OXFORD

Bioimage informatics

Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks

Philippe Weitz [1], Yinxi Wang[1], Kimmo Kartasalo [1,2], Lars Egevad[3], Johan Lindberg[1,4], Henrik Grönberg[1], Martin Eklund [1] and Mattias Rantalainen[1,5,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden, [2]Faculty of Medicine and Health Technology, Tampere University, 33100 Tampere, Finland, [3]Department of Oncology and Pathology, Karolinska Institutet, 17177 Stockholm, Sweden, [4]Science for Life Laboratory, 17177 Stockholm, Sweden and [5]MedTechLabs, BioClinicum, Karolinska University Hospital, 17176 Stockholm, Sweden

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Abstract

Motivation: Molecular phenotyping by gene expression profiling is central in contemporary cancer research and in molecular diagnostics but remains resource intense to implement. Changes in gene expression occurring in tumours cause morphological changes in tissue, which can be observed on the microscopic level. The relationship between morphological patterns and some of the molecular phenotypes can be exploited to predict molecular phenotypes from routine haematoxylin and eosin-stained whole slide images (WSIs) using convolutional neural networks (CNNs). In this study, we propose a new, computationally efficient approach to model relationships between morphology and gene expression.

Results: We conducted the first transcriptome-wide analysis in prostate cancer, using CNNs to predict bulk RNA-sequencing estimates from WSIs for 370 patients from the TCGA PRAD study. Out of 15 586 protein coding transcripts, 6618 had predicted expression significantly associated with RNA-seq estimates (FDR-adjusted P-value $<1\times10^{-4}$) in a cross-validation and 5419 (81.9%) of these associations were subsequently validated in a held-out test set. We furthermore predicted the prognostic cell-cycle progression score directly from WSIs. These findings suggest that contemporary computer vision models offer an inexpensive and scalable solution for prediction of gene expression phenotypes directly from WSIs, providing opportunity for cost-effective large-scale research studies and molecular diagnostics.

Availability and implementation: A self-contained example is available from http://github.com/phiwei/prostate_coexpression. Model predictions and metrics are available from doi.org/10.5281/zenodo.4739097.

Contact: mattias.rantalainen@ki.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Prostate cancer is one of the most common types of cancer and cause of cancer related deaths in men (Bray *et al.*, 2018). Molecular phenotyping is currently increasing in importance in both the research and clinical settings, as it enables detailed characterization of individual tumours and provides information that enables cancer precision medicine (Collins and Varmus, 2015). Molecular phenotyping can reveal molecular aetiology (Barbieri *et al.*, 2012; Gerhauser *et al.*, 2018; Taylor *et al.*, 2010), predictive and prognostic markers

(Abida *et al.*, 2019; Ren *et al.*, 2018), and enable molecular subtyping (Cancer Genome Atlas Network, 2012; Guinney *et al.*, 2015; Wirapati *et al.*, 2008). Gene expression profiling by RNA-sequencing offers a broad molecular phenotype of prostate cancer (The Cancer Genome Atlas Research Network, 2015; Stelloo *et al.*, 2018). In recent years, several gene expression-based prostate cancer assays for clinical use have been introduced. The Prolaris cell-cycle progression (CCP) score provides an assessment of disease aggressiveness, a 10-year risk of metastasis after therapy, risk of recurrence

after prostatectomy and disease-specific mortality under conservative management based on the mean mRNA expression of 31 genes in either biopsy or prostatectomy tissue (Bishoff *et al.*, 2014; Cooperberg *et al.*, 2013; Cuzick *et al.*, 2012). Other mRNA-based diagnostic tests are the Oncotype Dx genomic prostate score (Cullen *et al.*, 2015; Eure *et al.*, 2017; Klein *et al.*, 2014; Knezevic *et al.*, 2013; Van Den Eeden *et al.*, 2018), the Decipher Biopsy and Post-operative scores (Erho *et al.*, 2013; Marrone *et al.*, 2015; Nguyen *et al.*, 2017). It has also been shown that gene expression is associated with prostate cancer grades (Hamzeh *et al.*, 2019; Penney *et al.*, 2011). However, molecular phenotyping remains costly and time-consuming. There is therefore a demand for tools that can be used to cost-efficiently identify the molecular characteristics of large cohorts of patients retrospectively in research studies, as well as patients in the clinic. This has the potential to identify both novel biomarkers as well as help prioritizing patients that may benefit from more comprehensive molecular phenotyping.

With the advent of digital pathology, where histopathology slides are digitized as part of the routine workflow, computer-based image analysis can now be applied to analyse morphological patterns in histopathology images. It has been demonstrated that computer vision models can be applied to predict molecular characteristics from tissue morphology, including mutations, molecular subtypes (Kather *et al.*, 2020; Schaumberg *et al.*, 2018) and gene expression (Fu *et al.*, 2020; Schmauch *et al.*, 2020). Compared with conventional bulk DNA- or RNA-sequencing, these models also capture spatially resolved intra-tumour heterogeneity (He *et al.*, 2020; Wang *et al.*, 2021). While previous studies have demonstrated the feasibility to predict molecular phenotypes from haematoxylin and eosin (H&E)-stained whole slide images (WSIs), the majority of these models are pan-cancer models (Fu *et al.*, 2020; Schmauch *et al.*, 2020) based on tumours originating from a range of organs. Although it can be assumed that some morphological patterns are shared among these tumours, it is unlikely that morphological patterns in general share their specific association with gene expression features across different cancers. Hence, cancer-specific models are almost certainly required to achieve optimal prediction performance.

To date, no comprehensive analysis of the potential of computer vision models for whole-transcriptome analysis in prostate cancer has been reported. We therefore conducted a transcriptome-wide analysis of gene expression prediction modelling specifically for prostate cancer using data from the TCGA PRAD (The Cancer Genome Atlas Research Network, 2015) study, applying a rigorous performance estimation strategy. We developed a novel computationally efficient modelling approach that exploits the co-expression patterns in gene expression data. This methodology can be deployed on relatively constrained computational infrastructure. Previous studies with this objective either relied on convolutional neural networks (CNNs) as feature extractors, with secondary models fitted to the CNN features, or on single transcript CNNs (Wang *et al.*, 2021). These approaches are either limited in capacity to learn domain-specific representations, or are computationally very costly. We therefore propose to jointly predict individual expressions in clusters of co-expressed (correlated) genes with multi-output models. This allows exploiting potential shared patterns and investigating the possibility of predicting transcripts and pathways that have previously been implicated in prostate cancer. To demonstrate a clinically relevant application, we show that this approach can be applied to predict the prognostic CCP score (Bishoff *et al.*, 2014; Cooperberg *et al.*, 2013; Cuzick *et al.*, 2012).

