then used the adaptation strategy to
31 compare datasets from organoids and mouse models (dataset IDs 27-31) to the *Prensner* dataset to
32 investigate whether they would resemble tissue or cell-lines. None of the model-systems analyzed
33 seemed to produce CS ssGSEA scores at levels similar to prostate epithelial tissue or low-grade primary
34 cancer tissue. Instead their CS scoring range typically compare to the cell-lines from *Prensner* (Figure
35 6D). Overall, these results call attention to the research challenge for current prostate model systems
36 to recapitulate the function of normal prostate epithelium in terms of citrate and spermine secretion.

37 ***Figure 6: Model systems.*** *A) CS signature ssGSEA scores for all sample types in Prensner (dataset ID*
38 *10), including 58 prostate normal and cancer cell-line samples. Cell-line samples, in general, have*
39 *very low CS signatures scores. B) Average CS signature ssGSEA scores for different cell-types in*
40 *Prensner. The ssGSEA scores were highly reproducable within samples from the same cell-type.*
41 *Androgen responsive cell-types (MDA-Pca-2b, LNCaP and VCaP) have somewhat higher ssGSEA*
42 *scores than androgen resistant cell-types (DU145 and PC3). Prostate normal cell-types are similar to*
43 *cancer cell-types. Number of samples of each cell-type are: MDA-Pca-2b (1), DU145(10), PC3(2),*
44 *LNCaP(7), VCaP(10), RWPE(9), PrEC(4). C) CS signature ssGSEA scores in 1019 cancer cell-types from*
45 *CCLE (dataset ID 24). Androgen responsive prostate cancer cell-types score higher than cancer cell-*

14

1   *types from other cancers, while androgen resistant cell-lines score similar to cell-types from other*

2   *cancers. D) : CS Signature ssGSEA scores in four datasets with prostate and prostate cancer derived*

3   *organoids and model systems (dataset IDs 27-30) and one dataset with prostate tumors from mouse*

4   *models (Aytes, dataset ID 31). The ssGSEA scores were adapted and normalized to CS signature GSEA*

5   *scores from Prensner. All model systems have consistently low GSEA scores, similar to prostate*

6   *cancer cell-lines. In total 64 model system samples and 384 mouse tumor samples were analyzed.*

7   *Dataset descriptions: **27**: 2D and 3D cultures derived from prostate cancer cell-line LNCaP (9 samples)*

8   *and normal cell-line RWPE (9 samples), in total 18 samples. **28**: Organoids from primary prostate*

9   *epithelial cells in mono and co-culture with stromal cells (7 samples), stromal cells in co-culture (4*

10   *samples) and macrodissected tumor tissue (2 samples), in total 13 samples.**29**: Cells from benign*

11   *prostatic bulk (3+3 samples), basal (3+3 samples) and luminal (3+3 samples) tissue cultured in two*

12   *different media, in total 18 samples. **30**: Mouse prostate organoids with WT (6 samples) and mutated*

13   *(6 samples) gene SPOP, in total 12 samples. **31**: Prostate tumor tissue from mouse models with*

14   *different genetic modifications, in total 384 samples. (See also Supplementary Figures 6-9).*

15

16   **Discussion**

17   **Expanding the view of zinc-binding proteins and zinc-transporters in the prostate**

18   Six metallothioneins were identified as part of the CS gene signature in this study and constituted a

19   functionally distinct cluster particularly associated with zinc-binding. Metallothioneins are low

20   molecular weight, metal binding proteins localized to the Golgi, and have different expression in

21   normal prostate and prostate cancer [51, 52] . In addition, expression levels can change in response

22   to zinc-stimuli in prostate cell-lines [53]. It has also been speculated that metallothioneins can affect

23   mitochondrial function, since they are small enough to enter the mitochondrial membrane bilayer

24   carrying zinc [54]. This suggestion fits with the GO associations of metallothioneins to mitochondria,

25   respiratory complex and the electron transport chain. Otherwise, the other potential zinc-binding

1   genes discovered in this study have not been studied previously in the context of citrate accumulation
2   and secretion in the prostate.

3   The zinc-transporter SLC39A1 (ZIP1) was not a part of the CS-signature, though this gene has
4   previously been shown to be important for zinc-homeostasis in prostate cells. There could be several
5   reasons for this discrepancy. First, this study only measures genes at the transcript level, while it is
6   possible that SLC39A1 is regulated at the protein level [6]. However, expression levels from genes
7   displayed in NCBI-gene resource show that SLC39A1 is not prostate specific, but expressed in most
8   tissues. Though SLC39A1 has the ability to transport zinc within prostate tissue, it may not be the main
9   determinant of prostate specific zinc regulation *in vivo*. In this context is should also be noted that
10   none of the metallothioneins displayed prostate specific expression in the NCBI-gene database. Of the
11   zinc transporters, most prostate specific expression is observed for SLC39A2, SLC39A6, SLC39A7 and
12   SLC39A10, where SLC39A10 was part of the CS signature. SLC39A6 and SLC39A7 (though not part of
13   the signature) were also positively correlated with citrate in our analysis, while SCL39A1 and SLC39A2
14   showed no correlation. Of the potential zinc-binders, AZGP1 show the most prostate-biased
15   expression, but the function of this gene is unknown. Overall, more targeted research experiments
16   are needed to identify the genes that control zinc-levels in the prostate, and how they perform their
17   function.

