Integrating weighted correlation network analysis and machine learning identifies common trajectories of prostate cancer

Raheleh Sheibani-Tezerji[1,2]*, Carlos Uziel Perez Malla[1,2]*, Anna Malzer[1,2], Katarina Misura[1,2], Jessica Kalla[2], Loan Tran[1,2], Gabriel Wasinger[2] and Gerda Egger[1,2,3]**

[1]Ludwig Boltzmann Institute Applied Diagnostics, Vienna, Austria

[2]Department of Pathology, Medical University of Vienna, Vienna, Austria

[3]Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria

*equal contribution

**Correspondence to: gerda.egger@meduniwien.ac.at

Abstract

The identification of molecular determinants of prostate cancer development and progression represents a major clinical and biological undertaking. We integrated transcriptomic data of primary localized and advanced prostate cancer from two cancer databases. Transcriptomic analysis showed upregulation of genes encoding for cell surface proteins in metastatic tumors, which were associated with cell-matrix components and chemokine signaling. Integration of machine learning and weighted correlation network modules identified the EZH2-TROAP axis as the main trajectory from initial tumor development to lethal metastatic disease. In addition to known biomarkers, we identified several targets including *HROB* and *GEN1* that are implicated in DNA damage repair pathways or prognostic markers including *SBK1*, *DLX2* and *E2F*2, whose biological role has not yet been elucidated for prostate cancer. Our results demonstrate the usability of complex bioinformatics approaches to identify biological drivers of prostate cancer progression and their suitability for clinical applications as prognostic biomarkers.

Introduction

Prostate cancer (PCa) is the second most common cancer in men with 1.41 million cases worldwide in 2020, accounting for an incidence of 23.6 and mortality of 3.4 per 100,000 cases (Sung *et al*, 2021). Risk factors include age, family history, race, germ line mutations and clinical predisposition (e.g. Lynch syndrome) (Sung *et al.*, 2021). Clinically, PCa can range from indolent and slowly progressive disease, which can be cured with surgery and/or radiation therapy, to aggressive disease including metastatic castration-resistant PCa (mCRPC) (Shen & Abate-Shen, 2010). Approximately 20-40% of patients experience biochemical recurrence (BCR) defined by rising prostate specific antigen (PSA) levels after radical prostatectomy due to local recurrence or metastasis. Metastasis most frequently involves bone, distant lymph nodes, liver and lungs and represents the main cause of disease related mortality.

PCa is highly heterogeneous and similar to other cancers, higher risk of PCa and its progression is led by somatic or germline genetic and epigenetic aberrations (Abeshouse *et al*, 2015). COSMIC data on mutation frequency in PCa patients highlights mutations in *LRP1B* (38%), *TP53* (23%), *FHIT* (23%), *ERBB4* (22%), *CAMTA1* (20%) as the top five recurring, together with mutations in *AR*, *SPOP*, *PTEN*, *ATM*, *SMARCA4* and others (Forbes *et al*, 2010). Whole exome sequencing studies identified multiple altered genes and pathways in PCa, including androgen signaling, DNA damage repair (DDR) and DNA mismatch repair genes (MMR), together with phosphoinositide 3-kinase (PI3K)/AKT signaling, the WNT/β-catenin pathway and MAPK signaling. Furthermore, DDR defects with *BRCA1*, *BRCA2* and *ATM* mutations being the most common to occur, were present in 10% of the primary tumors but in 25% of metastases, highlighting that pathways were more frequently changed in metastatic lesions compared to primary tumors (Abeshouse *et al.*, 2015; Armenia *et al*, 2018; Robinson *et al*, 2015). Additionally, mutations in *TP53*, *AR*, *PTEN*, *APC* and *STAT3* are common in PCa metastases (Abdulghani *et al*, 2008; Armenia *et al.*, 2018). Several, integrative molecular

studies of primary and metastatic PCa proposed classification of PCa into molecular groups related to disease progression and clinical outcome (Abeshouse *et al.*, 2015; Abida *et al*, 2019; Kamoun *et al*, 2018; Robinson *et al.*, 2015; Taylor *et al*, 2010; Varambally *et al*, 2005).