2 Materials and methods

2.1 Study materials
This study is based on image and expression data from the publicly available TCGA PRAD (The Cancer Genome Atlas Research Network, 2015) dataset, which consists of 403 patients with 449 WSIs of formalin-fixed paraffin embedded H&E-stained sections of resected prostate tumours. These patients originate from 27 cancer centres and organizations, each of which contributed between 1 and 62 patients. From these 403 patients, 399 patients with adenomas and adenocarcinomas were included in this study, whereas 4 patients with ductal and lobular neoplasms were excluded. Of these patients, 389 with matching WSIs and gene expression data from tumour tissue available through TCGA were further selected. For patients with multiple WSIs available, we included one at random. A further nine patients were excluded due to a prior systemic treatment or synchronous malignancies. The patient selection is shown in Supplementary Figure S1. We tiled and preprocessed the WSIs of these 380 cases as described in the Supplementary Materials and Methods, and identified cancer regions using a cancer detection model that we developed with transrectal core needle biopsy data from the STHLM3 prostate cancer diagnostic study (Grönberg *et al.*, 2015; Ström *et al.*, 2020) (see Supplementary Materials and Methods). Supplementary Figure S2 shows the performance of the cancer detection model in the STHLM3 biopsy test set and an annotated subset of the TCGA PRAD study. Further information on the cancer detection model is provided in the Supplementary Materials and Methods. Subsequently, 10 patients whose largest contiguous tumour area was below 1 mm^2 were excluded. The remaining 370 patients were included in this study. For each WSI, we only included tiles that we predicted to be malignant. We then randomly selected 92 (24.86%) of these 370 patients as a held-out test set. To this end, we computed 500 random splits stratified on the International Society of Urological Pathology (ISUP) grading system (Epstein *et al.*, 2016) and selected the split with the best matching age distributions as determined by a Kolmogorov–Smirnov test. The remaining 278 patients, which we will refer to as the development set, were further split into 10 cross-validation (CV) folds. The demographic and clinical characteristics of both the CV set and the held-out test set are provided in Supplementary Table S1.

2.2 Gene selection
The TCGA PRAD RNA expression data includes 60 843 transcripts. The biomaRt (Durinck *et al.*, 2005) *hsapiens_gene_ensembl* currently lists 22 802 transcripts as protein coding. Of these, expression levels were available for 19 601 transcripts in the TCGA PRAD dataset, which were selected for further analyses. We only included genes for which there are at least three counts in at least 10% of patients, since less frequently expressed genes may not be possible to model with the number of samples in this study. This further excludes 4015 transcripts, resulting in a set of 15 586 included transcripts. In subsequent analyses, normalized expression values were used [log2 of the upper quartile normalized fragments per kilobase of transcript per million mapped reads (FPKM-UQ) as preprocessed with HTSeq (Anders *et al.*, 2015)].

2.3 Identification of sets of co-expressed transcripts
In order to reduce the computational complexity of predicting expression levels of 15 586 transcripts, we propose a novel approach based on clustering the transcripts based on their co-expression. Transcripts were assigned to clusters only based on the development data in order to preserve independence of the test set. Clustered transcripts were subsequently jointly predicted with multi-output CNN models, with the expression values of the transcripts in each cluster as the response variables, such that a cluster consisting of *n* transcripts is predicted by a CNN with *n* outputs, one for each transcript. Supplementary Figure S3 shows this modelling approach. Supplementary Figure S4 depicts the number of transcripts initially included in each cluster, the number of transcripts in each cluster brought forward for validation in the test set, and the average absolute Spearman correlation for all gene pairs within the clusters. The clustering is described in more detail in the Supplementary Materials and Methods.

2.4 Model optimization and performance evaluation
We compared the joint cluster prediction with three alternative modelling approaches. In the first, we optimized a CNN to jointly predict the expression of all 15 586 genes in one single model. In the

second approach, we extracted a feature vector for each tile using an ImageNet (Russakovsky *et al.*, 2015) pretrained ResNet18 (He *et al.*, 2016) model and fitted boosting models with *LightGBM* (Ke *et al.*, 2017) (lgbm) to predict gene expression with one boosting model per gene. To reduce the computational cost during model selection, these models were compared in a randomly selected subset of 10 clusters that contains 2636 transcripts. To evaluate the clustering, we randomly reassigned all genes of this subset into 10 random clusters of matched sizes to investigate whether representations learned with the combination of gradients of co-expressed genes yield improved performance compared to a random combination. Furthermore, we optimized single transcript prediction CNNs for a subset of 50 transcripts that were randomly sampled out of the 2636 transcripts. While single transcript CNNs are not a viable option for a transcriptome-wide analysis in this study considering the available computational resources, it is nevertheless an important baseline for interpreting the prediction performance of the proposed method.

For each transcript, we centred and scaled all expression values by the mean and variance of the respective training data before training the corresponding model. We then assigned this slide-level expression as response for all the tiles of the respective slide. The mean prediction across tiles was used to assign a slide-level predicted expression value. In order to preserve the independence of the validation folds and to reduce computational cost, hyperparameters were tuned in five different fold allocations for this subset of 10 clusters for the CNN models, whereas hyperparameters of the single-gene lgbm models were optimized on a random subset of 200 out of these 2636 transcripts. This validation procedure is comparable to a nested 10-fold CV, as shown in Supplementary Figure S5. For 5 of the 10 splits, we excluded the respective outer validation fold, used 2-folds as inner validation folds and 7-folds for model fitting. Predictions for each of the validation folds were concatenated to obtain an independent prediction for each patient in the development set. Further details of the model optimization are provided in the Supplementary Materials.