18   Previous studies have shown that prostate epithelial cells can utilize glucose and aspartate as the
19   carbon sources for citrate production. The gene *GOT2* (*glutamic-oxaloacetic transaminase 2*, also
20   named *mAAT-mitochondrial aspartate aminotransferase* in previous reports) was shown to be
21   responsible for the conversion of aspartate to oxaloacetate and citrate in the mitochondria. *GOT2* was
22   a part of the CS signature, in addition to *SMS* (*spermine synthase*), and both *GOT2* and *SMS* showed
23   high tissue-expression in prostate according to NCBI-Gene. For these two genes our results fit with
24   previous mechanistic knowledge [9, 55].

25   Very little is known about the role of branched chain amino acid catabolism for normal prostate
26   function. All of the three amino acids leucine, isoleucine and valine are upregulated in prostate cancer
27   [12], indicating their relevance for cancer transformation, and this fits with our finding that the
28   degradation of these amino acids is important to maintain normal prostate function. Branched chain
29   amino acids can possibly work as precursors for citrate production [56], and leucine can act as a sensor
30   for mTOR-pathway activation [57], which is generally regarded as an important pathway in prostate
31   cancer development [58].

32   **Expression of CS signature genes are with prostate cancer progression**

33   There has been a discussion on whether the reduced citrate levels observed in prostate cancer are
34   merely a result of a reduction in luminal glands leading to reduced amounts of prostatic fluid [59].
35   However, our results clearly show that changes in citrate and spermine levels are accompanied by
36   changes at the gene expression level, and this study supports the hypothesis that changes in genes
37   responsible for citrate and spermine secretion are important for the cell transformations leading to
38   cancer. Further research into these mechanisms can lead to new discoveries and treatment targets in
39   the management of prostate cancer.

40   **The challenge of finding a model system for normal prostate**

41   We could not find CS signature enrichment comparable to tissue samples in any of the cell-lines and
42   model systems we investigated. Androgen responsive cell-lines seem to maintain some traces of
43   prostate specific functions compared to other model systems, but their CS signature enrichments are

16

1   far below that of *in vivo* tissue and indicate that most parts of the prostate-lineage specificity is lost.
2   Thus, these observations question the relevance of these model systems for studying normal prostate
3   function and transitions from normal cells to prostate cancer cells. The secretory function can possibly
4   be triggered by proper stimuli like dihydrotestosterone (DHT), testosterone or prolactin [60-62]. For
5   example, LNCaP (but not PC3) cells were able to secrete citrate when stimulated by DHT [61].
6   However, it was also observed that the rate of citrate consumption by the TCA cycle increased
7   proportionally in the same experiment [61]. Since these experiments were performed before the era
8   of transcriptomics, it would be interesting to repeat these experiments with accompanying
9   transcriptomics analysis to identify the genes that change during such stimuli.

10   **Signature refinement produces more robust gene signatures with improved integrity**

11   In this manuscript, we have argued that improved gene signature integrity (assessed by CMS) can be
12   achieved by integrating data from multiple prostate cancer datasets. The aim was to remove noise
13   and produce better and more robust gene signatures in terms of biological interpretation. We found
14   several evidences that more robust signatures were achieved by integration of datasets. First, we
15   observed that the refined signature improved CMS scores compared to the initial signature in
16   independent normal prostate samples (Figure 2D). Second, when we used the initial gene signature
17   for ssGSEA in the *Hsu* dataset, we observed no clear separation between prostate and other
18   metastatic samples (Figure 4D, Supplementary Figure 10). Third, in the *FANTOM* dataset, the one
19   Prostate Adult Tissue sample were only ranked as the highest scoring sample after signature
20   refinement, but not when the initial signature was used (Figure3D, Supplementary Figure 11).

21   The refinement procedure generated an improved CS signature ssGSEA scores which was particularly
22   pronounced for prostate samples. We also tested the effect of refinement ion gene expression data
23   from other tissue types in the *TCGA-complete* dataset (33 cancer and 22 normal tissue types). Other
24   cancer types showed both elevated and decreased GSEA scores after refinement, and elevations were
25   consistently lower than in prostate samples (Supplementary Figure 12).

26   The best CS signature ssGSEA score improvement after refinement were also observed particularly for
27   the metabolites citrate and spermine. For this test, we redid the procedure to create initial and refined
28   signatures for all metabolites and lipids in the metabolite data from *Bertilsson* (23 metabolites and 17
29   lipid signals). We then calculated ssGSEA scores in all cancer and normal tissue types in the *TCGA-
30   complete* dataset for each metabolite/lipid signature (Supplementary Figures 13-16). The most
31   elevated ssGSEA scores were for citrate and spermine in prostate compared to other tissue types,
32   while other metabolites showed both elevated and decreased ssGSEA score for other tissue types
33   compared to prostate. This observation was similar for both cancer and normal tissue samples. Thus,
34   the effect of refinement was highest for citrate and spermine specifically in the prostate. Interestingly,
35   we noted that other metabolites/lipid signatures were elevated in prostate, particularly putrescine (a
36   precursor for spermine in the polyamine pathway) a few lipid signals and glucose, indicating prostate
37   specific regulation of other metabolites than just citrate and spermine. These observations were
38   confirmed in the *FANTOM* dataset (dataset ID 23) (Supplementary Figure 17).