Currently, only a handful of PCa biomarkers are in clinical use. The most prominent is prostate specific antigen (PSA), which is used for screening, diagnosis, monitoring and risk prediction of PCa. PSA screening has limitations due to overdiagnosis and overtreatment, leading to unnecessary prostate biopsies and aggressive treatment regardless of the risk (Loeb *et al*, 2014). Additionally, prostate-specific membrane antigen (PSMA), a transmembrane protein, is highly overexpressed in PCa (100- to 1000-fold) and is a theranostic target of PCa, both for diagnosis and treatment in nuclear medicine (Adnan & Basu, 2023). Expression levels of PSMA are positively correlated with more aggressive disease, high PSA, high Gleason scores and early recurrence.

PCa risk stratification at diagnosis and treatment decisions are currently based on clinical parameters including Gleason score, PSA level and tumor staging. In addition, several risk stratification methods to predict BCR, metastasis or for therapy stratification have been tested and validated based on gene expression signatures, some of which are already in clinical use (Cuzick *et al*, 2011; Erho *et al*, 2013; Klein *et al*, 2014).

Numerous studies have performed meta-analyses of publicly available datasets and databases such as The Cancer Genome Atlas (TCGA) (Weinstein *et al*, 2013). Aside from biomarker discovery, access to public data provides significant insight into understanding biological pathways, modifications and functions of established biomarkers in certain patients. *In silico* data mining is a useful tool to decipher relevant disease biomarkers and to generate research hypotheses. In particular, various machine (deep) learning algorithms have been developed and also applied to PCa data, which have high potential to interrogate large amounts of multi-

dimensional data (Elmarakeby *et al*, 2021; Goldenberg *et al*, 2019; Inza *et al*, 2010; Larrañaga *et al*, 2006; Sapoval *et al*, 2022).

Here, we integrated publicly available gene expression datasets of localized PCa with matched adjacent normal tissue from The Cancer Genome Atlas (TCGA) prostate adenocarcinoma (PRAD) dataset (Weinstein *et al.*, 2013) with mCRPC samples from two recent studies (Abida *et al.*, 2019; Robinson *et al.*, 2015), which were accessed via the database of Genotypes and Phenotypes (dbGap) (Mailman *et al*, 2007). In order to identify prognostic biomarkers for PCa development and disease progression to metastasis we performed pair-wise differential gene expression analysis identifying deregulated genes between localized primary PCa (prim) and normal adjacent tissues (norm), as well as between primary PCa and mCRPC (met) specimens. Differentially expressed genes were subsequently used to infer significant targets based on several computational methods including pathway analysis, random forest machine learning, weighted correlation network analysis (WGCNA) and logistic regression models. A distinct focus was set on genes that encode for cell surface and extracellular proteins, in order to identify potential diagnostic and therapeutic targets. Using this strategy, we identified several well established prognostic PCa biomarkers in addition to novel targets, which showed association with DNA damage repair pathways, cell cycle and correlated disease-free survival of the patients. These markers might represent important biological drivers of PCa progression and prove useful as diagnostic and prognostic markers as well as therapeutic targets in the future.

Results

To infer PCa specific gene expression signatures that reflected general transcriptional trajectories of PCa development and progression irrespective of molecular subtype, we considered RNAseq data of 500 primary prostate adenocarcinomas with 52 matched adjacent

normal tissue samples from the PRAD dataset of the TCGA (Weinstein *et al.*, 2013), and 57 mCRPC samples from the dbGAP databases, which were recently published (Abida *et al.*, 2019; Robinson *et al.*, 2015) (Fig 1A and Table EV1). Metastatic samples were derived from different sites including lymph nodes, bone, liver and other soft tissues. (Fig 1B).