After selecting the best performing out of the four investigated modelling approaches based on their performance on the outer validation folds, we performed the inner CV with the 40 remaining clusters with the best performing model to determine an optimal resolution out of $40\times$, $20\times$ and $10\times$ for each of the remaining transcripts. We then fitted one CNN for each cluster and resolution level with 9 training folds that include the 2 inner validation folds and the prediction performance of each of the 15 586 transcripts was evaluated on the respective outer validation folds. While each cluster was predicted entirely at every considered resolution level, we only used the prediction at the resolution level that was previously determined as optimal for the respective transcript.

Spearman rank correlations between the slide-level predictions and the RNA-seq expression level values were used as the primary performance metric. Genes with FDR (Benjamini and Hochberg, 1995) adjusted *P*-value <0.0001 were brought forward for validation in the test data. To obtain predictions in the test data, we predicted all test set tiles with all 500 models from the 10-folds and 50 clusters for all 3 resolution levels, and averaged over the 10 predictions (one from each of the 10 CV models) per tile at the resolution level that was selected for each gene.

2.5 Gene set enrichment analysis

Gene set enrichment analysis (Subramanian *et al.*, 2005) (GSEA) was applied to investigate whether any specific biological functions were implicated with transcripts that were associated with morphology. The Reactome (Jassal *et al.*, 2020) pathway knowledge database was used in the analysis together with genes (15 586) ranked by their respective *p*-values from Spearman correlations between CNN predictions and RNA-seq expression estimates. GSEA was performed on *P*-values from the CV data rather than the test data, since ranked enrichment analysis can identify significantly enriched gene sets even if a proportion of the included genes did not meet any significance thresholds.

2.6 CCP score

In order to investigate potential clinical applications of this modelling approach, we computed the CCP score (Bishoff *et al.*, 2014; Cooperberg *et al.*, 2013; Cuzick *et al.*, 2012), both from the TCGA RNA-seq expression data and from model predictions. The CCP is a commercial prognostic test that is intended to support clinical decision making and is computed by taking the mean of 31 highly correlated gene expression levels. We evaluated the prediction performance by computing an RNA-seq-based CCP and assessed the Spearman correlation between this score and a CNN-based score that was computed as the mean of all CCP genes that met the validation criterion for the test set (FDR-adjusted *P*-value <0.0001 in the CV data). In order to evaluate whether the prognostic performance of the CNN-predicted CCP is comparable to the CCP based on the RNA-seq data, we performed univariate hazard analysis with Cox proportional hazard models with time to biochemical recurrence (BCR) as the outcome.

3 Results

We developed and applied a new approach for transcriptome-wide prediction of prostate cancer gene expression using deep CNN models. Prediction performance was validated in a held-out test set.

3.1 Comparison of modelling strategies

We first evaluated four CNN-based modelling approaches for the prediction of gene expression in a subset of 2636 transcripts from 10 randomly drawn clusters (see Section 2). The cluster-based approach, which exploits shared representations for co-expressed genes, achieved the highest average Spearman correlation (0.243) as well as the highest number (1191 out of 2636, 45.18%) of significant correlations (FDR-adjusted *P*-values <0.0001). Predicting genes in randomly assigned clusters resulted in 1030 (39.07%) significant correlations. Fitting lgbm boosting models to ImageNet ResNet18 features with one boosting model per gene or predicting all selected 15 586 genes jointly with a single CNN resulted in 693 (26.29%) and 0 (0%) significant correlations out of 2636 genes, respectively. The distribution of Spearman correlations for each modelling approach is visualized in Figure 1a. The *P*-value from one-sided Wilcoxon rank sum test for the proposed 'corr cluster' method compared to the second-best method, 'rnd cluster', is below 0.0001. Figure 1b shows a comparison of Spearman correlations between a randomly sampled subset of 50 transcripts between the proposed method and CNNs that were optimized to predict single transcripts. The mean difference in Spearman correlation is 0.024. The *P*-value from a paired one-sided Wilcoxon rank sum test that compares the distributions is below 0.01, indicating higher correlations for the correlated cluster CNN. Average training times per gene were also assessed and are depicted in Figure 1c, revealing substantially shorter times for the cluster-based approaches with 11.39 s per transcript, compared to 33.18 s per transcript-wise *lgbm* model. Single transcript CNNs require $\sim$3550 s per transcript.