39   In summary, we conclude that the identification followed by refinement procedure used in this study
40   created a robust and unbiased gene signature biologically relevant for in vivo spermine and citrate
41   secretion in the prostate. In a wider perspective, the module approach may represent a general
42   strategy to find robust genesets related to other –omics data (for example metabolomics, proteomics
43   or lipidomics), as long as these -omics data are accompanied with gene expression data.

1 **Conclusion**

2 The genes and mechanisms governing normal prostate function, in particular the ability to accumulate
3 and secrete large amounts of citrate and spermine, are mostly unknown. We have used bioinformatics
4 on an extensive collection on prostate datasets to discover genes relevant for secretion of citrate and
5 spermine in the prostate. Based on this we have created a 150 gene signature which can be used to
6 assess the degree of citrate and spermine secretion in a single sample. The bioinformatics approach
7 enabled us to validate and explore the gene signature in a large number of contexts based on public
8 data, including normal vs cancer enrichment, spatial colocalization on tissue, tumor grade, metastasis
9 and different model systems. The employed approach is generalizable to other types of data when
10 measured together with gene expression. The signature showed an inverse association to cancer
11 progression from normal through low and high grade to metastasis. The genes in the signature were
12 enriched for zinc-binding, mitochondrial function, respiratory complex, secretion through the ER –
13 golgi - lysosome – exosome pathway and branched chain amino acid catabolism, pointing to known,
14 suggested and novel mechanisms for normal prostate function. The lack of suitable model systems is
15 a challenge, and need to be established to study these findings further.

16 **Methods**

17 **Datasets**

18 A list of all datasets with description and references are given in Supplementary File 1

19 **Module refinement procedure**

20 A schematic representation of the procedure with stages 1-3 is shown in Figure 7A. The procedure
21 make use of the following abbreviations: **M**=Module, **G**=several genes, **g**=single gene, **D**=Dataset,
22 **P**=gene expression Profile, **TB**=Table, **av**=average, **in**=initial, **rf**=refined, **c**=candidate, **m**=missing,
23 **cr**=correlation, **ic**=intra-correlation, **CMS** = Correlation Module Score (CMS), **MGCR** = Module Gene
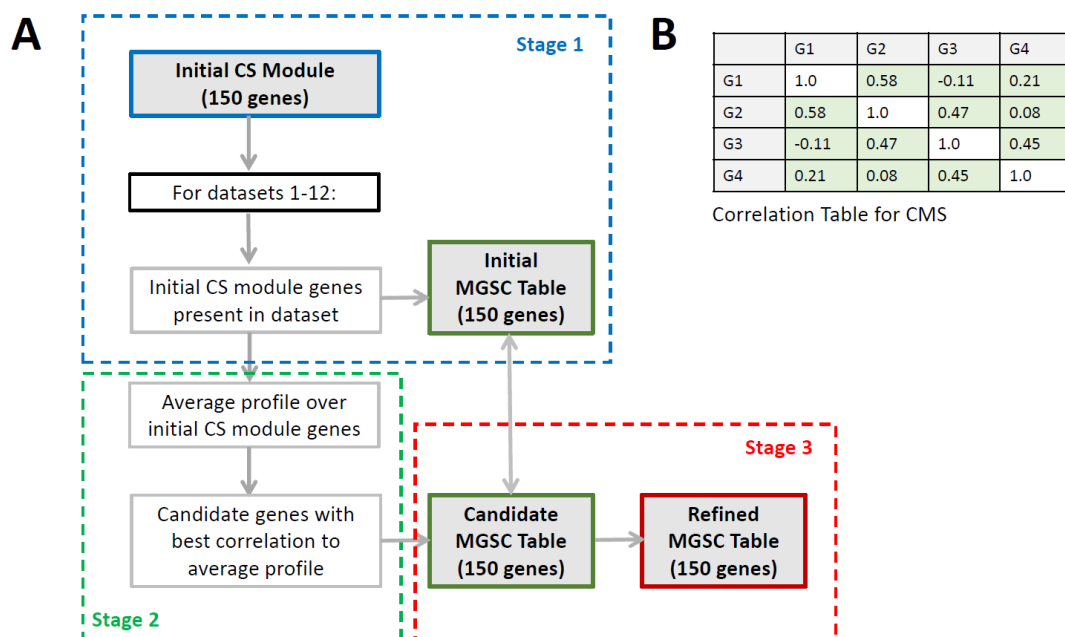24 Contribution Score (MGCR).

25

1 ***Figure 7: Module refinement procedure and Correlation Module Score (CMS)*** *A) Flowchart of the*
2 *module refinement procedure. Each stage 1-3 are explained in the main text. B) Example of a*
3 *correlation table used to calculate the Correlation Module Score (CMS). This gene module has four*
4 *genes (G1-G4) and a CMS of 0.32.*

5

6 **Correlation Module Score (CMS)**

7 To assess how well the correlations between genes in a module (or geneset) is preserved for all (or a
8 subset of) samples in a particular dataset, we created the Correlation Module Score (abbreviated
9 CMS). The CMS for a module in a dataset is calculated by first creating a table of intra-correlations
10 between expression profiles for all genes in the module for that dataset (Figure 7B). Here the
11 expression profile for a gene is the vector consisting of gene expression values over all samples in that
12 dataset. For a 150-gene module, this will create a table with 150*150 = 22500 correlation values (a
13 table for a hypothetical 4-gene module is shown in Figure 7B). The CMS is then calculated as the
14 average over all correlations included the table (but excluding self-correlations). A high CMS
15 represents a high integrity of the gene module in the dataset, indicating the preservation of strong
16 correlations between the module-genes in the datasets. Likewise, a low CMS would indicate weak
17 relations between the genes in the module, and the loss of gene module integrity. We further assume
18 if a gene module has a high CMS in a dataset, it would indicate that the biological process the module
19 represents is important or active for samples in this dataset. In this study the datasets are patient
20 cohorts with normal and cancer tissue samples from the prostate.