Global scaling of gene expression profiles unveils (dis)similarities between normal and tumor samples

In order to assess the relation of the different sample types based on their gene expression signatures, we applied principal component analysis, which confirmed that most of the variation in expression levels was based on the different sample groups. We identified a clear separation of metastatic (met) samples from primary (prim) and normal adjacent (norm) samples, whereas some overlap was apparent between norm and prim tissues, as indicated by ~21% of data variance of the two major principal components (PC1=~12%, PC2=~9%) (Fig 2A). The remaining components were sorted in descending order of their contribution to the variance, (PC3= 7%, PC4=4%) (Fig EV1). Additionally, unsupervised hierarchical clustering of normalized gene expression data showed as well a clear separation of normal and metastatic samples from primary tumors (Fig 2B).

Identification of differentially expressed genes

Next, to get a deeper understanding of the transcriptional changes between the three groups including norm, prim and met samples, we performed pairwise differential gene expression analysis in the combined TCGA and dbGAP datasets (Fig 3A). For consecutive downstream analyses, we only considered differentially expressed genes (DEGs) with unique gene IDs (e.g ENTREZID) and, unless otherwise specified, filtered by Padj < 0.05 and absolute LFC > 1.

The prim/norm analysis resulted in 3,554 DEGs, 1,670 of which were up- and 1,884 were down-regulated (Table EV2A). In the met/prim comparison we found 6,230 significantly deregulated genes with 4,532 up- and 1,698 down-regulated (Table EV2B). Interestingly, 606

up- and 350 down-regulated DEGs were shared between prim/norm and met/prim, respectively, which might include genes important both for tumor initiation and progression (Table EV2, C and D).

To define top gene expression signatures specific for each sample type, we performed hierarchical clustering analyses using DEGs identified in pairwise comparisons of prim/norm and met/prim samples. Furthermore, samples were grouped into two individual sample classes by consensus partitioning. For the prim/norm comparison, some normal samples overlapped with primary tumors, whereas the metastatic samples were more clearly separated from the primary tumors in individual subgroups. The DEGs with zero variance and too low variance (sd <= 0.05 quantile) were removed, resulting in 3,093 and 5,367 signature genes from prim/norm and met/prim, respectively, which were grouped into two separate gene expression clusters in both analyses (Fig 3, B1 and C1).

To explore the systematic features and biological functions of the differentially expressed signature genes, gene ontology (GO) enrichment analysis was applied to the two gene clusters of prim/norm and met/prim (Table EV3, A and B). Significant GO terms (Padj < 0.05) were then clustered based on their similarities. Furthermore, GO enrichment results were visualized as word clouds to summarize the biological functions in each cluster (Fig 3, B2 and C2). Pathways associated with development were enriched from down-regulated genes both in prim/norm and met/norm, whereas GO terms associated with cell cycle and mitosis were both enriched for up-regulated genes in both pairwise comparisons. Furthermore, metabolic processes and signaling pathways were enriched in up- and down-regulated gene expression clusters of both comparisons. Together, these data indicate that similar pathways are affected by differentially expressed genes in the transition from normal to primary tumors and to metastatic disease.

Functional gene set association analysis of DEGs

To identify candidate diagnostic or prognostic PCa biomarkers as well as putative therapeutic targets, we subsequently focused our analyses only on up-regulated DEGs (Padj < 0.05 and LFC > 1) from the pairwise comparisons prim/norm and met/prim. Ingenuity Pathway Analysis (IPA) (Krämer *et al*, 2014) identified several known PCa signaling pathways including TGF-ß, p38-MAPK, cAMP and PKA signaling in prim/norm samples and cAMP, PKA, PI3K-AKT, CREB and G-protein coupled receptor signaling in the met/prim analyses (Table EV4, A and B). Using the biomarker analysis tool included within the IPA software, a total of 34 genes were identified from the up-regulated DEGs of prim/norm samples and 114 from the up-regulated DEGs of met/prim tumors (Table EV5, A and B), 21 of which were found in both comparisons (Table EV5C). Among the biomarker candidates we found cytokines and growth factors, enzymes and peptidases, transmembrane receptors and transporters as well as transcription regulators and other mostly secreted factors. Interestingly, several factors associated with innate immunity and acute phase response proteins including albumin, C-reactive protein, fibrinogen, prothrombin, or haptoglobin were up-regulated in primary and metastatic PCa. Some of these factors were already suggested as biomarkers for non-invasive diagnostics in liquid biopsies (Arthur *et al*, 2018; Wang & Xing, 2021). Importantly, several of the identified biomarkers represent drug targets, some of which are already in clinical use or tested in clinical studies including androgen receptor (AR) targeting drugs, EZH2 inhibitors, matrix metallopeptidases (MMP), transthyretin (TTR) or the RET proto oncogene.