3.2 Transcriptome-wide prediction of prostate cancer expression values

Based on the model comparison in the previous section, the cluster-based method was selected for the transcriptome-wide analysis across all 15 586 transcripts. First, the prediction performance across all transcripts was assessed in (nested) CV (Fig. 1). Out of the 15 586 predicted gene expression levels, 6618 (42.5%) were associated with the corresponding RNA-sequencing-based estimates [Spearman correlation, FDR-adjusted *P*-value $<1 \times 10^{-4}$, adjustment with the method described by Benjamini and Hochberg (1995)]. The 6618 significant transcripts were brought forward for validation in the held-out test (92 patients). Out of the 6618 transcripts, 5419 (81.9%) had a Benjamini and Hochberg (BH)-adjusted *P*-value <0.01 in the test set. Based on this criterion, the lowest significant correlation between predicted expression and RNA-seq-based expression measurements was 0.274. The distributions of Spearman correlations are depicted in Figure 1a for the entire CV
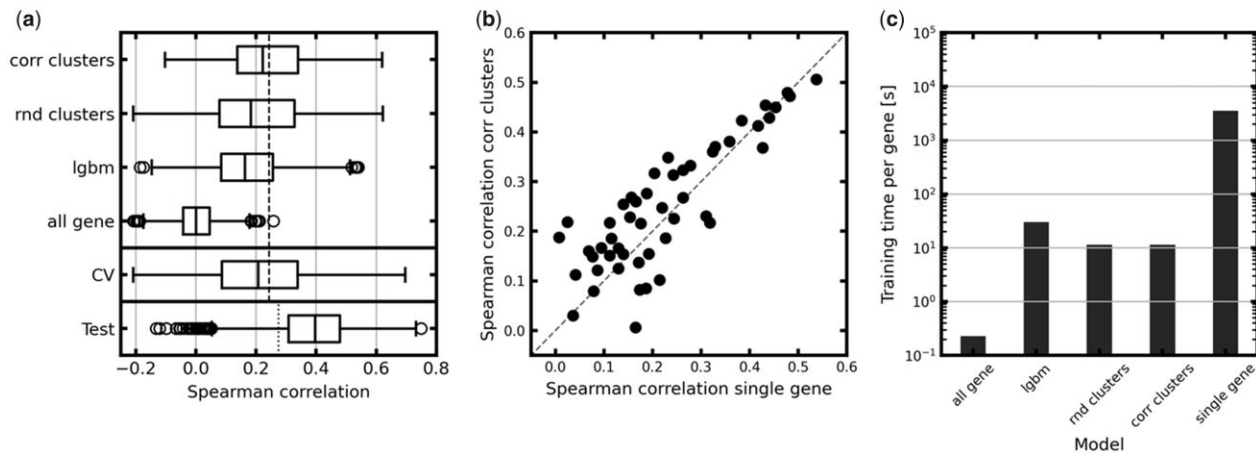
Fig. 1. Performance of modelling approaches. (**a**) Boxplots of distributions of Spearman correlation coefficients for different modelling approaches and validation sets, as well as a comparison of computational efficiency. Vertical dashed lines indicate the significance threshold for adjusted *P*-values of 0.0001 in the validation set and vertical dotted lines indicate the corresponding threshold in the test set of 0.01. corr clusters refers to correlation-based clustering, rnd clusters to random cluster assignments, lgbm to prediction with boosting models based on ResNet18 features and all gene to a cnn that predicts all 15 586 selected genes at once (distribution shown only includes compared 2636 genes). CV denotes the boxplot of Spearman correlations between gene expression and the respective CNN prediction for all 50 clusters comprising 15 586 genes in the validation data, using the corr clusters method. A total of 6618 genes had an adjusted *P*-value lower than 0.0001. Test denotes the boxplot of Spearman correlations of the 6618 selected genes in the held-out test set, with 5419 adjusted *P*-values below 0.01. (**b**) Comparison between a Spearman correlation for 50 randomly sampled transcript that were predicted with single transcript CNNs and the proposed method. (**c**) Comparison of the training time per gene for different modelling approaches. Fitting one CNN per transcript requires ~300 times more training time as compared to the proposed cluster-based method

data and test set, respectively. Supplementary Figure S6a–d shows area under the receiver operating characteristic curves (AUCs), sensitivities and specificities for classification whether expression is higher than the transcript-wise median, as well as a comparison of Pearson and Spearman correlation for the 15 586 transcripts in the CV data and the 6618 selected transcripts in the test data. For a subset of 78 out of the 92 test set cases, PSA, ISUP grade and age are available. When adjusting for these potential covariates with a linear regression model and CNN predictions as the exogenous variable and RNA-seq expression estimates as the endogenous variable, 4690 (70.9%) transcripts out of the 6618 that were brought forward for evaluation in the test set, are statistically significant after BH-adjustment. Out of these, 4512 (68.2%) transcripts satisfy both criteria. In a univariate analysis with linear regression models, 5257 (79.4%) of predicted transcripts were significantly marginally associated with RNA-seq estimates. Supplementary Figure S6e shows a comparison of the Spearman correlations associated with significance determined by correlation and multivariable analysis. All performance metrics for each transcript both for the CV data and the test data are available through the online Supplementary Material. Further details of the multivariable analysis, including an analysis of tumour cellularity, are provided in the Supplementary Materials and Methods. The gene with the highest Spearman correlation between RNA-seq and CNN prediction in the test set was BRICD5, with a correlation of 0.749. Figure 2a shows scatter plots for the gene BRICD5 together with example tiles with low and high predicted expressions (Fig. 2b and c). BRICD5 belongs to the BRICHOS family, which is assumed to act as a chaperone in protein folding (Johansson *et al.*, 2009).

3.3 Genes associated with molecular mechanisms of prostate cancer

Among the significantly predicted transcripts, several of the corresponding genes have previously been reported to be associated with molecular mechanisms of prostate cancer. Out of the 20 genes included in an expression-based androgen receptor (AR) activity score (The Cancer Genome Atlas Research Network, 2015), two were significantly predicted from WSIs: GNMT, and MPHOSPH9 with respective correlations of 0.51, and 0.324. The relationship between predicted and RNA-seq expression estimates for GNMT is shown in Figure 2d, with examples of low and high expression in Figure 2e and f. Further significantly predicted genes in the androgen signalling pathway were NCOR1 (0.468), the gene encoding the AR

(0.322) and NCOA2 (0.31), which has previously been found to be over-expressed in 8% of primary tumours and 37% of metastases (Taylor *et al.*, 2010). FOXA1 and SPOP expression predictions were not significantly associated with their expression (Spearman correlations of 0.013 and 0.22 in CV). However, a human paralog of SPOP, SPOPL, which can act as a negative regulator of SPOP (Clark and Burleson, 2020) was correlated with 0.526.

Expression of the DNA repair genes CDK12 (examples in Fig. 2g–i), which is frequently mutated in metastatic prostate cancer (Grasso *et al.*, 2012), and ATM show Spearman correlations of 0.577 and 0.56 between predicted and RNA-seq expression. The DNA mismatch repair genes MSH2 and MSH6 (0.383 and 0.305) have been found to be frequently mutated in hypermutated microsatellite unstable advanced prostate cancers (Pritchard *et al.*, 2014).

While PTEN did not meet the inclusion criterion due to low expression, multiple established tumour suppressor genes had a significant association between RNA-seq estimates of gene expression and prediction. ZFHX3, which could be predicted with a correlation of 0.6, is a tumour suppressor gene that down-regulates proliferation via MYC in prostate cancer (Hu *et al.*, 2019). Other significantly associated tumour suppressor genes include APC, Rb1, KMT2D and KMT2C, with Spearman correlations of 0.6, 0512, 0.512 and 0.484.