21 **Refinement Procedure - Stage 1**

22 The purpose of Stage 1 in the module refinement procedure is to find the contribution of each gene
23 in the initial CS module from *Bertilsson* to the overall CMS for the module. To represent this
24 contribution we introduce the Module Gene Contribution Score (MGCS), which is the average over all
25 correlations between that gene and each of the other genes in the module (or, the overall contribution
26 of that gene to the overall CMS for the module). In Figure 7B this will be the average of the rows
27 (excluding the diagonal), where each average value represent the MGCS score for a gene. A high MGCS
28 score will indicate that the expression profile of the gene fits better with expression profile of other
29 genes in the module, while a low MGCS score indicates a gene expression profile that deviates from
30 other genes in the module.

31 Procedure:

32 S1: For each dataset **D** (dataset ID 1-12 with normal and cancer tissue samples from the prostate):

33 Find the genes from the initial CS module, **in-M-G,** present in the dataset **D, in-M-G-D**

34 Calculate the intra-correlation table, **ic-TB-D,** for the initial CS module gene profiles, **in-M-G-P-D**, in
35 dataset **D**

36 Calculate the corresponding CMS score, **CMS-D**, from the intra-correlation table, **ic-TB-D**, in dataset
37 **D**

38 Calculate the MGCS score for each initial CS gene **in-g** in **CMS-D, in-g-MGSC-D**

39 For missing initial CS module genes, **m-in-M-G-D**, in dataset **D**: Set the **in-g-MGSC-D** to a *dataset*
40 *specific missing gene constant* equal to the **CMS-D** multiplied by 0.5

19

1 Store the **in-g-MGCS-D** for each initial CS module gene in each dataset **D** in an *Initial MGSC table,* **in-**
2 **MGCS-TB** for comparison in Stage 3.

3 Store MGSC for missing genes in each cohort, **m-MGSC**, for **Stage 2.**

4 **Refinement Procedure - Stage 2**

5 The purpose of Stage 2 is to calculate, for all candidate genes **c-g** present in any one of the 12 datasets
6 **D**, a score which describes how well each candidate gene expression profile fit with the expression
7 profiles of genes in the initial CS module. This is done by comparing each candidate gene expression
8 profile **c-g-P-D** to the average expression profile, **in-M-G-P-D-av**, of the genes in the initial CS module
9 for each dataset **D**. We regard any gene present in any of the 12 datasets **D** as candidate genes.

10 Procedure:

11 Define candidate genes, **c-G**, by the union of all genes present in any dataset **D**

12 For each dataset **D**:

13 Find the genes from the initial CS module, **in-M-G**, present in dataset **D**

14 Calculate the average expression profile, **in-M-G-P-D-av**, over expression profiles for all initial
15 module genes **in-M-G-P-D** in dataset **D**

16 For each gene **g** present in dataset **D**:

17 Calculate the correlation between the expression profile of gene **g** and the average profile **in-M-G-P-**
18 **D-av** for dataset **D**

19 For genes not present in this dataset, the correlation is set to the *cohort specific missing gene*
20 *constant* in **m-MGSC** from **Stage 1**

21 Store all correlations in a table, **cr-TB** ,for evaluation in **Stage 3**

22 **Refinement Procedure - Stage 3**

23 The purpose of Stage 3 is to identify the genes that correlates best with genes in the initial CS module
24 over all 12 datasets, but that were not a part of the initial CS module to begin with. From this analysis
25 we select the 150 best candidates, and compare their MGCS to the MGSC for the genes in the initial
26 CS module from **Stage 1.** If the MGSC for a new candidate gene is higher than the MGSC for an initial
27 CS module gene, the initial gene is replaced by the candidate gene. The final output is a refined CS
28 module, where the robustness of gene correlations across all 12 prostate cancer datasets are taken
29 into consideration in the selection of genes for the module. This refined module is defined as the CS
30 gene signature.

31 Procedure:

32 Exclude initial CS module genes **in-M-G** from the correlation table **cr-TB** from **Stage 2**

33 Calculate the average correlation, **c-g-cr-av**, for each of the remaining candidate genes over all 12
34 datasets in **cr-TB**.

35 Sort genes in correlation table **cr-TB** by average correlation **c-gr-cr-av**.

36 Select 150 candidate genes, **c-g-150**, with the highest average correlation, **c-gr-cr-av**

37 For each dataset **D**:

1   For each of the top 150 candidate genes, **c-g-150**:

2   Calculate the average correlation of gene **c-g-150** to all genes from the initial CS module, **in-M-G-D**,
3   and use the average correlation as the candidate gene MGSC, **c-g-MGSC-D**. This is a measure of the
4   fit for the candidate gene, **c-g-150**, to the initial CS module in dataset **D**.