For prim/norm samples, we identified two biomarkers associated with folate cleavage (*FOLH1*) and transport (*SLC19A1*), which might indicate an important dependency of PCa on folate metabolism. In fact, the *FOLH1* gene encodes for the prostate specific membrane antigen PSMA, which is an established tracer for positron emission tomography/computed tomography (^{68}Ga PSMA PET/CT) and for radionuclide therapy (^{177}Lu PSMA) (Adnan & Basu, 2023). In the met/prim we additionally identified *ADAM12*, a member of the A disintegrin and

metalloproteases (*ADAM*) protein family, which is frequently up-regulated in solid cancers and involved in cell migration and invasion (Wang *et al*, 2018).

Cell surface proteins such as receptors or transporters represent suitable targets for specific diagnostic and therapeutic applications such as molecular imaging or targeted therapies. Therefore, we utilized the Cell Surface Protein Atlas tool (CSPA) (Bausch-Fluck *et al*, 2015) in order to identify up-regulated cell surface proteins (CSPs) in the pairwise comparisons described above. The CSP analysis resulted in a total of 246 proteins, of which 201 were significantly up-regulated in met/prim tumors and 45 were significantly up-regulated in prim/norm samples (Table EV6, A and B). In addition, 22 target genes were found to overlap in both comparisons (Table EV6C), suggesting an important role of these factors for tumor progression and potentially prognosis. Metastatic tumors showed top up-regulated CSPs including several cell adhesion proteins, proteins regulating cell-cell and cell-matrix interactions or migration, as well as chemokine receptors (Fig 4A). The up-regulation of matrix associated factors including *OLFML2A/B*, *ADAM12*, *NTN3, LRRC15/32, CDH17, PCDH19, VTN, ICAM5* highlight the relevance of the tumor microenvironment for progression to metastatic disease. Interestingly, several of the identified genes are related to neuronal development or function including *NTSR1, NTN3, SERPINE2, SERPINI1.*

To gain further insight into the functional aspects of up-regulated CSPs mainly in metastatic tumors, the identified genes were used for interactive pathway analysis (Paczkowska *et al*, 2020). The defined active pathways were used as input for the enrichment map analysis (Merico *et al*, 2010) of Cytoscape (Cline *et al*, 2007) for network visualization of similar pathways. Auto-annotated enrichment maps of up-regulated CSPs highlighted their potential functional implication for PCa. The pathways were grouped into four major classes including cell adhesion and migration, signaling, transporters and receptors and regulation of neuron projection (Fig 4B). Together these data suggest that PCa metastasis is highly dependent on

proteins that are expressed at the surface of tumor cells or factors secreted into the extracellular space, which might represent efficient diagnostic and therapeutic targets for aggressive PCa.

Identifying the most relevant DEGs using explainable machine learning yields multiple cancer-related biomarkers

In order to reduce the number of DEGs and sort them by their importance to aid in biomarker identification, we trained explainable machine learning binary classifiers to distinguish between pairs of sample types, as defined by our experimental subgroups. We used shapely additive explanations (SHAP) values (Scott M. Lundberg, 2017), averaged over 10,000 bootstrap training iterations of random forests, to rank our up-regulated DEGs. SHAP values, which are generated for each sample and gene individually, indicate the impact that each gene has on the final model output, with higher absolute values demonstrating an increased influence. We achieved a mean validation F1-score across all bootstrap iterations of 0.98 ± 0.01 for prim/norm and 0.99 ± 0.01 for met/prim.