The PI3K pathway is up-regulated in 30–50% of prostate cancers and has been identified as a therapeutic target (Morgan *et al.*, 2009). PIK3CA and PIK3R1 were predicted with Spearman correlations of 0.458 and 0.407. The GTPase HRAS is upstream of the PI3K pathway and has a Spearman correlation of 0.568. MED12 is a subunit of the Mediator kinase complex and is essential in the transcription of protein coding genes. It is frequently over-expressed in castration-resistant distant metastatic and locally recurrent prostate cancers as compared to androgen-sensitive prostate cancers or benign prostatic tissue (Shaikhibrahim *et al.*, 2014) and could be predicted with a Spearman correlation of 0.454.

3.4 Gene set enrichment analysis

GSEA revealed 12 significantly enriched pathways that belong to the functional groups of the cell cycle, RNA metabolism, the immune system, the metabolism of proteins, signal transduction, haemostasis, chromatin organization, the circadian clock and metabolism. Brief description of the identified pathways, their adjusted *P*-values as well as the distribution of Spearman correlations
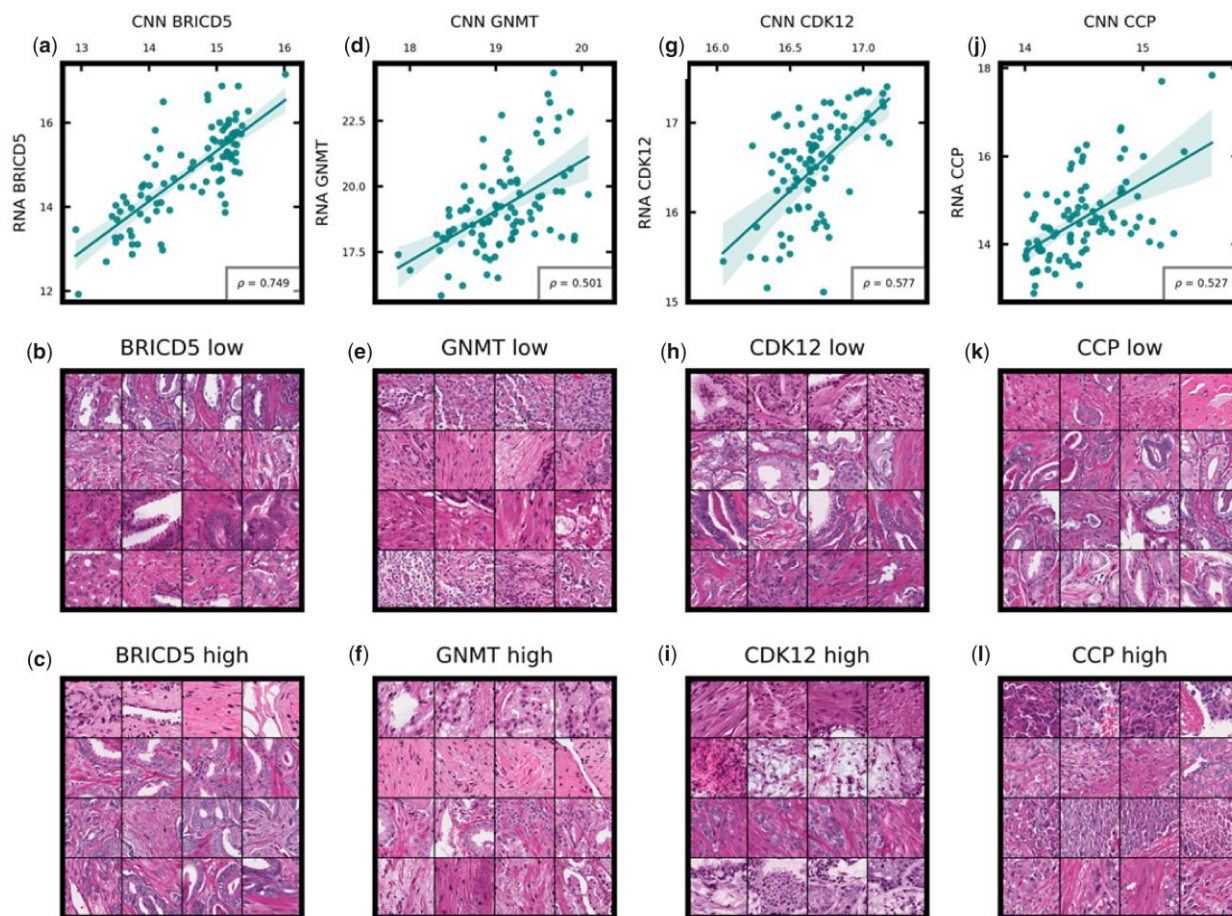
Fig. 2. Comparison between predicted and RNA-seq expression. The lower two rows provide examples of tiles with low and high predicted expression for selected genes. Each panel in the lower two rows contains 16 example images, divided by black lines. Each row in the subplots contains four tiles by the same patient, with four rows corresponding to four different patients. The edge length of each of the 16 tiles is 110.88 μm. (**a**) Scatter plot between CNN prediction and RNA-seq estimates of expression for the best predicted gene BRICD5 with a Spearman correlation of 0.749. (**b**) Examples of tiles with low predicted BRICD5 expression. (**c**) Example tiles with high predicted expression. (**d–f**) Corresponding plots for GNMT with a Spearman correlation of 0.501. GNMT is part of the androgen signalling pathway. (**g–i**) The respective relationship and examples for the DNA repair gene CDK12, with a Spearman correlation of 0.577. The corresponding plots for the CCP score are displayed in (**j–l**), with higher expression being associated with higher proliferation, ISUP grade and poorer prognosis

between CNN predictions and sequenced expression levels are depicted in Supplementary Figure S7. The most significantly enriched pathway, R-HSA-113510 with an adjusted *P*-value of 0.005, regulates DNA replication through the Rb1 E2F pathway. This pathway has previously been found to be frequently mutated in prostate cancer (Grasso *et al.*, 2012). Besides the tumour suppressor gene Rb1, this pathway also contains the CCP gene RRM2, which encodes a reductase that catalyzes the formation of deoxyribonucleotides from ribonucleotides. Both the second and third most strongly associated pathways, R-HSA-6782315 and R-HSA-72200, serve the metabolism of RNA. R-HSA-6782315, with an adjusted *P*-value of 0.07, is involved in tRNA modification in the nucleus and cytosol and has previously been implicated in human diseases, including cancer (Torres *et al.*, 2014).