5   For candidate genes not present in dataset **D,** the MGSC is set to the *cohort specific missing gene*
6   *constant* in **m-MGSC** from **Stage 1**

7   Combine the candidate MGSC table **c-g-MGSC-D** with the initial MGSC table **in-g-MGSC-D** from **Stage**
8   **1.** The combined table , **comb-TB**, will contain 300 genes, 150 from the initial CS module and 150
9   new candidates.

10   Calculate the average MGSC score, **g-MGSC-av,** for each gene over all 12 datasets for the combined
11   table **comb-TB**.

12   Sort genes by average MGSC score, **g-MGSC-av**.

13   Select the 150 genes with the highest average MGSC score, **g-MGSC-av**, as genes for the refined CS
14   module, **rf-G-M**.

15   The number of genes from the initial CS module present in each of the 12 datasets are shown in
16   Supplementary Table 1.

17

18   **Single sample Gene Set Enrichment Analysis (ssGSEA)**

19   Single sample Gene Set Enrichment Analysis (ssGSEA) was performed by sorting the genes in each
20   sample according to their expression level in descending order. Then a ssGSEA score was calculated
21   for each sample implementing a previously published algorithm [43]. We did not perform centering
22   and normalization of the expression values for ssGSEA in this study. This is because citrate and
23   spermine secretion, as well as the content of stroma tissue in the sample, are dominating biological
24   features of these samples where differences should be visible at absolute expression level. However,
25   we observed good correlations (0.86 – 0.97) for ssGSEA scores based on normalized and absolute
26   expression values. A few of the datasets were only presented with centered expression values in the
27   public resource, and for these we used centered values for ssGSEA. For the spatial transcriptomics
28   data from *Berglund* (dataset ID 20), we did use centered and normalized expression for ssGSEA in each
29   pixel. This was necessary due to many pixels with a high number of unexpressed genes, and/or very
30   few genes with abundant expression.

31   **Adapted and normalized ssGSEA**

32   The magnitude of GSEA scores depend on both the size of the gene set used, and the number of total
33   genes in the ranked list input to the GSEA calculations. To be able to compare ssGSEA scores from
34   *Prensner* to other datasets, we adapted the ssGSEA by only using the genes shared between the
35   datasets in each comparison. In this way, both the number and identity of genes, as well as the size of
36   the geneset will be identical for both datasets. For the comparison of several datasets in Figure 4D,
37   we also normalized the GSEA scores in *Prensner* to 0-1 scale for each comparison. The adapted and
38   normalized score calculated in each comparison were highly correlated (average 0.99), and had a small
39   standard deviation of the mean (0.004), and we conclude that they are comparable between the
40   datasets.

41   **GO analysis**

1   The GO analysis was based on gene annotation generated with the GAPGOM tool [49] [PMID:
2   30567492, GAPGOM in Bioconductor]. GAPGOM will predict gene annotation for a target gene by
3   identification of well-annotated genes showing correlated expression pattern with the target gene
4   across several experiments, and then estimate a predicted annotation for the target gene as a
5   consensus over the co-expressed genes, based on the hypothesis that co-expressed genes may be
6   involved in similar or related processes. This prediction of annotation terms may facilitate a richer
7   functional annotation of genes, in particular for genes where there is a lack of experimental
8   annotation data.

9   Since GAPGOM only allows us to do annotation prediction one gene at a time, GAPGOM was used in
10 conjunction   with   Snakemake
11 [https://snakemake.readthedocs.io/en/stable/project_info/citations.html] to create a software
12 pipeline. This pipeline was used to predict GO term annotation for all 150 genes in the CS-signature
13 on each of dataset 1-12. Since GO annotation consists of three different ontologies (MF, BP, CC), each
14 prediction is done separately for each. This then outputs a list for each gene in each dataset for each
15 GO ontology, in total 1800 lists. Each list contains one or multiple GO-term predictions, with the
16 following information for each prediction; GO-ID, Ontology, P-value, FDR/q-value (Bonferoni
17 normalized P-value), The description of the GO term, and the used correlation method for the
18 prediction. For each list, only GO terms with a q-value > 0.05 were selected. This generated a total of
19 21 080 GO terms for all lists, where 5354 GO terms were unique. A table was created with the number
20 of genes as rows and the number of unique GO-terms as columns. For each gene and GO-term, we
21 summed $-\log_{10}$ q-values from all datasets, producing an overall score for each GO-term and gene. We
22 made three tables, one for all 150 genes, one for the six Metallothioneins, and one for the 10 network
23 Hub-genes.

24 **Network analysis**

25 Correlation-based networks of the CS signature genes for each dataset are created as follows: The 150
26 genes are represented as nodes in the network and the pairwise Pearson correlation between genes
27 represents the interactions between nodes. To reduce the complexity and highlight the most powerful
28 interactions, only the 20 strongest outgoing links (in absolute Pearson correlation) from each node are
29 kept. In this way, the most central nodes will have 20 outgoing links and up to 130 ingoing links from
30 other nodes, while some of the non-central nodes will only have 20 links. By calculating the node-
31 degree (number of links to other nodes) for each node, genes with high node-degree reflects central
32 genes for driving the biological processes [63]. This network construction is performed on cancer
33 samples for all 12 datasets, and the 10 genes with the highest mean degree across the datasets are
34 considered as the top 10 hub genes.