Among the top ranked genes, we found relevant known cancer and PCa biomarkers for both examined comparisons, validating the suitability of our approach (Fig 5, A and B). For the prim/norm comparison, there were a total of 127 up-regulated DEGs above a SHAP threshold of 0.001 (Table EV7A). Among the top genes, we identified known PCa genes such as *EPHA10, DLX2, HPN, HOXC4, HOXC6* and *AMACR*. *EPHA10* encodes for an ephrin receptor, whose overexpression in PCa might have potential as a target for therapy (Nagano *et al*, 2014). *DLX2* is a novel marker of increased metastasis risk (Green *et al*, 2016). *HPN* is a type II transmembrane serine protease and a marker to distinguish normal tissue from PCa lesion validated by immunostaining (Ma *et al*, 2020). *HOXC4* and *HOXC6* are homeobox transcription factors commonly detected in PCa that co-localize with *HOXB13*, *FOXA1* and *AR*, three other transcription factors previously shown to contribute to the development of PCa

(Luo & Farnham, 2020). Finally, the racemase *AMACR* is a known good predictor of clinically significant PCa (Kotova *et al*, 2020).

For met/prim tumors, there were a total of 117 up-regulated DEGs above a SHAP threshold of 0.001 (Table EV7B). Among the top genes, we identified genes related to other cancers such as *NPFF*, a neuropeptide that was found to be a key biomarker for retinoblastoma after being identified as a hub gene in the protein-protein interaction (PPI) network (Zhao *et al*, 2021), and *PMM2*, an enzyme involved in glycosylation, and a prognostic marker for colon cancer (Cui *et al*, 2021).

To answer the question of whether there are common biomarkers in PCa progression, we calculated the overlap between the up-regulated DEGs above a SHAP threshold of 0.001 for met/prim and prim/norm, which resulted in two potential biomarkers: *EZH2* and *TROAP* (Table EV7C). The role of the polycomb repressive complex PRC2 protein *EZH2* has been well established for PCa progression to castration resistant prostate cancer or lineage plasticity of neuroendocrine PCa (Davies *et al*, 2021; Varambally *et al*, 2002). EZH2 also acts in a PRC2 independent fashion as an *AR* co-activator or by transcriptional activation of *AR* (Xu *et al*, 2012). On the other hand, *TROAP* is the trophinin associated protein that has been found to regulate PCa progression via the WNT3/survivin signaling pathways due to its essential role in cell proliferation (Ye *et al*, 2018).

We also compared our top DEGs (SHAP > 0.001) to 30 previously published gene expression signature sets which were recently re-evaluated by different ML tools for their prognostic potential (Li *et al*, 2022). *EZH2* and *TROAP*, were among the signature genes. Additionally, we identified 18 DEGs from prim/norm and 14 DEGs from the met/prim comparison (Table EV7D). Thus, our ML approach provided a significant number of candidate biomarker genes, some of which were already reported for their prognostic potential.

We additionally discovered relevant cell surface receptors among our top DEGs (SHAP > 0.001). In prim/norm we found *FGFRL1*, *LRFN1*, *NUP210*, *SMPDL3B* and *TMEM132A*, whereas for met/prim we found *ADAM12*, *MFAP3* and *YBX1*.

Clustering of DEGs based on weighted gene co-expression networks serves to identify groups of similar biological processes related to PCa

Next, we analyzed co-expression relationships between up-regulated DEGs to get further biological insights, by grouping genes that act together in order to perform similar biological functions. For prim/norm, we identified a total of six modules of varying sizes as well as hub genes in each module according to their node degree (Fig 6, A-C, Table EV8, A-F). We were able to identify genes related to PCa among the module hub genes, such as *TPX2* in M3, which enables importin-alpha family protein and protein kinase binding activity and is highly expressed in high-grade PCa as well as being significantly related to poor prognosis (Zhang *et al*, 2021), and *ERG* for M4, a gene linked to prostate cancer due to its overexpression in the presence of androgens (Adamo & Ladomery, 2016). Additionally, many of the most relevant genes in M3 (indicated by node size in Fig 6D) were previously identified by our ML analyses and showed high SHAP values (Fig 5B), such as *HOXC4*, *HOXC6*, *AMACR* and *DLX2*, increasing the relative biological significance of this module.