3.5 CCP score

Of the 31 genes that comprise the CCP, 29 were validated in the test set, which excludes CDC2 and CENPM. We therefore computed a CNN-based CCP score as the average of the 29 remaining CCP genes and compared it with an RNA-based CCP score that is based on all 31 transcripts. The Spearman correlations between the 29 CNN predictions and their RNA expression is depicted in Figure 3a and provided in Supplementary Table S2. The CNN CCP score has a Spearman correlation of 0.527 (bootstrapped 95% CI 0.357, 0.665) with its RNA-seq counterpart (Fig. 2j, examples of low and

high expression in Fig. 2k and l). The corresponding AUC for classifying whether the CCP is expressed above or below its median in the test set is 0.733. Figure 3b reveals a comparable relationship between ISUP grade and ranked CCP score both for the CNN prediction and RNA-seq. BCR is the only outcome with a sufficient number of events for time-to-event analysis in the TCGA PRAD study, with 50 (18%) and 20 (21.7%) patients with BCR events in the CV and the test set, respectively. The HR of the RNA-seq-based CCP was 1.68 (1.256, 2.246) in the CV and 1.351 (0.956, 1.909) in the test data. For the CNN-predicted CCP, the respective HR values were 2.579 (1.412, 4.713) and 2.943 (1.055, 8.212) (Fig. 3c). There is an insufficient number of events for multivariable analysis in the test data. We performed multivariable analysis in a subset of 238 patients from the CV data for which ISUP, PSA and age are available, which includes 50 recurrences. Supplementary Figure S8 shows the multivariable CPH-model coefficients, which are also provided in Supplementary Table S5. Neither the predicted nor the RNA-based CCP are statistically significant in the multivariable analysis. Figure 3d depicts CNN CCP predictions overlayed over representative example WSIs for cases of all ISUP grades.

4 Discussion

In this study, we performed the first transcriptome-wide gene expression prediction specifically for prostate cancer and identified a
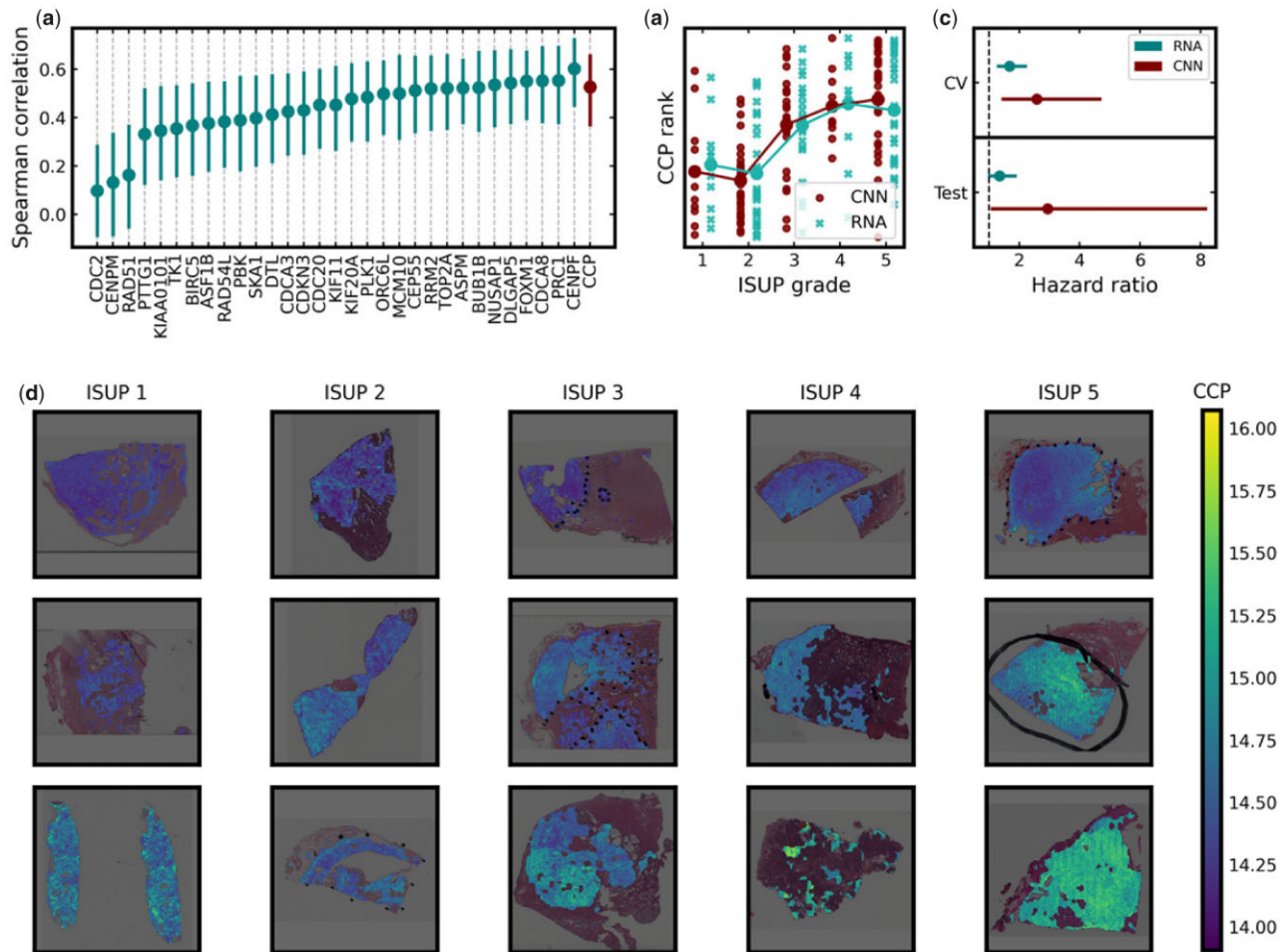
Fig. 3. Comparison between the cell-cycle progression (CCP) score based on RNA-seq and CNN predictions. (**a**) Spearman correlation between sequenced and predicted gene expression in the test set with bootstrapped confidence intervals. (**b**) Ranked CCP scores per ISUP grade both for RNA CCP as well as CNN CCP. (**c**) Univariate hazard analysis for time to first BCR for the RNA-seq-based and predicted CCP score in the CV data and the test set. The HR of the RNA-seq-based CCP is 1.68 (1.256, 4.713) in the CV and 1.351 (0.956, 1.909) in the test data. For the predicted CCP, the respective HR values are 2.579 (1.412, 4.713) and 2.943 (1.055, 8.212) in the test set. (**d**) Examples of WSIs per ISUP grade with overlaid local CCP score predictions. Penmarks in the WSIs originate from the diagnostic workflow before WSI digitization and likely indicate cancer regions

set of 5419 genes whose expression is associated with morphological changes that are detectable by current computer vision models in the TCGA PRAD dataset. We furthermore evaluated this approach to predict a prognostic gene expression-based proliferation score. To this end, we optimized CNN models to predict 15 886 frequently expressed protein coding genes and assessed four different computationally efficient modelling approaches.