35 **Supplementary Material**

36 **Supplementary File 1: Overview of datasets with references used in this study.**

37 **Supplementary File 2: Supplementary tables and figures**

38 **Supplementary File 3: CS signature genes and manually curated functional annotation**

39 **Supplementary File 4: Top 20 GO terms for all CS signature genes, six Metallothioneins and 10**
40 **network Hub-genes in the categories Biological Process (BP), Molecular Function (MF) and Cellular**
41 **Component (CC).**

Acknowledgements

Author Contributions

MBR, MBT and HB conceived the idea and developed the concept. MBR, MBT and HB performed data curation. MBR developed methods and performed overall analysis. FD and CM developed and performed the GAPGOM analysis. MH performed network analysis. MBR, MBT and TFB acquired funding. MBR, MBT, SB, FD and TFB performed supervision and provided intellectual input. MBR, MBT, SB, HB, FD and TFB prepared the original draft. MBR prepared figures. All authors wrote and reviewed the manuscript.

Competing Financial Interests

The authors declare no competing financial interests

Ethical statement

The use of human tissue material and clinical data from the *Bertilsson* cohort was approved by the Regional Committee for Medical and Health Research Ethics (REC) for Central Norway, approval no 4-2007-1890. All experiments were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants. RNA-Seq data from the *Prensner* cohort was approved through dbGap (project #5870), and data were downloaded and stored according to the provided security requirements. All other data were downloaded from freely available and publicly open resources

References

1. Costello LC, Franklin RB: **The intermediary metabolism of the prostate: a key to understanding the pathogenesis and progression of prostate malignancy**. *Oncology* 2000, **59**(4):269-282.

2. Costello LC, Liu Y, Zou J, Franklin RB: **Mitochondrial aconitase gene expression is regulated by testosterone and prolactin in prostate epithelial cells**. *Prostate* 2000, **42**(3):196-202.

3. Costello LC, Franklin RB, Narayan P: **Citrate in the diagnosis of prostate cancer**. *Prostate* 1999, **38**(3):237-245.

4. Giskeodegard GF, Bertilsson H, Selnaes KM, Wright AJ, Bathen TF, Viset T, Halgunset J, Angelsen A, Gribbestad IS, Tessem MB: **Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness**. *PLoS One* 2013, **8**(4):e62375.

5. Braadland PR, Giskeodegard G, Sandsmark E, Bertilsson H, Euceda LR, Hansen AF, Guldvik IJ, Selnaes KM, Grytli HH, Katz B *et al*: **Ex vivo metabolic fingerprinting identifies biomarkers predictive of prostate cancer recurrence following radical prostatectomy**. *Br J Cancer* 2018, **118**(6):e11.

6. Costello LC, Franklin RB: **A comprehensive review of the role of zinc in normal prostate function and metabolism; and its implications in prostate cancer**. *Arch Biochem Biophys* 2016, **611**:100-112.

7. Kolenko V, Teper E, Kutikov A, Uzzo R: **Zinc and zinc transporters in prostate carcinogenesis**. *Nature reviews Urology* 2013, **10**(4):219-226.

8. Franklin RB, Costello LC: **Glutamate dehydrogenase and a proposed glutamate-aspartate pathway for citrate synthesis in rat ventral prostate**. *J Urol* 1984, **132**(6):1239-1243.

9. Costello LC, Akuffo V, Franklin RB: **Testosterone stimulates net citrate production from aspartate by prostate epithelial cells**. *Horm Metab Res* 1988, **20**(4):252-253.

10. Bettuzzi S, Davalli P, Astancolle S, Carani C, Madeo B, Tampieri A, Corti A: **Tumor progression is accompanied by significant changes in the levels of expression of polyamine metabolism regulatory genes and clusterin (sulfated glycoprotein 2) in human prostate cancer specimens (vol 60, pg 28, 2000)**. *Cancer Research* 2000, **60**(5):1472-1472.

11. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer**. *Cancer Research* 2002, **62**(15):4427-4433.

12. Tessem MB, Bertilsson H, Angelsen A, Bathen TF, Drablos F, Rye MB: **A Balanced Tissue Composition Reveals New Metabolic and Gene Expression Markers in Prostate Cancer**. *PLoS One* 2016, **11**(4):e0153727.

13. Lynch MJ, Nicholson JK: **Proton MRS of human prostatic fluid: correlations between citrate, spermine, and myo-inositol levels and changes with disease**. *Prostate* 1997, **30**(4):248-255.

14. Cheng LL, Wu C, Smith MR, Gonzalez RG: **Non-destructive quantitation of spermine in human prostate tissue samples using HRMAS 1H NMR spectroscopy at 9.4 T**. *FEBS Lett* 2001, **494**(1-2):112-116.

15. Bertilsson H, Tessem MB, Flatberg A, Viset T, Gribbestad I, Angelsen A, Halgunset J: **Changes in Gene Transcription Underlying the Aberrant Citrate and Choline Metabolism in Human Prostate Cancer Samples**. *Clin Cancer Res* 2012, **18**(12):3261-3269.

16. Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, Mercola D, McClelland M: **In silico estimates of tissue components in surgical samples based on expression profiling data**. *Cancer Res* 2010, **70**(16):6448-6455.