To determine the biological processes in which the co-expressed genes are involved in, we performed functional enrichment analysis for the Reactome dataset on the genes in each module, which we summarized using wordclouds (Fig 6D). Among the top enriched pathways in module M2, which was defined by *MRPL41* as a hub gene, we found processes related to chromatin, epigenetic regulation and DNA repair (Table EV9A). Importantly, WNT signaling, which is a key driver of PCa tumorigenesis, metastasis and therapy resistance was also represented within this module. Regarding the enriched pathways in M3 (Table EV9B), we identified a common theme surrounding mitosis and cell division. Additionally, we detected

transcriptional regulation by TP53, which is one of the most common genetic mutations in PCa (Ecke *et al*, 2010).

For met/prim we identified a total of nine modules of varying sizes as well as hub genes in each module defined by node degree (Fig 7, A-C, Table EV10, A-I). We again discovered genes related to PCa among the module hub genes, such as *FAM229A* in M5, whose knockout led to enhanced potential of invasion and metastasis (Wang *et al*, 2022b) and *TPX2* in M4. It is worth mentioning that M3 from prim/norm and M4 from met/prim not only shared the same hub gene, but also 112 DEGs (26% of all genes present), which is a considerable amount (Table EV10J) and suggests that both modules encode similar biological functions.

Reactome analysis (Gillespie *et al*, 2021) of the genes in each module identified top enriched pathways associated with the extracellular matrix remodeling in module M1, immune system related pathways in M3, or respiratory electron transport in M5 (Fig 7D, Table EV11, A-E). Interestingly, module M4 in met/prim contained all 23 pathways identified for M3 in prim/norm, and additional pathways associated with mitotic and cell cycle processes and pathways related to TP53. Moreover, in M4 the gene ratio in these shared pathways (i.e. the fraction of DEGs found in the gene set) was considerably bigger, which implied an increased activity of these pathways in metastatic tumors (Table EV11F).

Finally, we identified cell surface receptors within the most interesting modules. In prim/norm M3 we found *ASPN*, whereas for met/prim M4 we found *F5* and *TNFSF18*. Receptors *CELSR3, COL11A1, CSPG5, NTM* and *TMEM132A* were present in both pairwise comparisons.

Together, these data highlight the potential of co-expression analysis for the identification of relevant modules that reflect major biological pathways promoting for malignant transformation, tumor progression and metastasis. As a central theme, these analyses manifested the importance of transcriptional programs regulating mitosis and cell cycle control for cancer development and progression.

Integrative analysis for the mining of PCa candidate biomarkers

In order to obtain a conclusive, robust estimation of potential PCa biomarkers, we combined the outputs of our methods, forming an integrative approach to biomarker discovery for primary and metastatic prostate tumors.

First, we looked at the overlap between WGCNA and ML analyses focusing on modules M3 from prim/norm and M4 from met/prim, which shared the most biological functions. We considered ML genes above a SHAP value of 0.001 for the corresponding comparison. Interestingly, this was equal to the overlap between the genes above a SHAP value of 0.001 for both comparisons and also resulted in *EZH2* and *TROAP*, again highlighting the central role of these two proteins for PCa progression. Both genes, which were significantly up-regulated in the pairwise comparisons, showed a gradual increase of expression from normal to primary and metastatic samples (Fig 8A). Additionally, we evaluated the capability of these biomarkers with a simple logistic regression model using 10-fold cross-validation to predict the sample type (norm/prim/met) in each comparison (Table EV12A). Validation metrics, such as the F1-score, showed that both *EZH2* (prim/norm=0.93 ± 0.03, met/prim=0.94 ± 0.03) and *TROAP* (prim/norm=0.88 ± 0.03, met/prim=0.93 ± 0.02) are highly suitable predictors of localized or metastatic tumors.