As compared to fitting one CNN per gene, the co-expression-based modelling approach proposed here reduces the number of models that need to be fitted from 15 586 to 50, which roughly translates to a 300-fold reduction in computational cost. This increases computational efficiency substantially and reduces hardware requirements and costs, while not reducing prediction performance as compared to CNN models that were optimized to predict single transcripts. Using correlated instead of randomly assigned clusters for joint prediction proved to be a computationally inexpensive way to increase model performance. We speculate that this may be because co-expression of genes is more likely to be associated with similar morphological features and therefore, representations learned in correlated clusters generalize across genes in each cluster. This study therefore provides strong indications that the prediction of transcripts in co-expressed clusters can enable end-to-end CNN model training without loss in performance for transcriptome-wide analyses. As opposed to training secondary models on extracted features, this has the benefit that task-specific representations can be learned, which could further improve prediction

performances particularly compared to secondary models once more training data becomes available.

Previous studies reported prediction of mRNA expression from WSIs of H&E-stained tissue with pan-cancer models, including in the TCGA PRAD cohort (Fu *et al.*, 2020; Schmauch *et al.*, 2020). The study presented by Schmauch *et al.* is difficult to compare to this study since it only relies on CV to assess prediction performance and reports Pearson correlation as the performance metric. Furthermore, the presented results include transcripts that are not known to encode proteins. Generally, the numbers of significantly predicted transcripts are in a similar order of magnitude. A direct comparison to the results by Fu *et al.* reveals a similar number of significantly predicted genes in the TCGA PRAD cohort. While a relatively high number of transcripts are found to be significantly predicted in these studies, effect sizes are relatively small for most transcripts, but for some of the transcripts the effect sizes are expected to be relevant for some purposes. How many of these correlations are sufficiently high to be useful depends on the context of an intended application.

This study has a few limitations. Although our results are based on data from a multi-centre study and while we applied a stringent validation approach with both a fully independent internal test set and a nested CV for model selection, we have not been able to perform validation in a fully independent cohort, since there are currently no additional studies available with both RNA-sequencing data and WSIs. Furthermore, although RNA-seq is now established

for gene expression estimation, orthogonal validation through polymerase chain reaction may be valuable. The size of this study is expected to be a limitation with respect to optimizing the models. We expect model performance to improve with more data both for already significantly predicted transcripts as well as with respect to the number of transcripts that can be predicted accurately. However, there are unknown upper limits to the correlations in this study since the tissue material used for bulk sequencing is not necessarily identical to the tissue sectioned and stained for the WSIs. This limits the correlations both due to noise in labels during training as well as when comparing predicted gene expression to bulk sequencing estimates. We based our models and predictions on regions of high tumour purity by identifying cancer regions with a cancer detection model. This means that the model is only defined for image tiles of cancer tissue and cannot be applied to WSIs of normal tissue sections. However, since the detection model was developed on biopsy data, it required additional calibration in the prostatectomy WSIs and we expect that cancer detection could potentially be improved further. Considering that we found tumour cellularity to not confound gene expression predictions, we nevertheless conclude that the cancer detection model is a useful component of the modelling approach.

In the set of genes that were significantly predicted in this study, there were many genes that are implicated in prostate cancer. Particularly, the expression of genes of the cell cycle and of genes involved in proliferation, such as the genes of the CCP score were predicted significantly. Transcripts of known tumour suppressor and DNA repair genes CDK12, ATM, Rb1, KMT2D and ZFHX3 were also predicted with high correlations. However, a surprisingly low number of genes from the androgen signalling pathway had a significant correlation between prediction and gene expression, despite the central role of androgen in prostatic carcinogenesis, with the exception of GNMT and a few other genes. Based on this, we can speculate that gene expression activity in the androgen signalling pathway has limited impact on tissue morphology. We identified 12 pathways that are enriched for genes that could be predicted from WSIs, including those related to cell cycle, metabolism of RNA and proteins, the immune system and signal transduction based on ranked GSEA. Some of these pathways had previously been implicated in prostate cancer. Further investigation into the relationship between the differential expression of the significantly correlated genes and their associated morphology may yield novel biological insight or candidates for diagnostic, prognostic or predictive biomarkers. Potential clinical use of computer vision-based gene expression prediction was investigated through an analysis of the prognostic CCP score. Rank-based analysis revealed that the predicted CCP score has a similar relationship to the ISUP grade as the sequencing-based score. Univariate time-to-event analysis with BCR as outcome revealed that both the RNA-seq-based and the CNN-predicted CCP were prognostic in the CV analysis, whereas only the CNN-predicted CCP was prognostic in the test set. This analysis was, however, based on a relatively low number of events and patients. Prediction of molecular phenotypes and cell-cycle score from histopathology images may prove clinically useful in low-resource environments in which molecular diagnostics are unavailable, or to analyse large cohorts of patients for which sequencing is too costly, including large-scale studies of archived slides that may not be suitable for RNA-sequencing.

In conclusion, our findings indicate that the expression of a large number of genes is significantly associated with morphological patterns. While considering the limitation that only approximate prediction of gene expression levels is possible from histopathology images, this study provides further evidence of a strong association between routine clinical H&E-stained histopathology slides and average tumour gene expression. We conclude that contemporary computer vision models offer an inexpensive and scalable solution for prediction of gene expression phenotypes directly from WSIs, providing opportunity for cost-effective large-scale research studies and molecular diagnostics.