17. Jia Z, Wang Y, Sawyers A, Yao H, Rahmatpanah F, Xia XQ, Xu Q, Pio R, Turan T, Koziol JA *et al*: **Diagnosis of prostate cancer using differentially expressed genes in stroma**. *Cancer Res* 2011, **71**(7):2476-2487.

18. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B *et al*: **Integrative genomic profiling of human prostate cancer**. *Cancer cell* 2010, **18**(1):11-22.

19. Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y, Perner S, Adami HO, Fall K, Mucci LA *et al*: **Molecular sampling of prostate cancer: a dilemma for predicting disease progression**. *BMC medical genomics* 2010, **3**:8.

20. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, Bergstralh EJ, Kollmeyer T, Fink S, Haddad Z *et al*: **Discovery and Validation of a Prostate Cancer Genomic Classifier that Predicts Early Metastasis Following Radical Prostatectomy**. *PLoS One* 2013, **8**(6):e66855.

21. Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, Egevad LA, Russell R, Ramos-Montoya A, Vowler SL *et al*: **Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study**. *Ebiomedicine* 2015, **2**(9):1133-1144.

22. Mortensen MM, Hoyer S, Lynnerup AS, Orntoft TF, Sorensen KD, Borre M, Dyrskjot L: **Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy**. *Scientific reports* 2015, **5**:16018.

23. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD *et al*: **Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression**. *Nat Biotechnol* 2011, **29**(8):742-749.

24. Penney KL, Sinnott JA, Tyekucheva S, Gerke T, Shui IM, Kraft P, Sesso HD, Freedman ML, Loda M, Mucci LA *et al*: **Association of Prostate Cancer Risk Variants with Gene Expression in Normal and Tumor Tissue**. *Cancer Epidem Biomar* 2015, **24**(1):255-260.

25. Kuner R, Falth M, Pressinotti NC, Brase JC, Puig SB, Metzger J, Gade S, Schafer G, Bartsch G, Steiner E *et al*: **The maternal embryonic leucine zipper kinase (MELK) is upregulated in high-grade prostate cancer**. *J Mol Med* 2013, **91**(2):237-248.

26. Aryee MJ, Liu WN, Engelmann JC, Nuhn P, Gurel M, Haffner MC, Esopi D, Irizarry RA, Getzenberg RH, Nelson WG *et al*: **DNA Methylation Alterations Exhibit Intraindividual Stability and Interindividual Heterogeneity in Prostate Cancer Metastases**. *Sci Transl Med* 2013, **5**(169).

27. Chandran UR, Ma CQ, Dhir R, Bisceglia M, Lyons-Weiler M, Liang WJ, Michalopoulos G, Becich M, Monzon FA: **Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process**. *Bmc Cancer* 2007, **7**.

28. Yu YP, Landsittel D, Jing L, Nelson J, Ren BG, Liu LJ, McDonald C, Thomas R, Dhir R, Finkelstein S *et al*: **Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy**. *J Clin Oncol* 2004, **22**(14):2790-2799.

29. Cai CM, Wang HY, He HSHS, Chen S, He LF, Ma F, Mucci L, Wang QB, Fiore C, Sowalsky AG *et al*: **ERG induces androgen receptor-mediated regulation of SOX9 in prostate cancer**. *Journal of Clinical Investigation* 2013, **123**(3):1109-1122.

30. Poisson LM, Sreekumar A, Chinnaiyan AM, Ghosh D: **Pathway-directed weighted testing procedures for the integrative analysis of gene expression and metabolomic data**. *Genomics* 2012, **99**(5):265-274.

31. Tomlins SA, Mehra R, Rhodes DR, Cao XH, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ *et al*: **Integrative molecular concept modeling of prostate cancer progression**. *Nature Genetics* 2007, **39**(1):41-51.

32. Hsu SD, Kim MK, Foye A, Silvestri A, Lyerly HK, Morse M, Petricoin E, Febbo PG: **Use of gene expression signatures to identify origin of primary and therapeutic strategies for patients with advanced solid tumors**. *J Clin Oncol* 2010, **28**(15).

33. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstrahle J, Tarish F, Tanoglidi A, Vickovic S, Larsson L *et al*: **Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity**. *Nature communications* 2018, **9**.

34. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S *et al*: **Gateways to the FANTOM5 promoter level mammalian expression atlas**. *Genome Biology* 2015, **16**.

35. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, Barretina J, Gelfand ET, Bielski CM, Li H *et al*: **Next-generation characterization of the Cancer Cell Line Encyclopedia**. *Nature* 2019, **569**(7757):503-+.

36. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang ZS, Liu HB, Degenhardt J, Mayba O, Gnad F, Liu JF *et al*: **A comprehensive transcriptional portrait of human cancer cell lines**. *Nature Biotechnology* 2015, **33**(3):306-+.

37. Sogaard CK, Moestue SA, Rye MB, Kim J, Nepal A, Liabakk NB, Bachke S, Bathen TF, Otterlei M, Hill DK: **APIM-peptide targeting PCNA improves the efficacy of docetaxel treatment in the TRAMP mouse model of prostate cancer**. *Oncotarget* 2018, **9**(14):11752-11766.