Data availability

WSIs, RNA-seq and clinical data for the TCGA PRAD cohort are publicly available from the GDC data portal (https://portal.gdc.can cer.gov/). Information on biochemical recurrence was obtained from xenahubs (https://xenabrowser.net/datapages/?dataset=TCGA.PRAD.sample Map/PRAD_clinicalMatrix&host=https://tcga.xenahubs.net). All patient-level expression predictions and computed correlations, as well as adjusted *P*-values, are available from https://zenodo.org/record/4739097#.YouJcu5Bwfs.

References

Abida,W. *et al.* (2019) Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci. USA*, **116**, 11428–11436.

Anders,S. *et al.* (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Barbieri,C.E. *et al.* (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.*, **44**, 685–689.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, **57**, 289–300.

Bishoff,J.T. *et al.* (2014) Prognostic utility of the cell cycle progression score generated from biopsy in men treated with prostatectomy. *J. Urol.*, **192**, 409–414.

Bray,F. *et al.* (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, **68**, 394–424.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Clark,A. and Burleson,M. (2020) SPOP and cancer: a systematic review. *Am. J. Cancer Res.*, **10**, 704–726.

Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

Cooperberg,M.R. *et al.* (2013) Validation of a cell-cycle progression gene panel to improve risk stratification in a contemporary prostatectomy cohort. *JCO*, **31**, 1428–1434.

Cullen,J. *et al.* (2015) A biopsy-based 17-gene genomic prostate score predicts recurrence after radical prostatectomy and adverse surgical pathology in a racially diverse population of men with clinically low- and intermediate-risk prostate cancer. *Eur. Urol.*, **68**, 123–131.

Cuzick,J. *et al.*; on behalf of the Transatlantic Prostate Group. (2012) Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. *Br. J. Cancer*, **106**, 1095–1099.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Epstein,J.I. *et al.*; Grading Committee. (2016) The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am. J. Surg. Pathol.*, **40**, 244–252.

Erho,N. *et al.* (2013) Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One*, **8**, e66855.

Eure,G. *et al.* (2017) Use of a 17-gene prognostic assay in contemporary urologic practice: results of an interim analysis in an observational cohort. *Urology*, **107**, 67–75.

Fu,Y. *et al.* (2020) Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer*, **1**, 800–810.

Gerhauser,C. *et al.* (2018) Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell*, **34**, 996–1011.e8.

Grasso,C.S. *et al.* (2012) The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, **487**, 239–243.

Grönberg,H. *et al.* (2015) Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol.*, **16**, 1667–1676.

Guinney,J. *et al.* (2015) The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, **21**, 1350–1356.

Hamzeh,O. *et al.* (2019) A hierarchical machine learning model to discover Gleason Grade-Specific biomarkers in prostate cancer. *Diagnostics*, **9**, 219.

He,B. *et al.* (2020) Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.*, **4**, 827–834.

He,K. *et al.* (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *Las Vegas, NV, USA*. pp. 770–778.

Hu,Q. *et al.* (2019) ZFHX3 is indispensable for ERβ to inhibit cell proliferation via MYC downregulation in prostate cancer cells. *Oncogenesis*, **8**, 28.

Jassal,B. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Johansson,H. *et al.* (2009) The Brichos domain of prosurfactant protein C can hold and fold a transmembrane segment. *Protein Sci.*, **18**, 1175–1182.

Kather,J.N. *et al.* (2020) Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer*, **1**, 789–799.

Ke,G. *et al.* (2017) LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, **30**, 3146–3154.

Klein,E.A. *et al.* (2014) A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur. Urol.*, **66**, 550–560.

Knezevic,D. *et al.* (2013) Analytical validation of the oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genomics*, **14**, 690.

Marrone,M. *et al.* (2015) A 22 gene-expression assay, Decipher® (GenomeDx Biosciences) to predict five-year risk of metastatic prostate cancer in men treated with radical prostatectomy. *PLoS Curr.*, **7**. http://currents.plos.org/genomictests/index.html%3Fp=23022.html

Morgan,T.M. *et al.* (2009) Targeted therapy for advanced prostate cancer: inhibition of the PI3K/Akt/mTOR pathway. *Curr. Cancer Drug Targets*, **9**, 237–249.

Nguyen,P.L. *et al.* (2017) Ability of a genomic classifier to predict metastasis and prostate cancer-specific mortality after radiation or surgery based on needle biopsy specimens. *Eur. Urol.*, **72**, 845–852.

Penney,K.L. *et al.* (2011) mRNA expression signature of Gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.*, **29**, 2391–96.

Pritchard,C.C. *et al.* (2014) Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. *Nat. Commun.*, **5**, 4988.

Ren,S. *et al.* (2018) Whole-genome and transcriptome sequencing of prostate cancer identify new genetic alterations driving disease progression. *Eur. Urol.*, **73**, 322–339.

Russakovsky,O. *et al.* (2015) ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, **115**, 211–252.

Schaumberg,A.J. *et al.* (2018) H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *bioArxiv*, **064279**.

Schmauch,B. *et al.* (2020) A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.*, **11**, 3877.

Shaikhibrahim,Z. *et al.* (2014) MED12 overexpression is a frequent event in castration-resistant prostate cancer. *Endocr. Relat. Cancer*, **21**, 663–675.

Stelloo,S. *et al.* (2018) Integrative epigenetic taxonomy of primary prostate cancer. *Nat. Commun.*, **9**, 4900.

Ström,P. *et al.* (2020) Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet. Oncol.*, **21**, 222–232.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Taylor,B.S. *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell*, **18**, 11–22.

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

Torres,A.G. *et al.* (2014) Role of tRNA modifications in human diseases. *Trends Mol. Med.*, **20**, 306–314.

Van Den Eeden,S.K. *et al.* (2018) A biopsy-based 17-gene genomic prostate score as a predictor of metastases and prostate cancer death in surgically treated men with clinically localized disease. *Eur. Urol.*, **73**, 129–138.

Wang,Y. *et al.* (2021) Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression-morphology analysis in breast cancer. *Cancer Res.*, **81**, 5115–5126.

Wirapati,P. *et al.* (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.*, **10**, R65.