38. Richards Z, McCray T, Marsili J, Zenner ML, Manlucu JT, Garcia J, Kajdacsy-Balla A, Murray M, Voisine C, Murphy AB *et al*: **Prostate Stroma Increases the Viability and Maintains the Branching Phenotype of Human Prostate Organoids**. *Iscience* 2019, **12**:304-+.

39. Zhang DX, Lin K, Lu Y, Rycaj K, Zhong Y, Chao HP, Calhoun-Davis T, Shen JJ, Tang DG: **Developing a Novel Two-Dimensional Culture System to Enrich Human Prostate Luminal Progenitors that Can Function as a Cell of Origin for Prostate Cancer**. *Stem Cell Transl Med* 2017, **6**(3):748-760.

40. Blattner M, Liu DL, Robinson BD, Huang D, Poliakov A, Gao D, Nataraj S, Deonarine LD, Augello MA, Sailer V *et al*: **SPOP Mutation Drives Prostate Tumorigenesis In Vivo through Coordinate Regulation of PI3K/mTOR and AR Signaling**. *Cancer cell* 2017, **31**(3):436-451.

41. Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM *et al*: **Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy**. *Cancer cell* 2014, **25**(5):638-651.

42. Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, de Hoon M, Severin J, Oki S, Hayashizaki Y *et al*: **Update of the FANTOM web resource: expansion to provide additional transcriptome atlases**. *Nucleic Acids Res* 2019, **47**(D1):D752-D758.

43. Markert EK, Mizuno H, Vazquez A, Levine AJ: **Molecular classification of prostate cancer using curated expression signatures**. *Proc Natl Acad Sci U S A* 2011, **108**(52):21276-21281.

44. Rye MB, Bertilsson H, Andersen MK, Rise K, Bathen TF, Drablos F, Tessem MB: **Cholesterol synthesis pathway genes in prostate cancer are transcriptionally downregulated when tissue confounding is minimized**. *Bmc Cancer* 2018, **18**(1):478.

45. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.

46. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.

47. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**. *Bmc Bioinformatics* 2013, **14**:128.

48. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A *et al*: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update**. *Nucleic Acids Res* 2016, **44**(W1):W90-97.

49. Ehsani R, Drablos F: **Measures of co-expression for improved function prediction of long non-coding RNAs**. *Bmc Bioinformatics* 2018, **19**(1):533.

50. Henry GH, Malewska A, Joseph DB, Malladi VS, Lee J, Torrealba J, Mauck RJ, Gahan JC, Raj GV, Roehrborn CG *et al*: **A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra**. *Cell reports* 2018, **25**(12):3530-3542 e3535.

51. Si M, Lang J: **The roles of metallothioneins in carcinogenesis**. *J Hematol Oncol* 2018, **11**(1):107.

52. Garrett SH, Sens MA, Shukla D, Flores L, Somji S, Todd JH, Sens DA: **Metallothionein isoform 1 and 2 gene expression in the human prostate: downregulation of MT-1X in advanced prostate cancer**. *Prostate* 2000, **43**(2):125-135.

53. Lin SF, Wei H, Maeder D, Franklin RB, Feng P: **Profiling of zinc-altered gene expression in human prostate normal vs. cancer cells: a time course study**. *J Nutr Biochem* 2009, **20**(12):1000-1012.

54. Franklin RB, Milon B, Feng P, Costello LC: **Zinc and zinc transporters in normal prostate function and the pathogenesis of prostate cancer**. *Front Biosci-Landmrk* 2005, **10**:2230-2239.

55. Juang HH, Costello LC, Franklin RB: **Androgen modulation of multiple transcription start sites of the mitochondrial aspartate aminotransferase gene in rat prostate**. *The Journal of biological chemistry* 1995, **270**(21):12629-12634.

56. Zhang S, Zeng X, Ren M, Mao X, Qiao S: **Novel metabolic and physiological functions of branched chain amino acids: a review**. *J Anim Sci Biotechnol* 2017, **8**:10.

57. Ananieva EA, Wilkinson AC: **Branched-chain amino acid metabolism in cancer**. *Current opinion in clinical nutrition and metabolic care* 2018, **21**(1):64-70.

58. Edlind MP, Hsieh AC: **PI3K-AKT-mTOR signaling in prostate cancer progression and androgen deprivation therapy resistance**. *Asian journal of andrology* 2014, **16**(3):378-386.

59. Tayari N, Heerschap A, Scheenen TWJ, Kobus T: **In vivo MR spectroscopic imaging of the prostate, from application to interpretation**. *Anal Biochem* 2017, **529**:158-170.

60. Liu Y, Franklin RB, Costello LC: **Prolactin and testosterone regulation of mitochondrial zinc in prostate epithelial cells**. *Prostate* 1997, **30**(1):26-32.

61. Franklin RB, Juang HH, Zou J, Costello LC: **Regulation of citrate metabolism by androgen in the LNCaP human prostate carcinoma cell line**. *Endocrine* 1995, **3**(8):603-607.

62. Costello LC, Liu Y, Franklin RB: **Testosterone stimulates the biosynthesis of m-aconitase and citrate oxidation in prostate epithelial cells**. *Mol Cell Endocrinol* 1995, **112**(1):45-51.

63. Barabasi AL, Oltvai ZN: **Network biology: Understanding the cell's functional organization**. *Nature Reviews Genetics* 2004, **5**(2):101-U115.