We also explored the DEGs filtered by SHAP > 0.001 in WGCNA M3 from prim/norm that were not overlapping with module M4 in met/prim (Table EV12B) to identify markers that were specific for primary PCa. We identified a total of 32 genes, 15 of which were not significantly up-regulated in met/prim (Fig 8B). Among these genes, we again found *HOXC6* as well as several other genes with known connections to PCa. *ARHGEF26* is known to be over-expressed in prostate cancer and to be linked to cancer progression (Wang *et al*, 2012). *LMX1B* can bind to *circGFRA1* and affect the tumorigenic properties and immune escape of PCa cells (Meng & Wu, 2022). Finally, *NKX2–3* expression was found to correlate with *PD-*

1, making it an interesting marker for the efficacy of anti-PD-1 therapy in PCa (Wu *et al*, 2021a). Once more we evaluated the diagnostic capability of each biomarker with a simple logistic regression model using 10-fold cross-validation to classify sample types in each comparison (Table EV12B). *SBK1* was the best performing biomarker of the pre-selected subset in prim/norm (Precision=0.93 ± 0.05, Recall=0.94 ± 0.03), followed by *HOXC6*, *NKX2-3* and *ARHGEF26*. Interestingly, the vast majority of the 32 candidate biomarkers achieved good predictive performance for both comparisons, making all of them interesting candidates for further analysis. However, it is worth noting that the performance of these markers was generally worse for the met/prim comparison, which was expected given that these genes were selected from the prim/norm comparison.

Lastly, we also examined the intersecting DEGs that were only present in the met/prim comparison in WGCNA M4 from met/prim that were not in M3 from prim/norm, again filtered by SHAP > 0.001 (Table EV12C). We identified a total of 24 genes, 19 of which were not significantly up-regulated in prim/norm (Fig 8C). We hypothesize that these genes, which were significantly up-regulated in met/prim, are good candidates for specific biomarkers able to differentiate between normal/primary and metastatic PCa. Of note, *ATAD5*, *HROB* and *GEN1* are involved in DNA replication, repair and maintenance of genome stability (Giovannini *et al*, 2020; Ho & West, 2022; Liu & Wang, 2022). Although their specific functions for PCa have not been elucidated, they might have important roles for PCa therapy including chemotherapy, radiation or PARP inhibitor treatment. We again estimated the capabilities of these potential biomarkers by fitting a simple logistic regression model using 10-fold cross-validation to label sample types in each comparison (Table EV12C). Results showed remarkable validation values. *CCZ1B* was the best predictor in met/prim (Precision=0.97 ± 0.02, Recall=0.97 ± 0.02), followed by *CCZ1* and *FAM72A*.

As an additional validation step, we performed differential expression analysis on the Prostate Cancer Transcriptome Atlas (PCTA) dataset (Li *et al*, 2021) after removing the samples from our study (i.e. those from TCGA-PRAD and SU2C-PCF) and assessed all of our candidate biomarkers for both comparisons (Table EV13, A-C and Fig EV2, A-C). We examined the differences in expression of our biomarkers between both datasets and found a considerable rate of agreement. More concretely, we found that for both datasets, 74% of the biomarkers were similarly expressed in prim/norm, whereas 79% were similarly expressed in met/prim (Table EV13, D and E).

Lastly, to determine the clinical relevance of the biomarkers discovered from the integrated analyses, we performed recursive partitioning-based survival analysis (Hothorn *et al*, 2006) of the 58 individual biomarkers using the web-based camcAPP tool (Dunning *et al*, 2017). Using the MSKCC dataset (Taylor *et al.*, 2010), which includes primary and metastatic PCa samples, we found that 22 of the markers were significantly associated with shorter time to biochemical recurrence and thus worse prognosis of the patients (Fig EV3, A-C).

To summarize, we provide strong evidence for three different groups of candidate biomarkers. The first group comprises general tumor progression biomarkers that had already been suggested by previous studies. The second group of biomarkers are specific for early stage PCa and could be usable for diagnostic purposes. Lastly, the third group includes specific biomarkers to differentiate between primary and metastatic PCa tumors, which might be relevant as prognostic biomarkers.

Discussion

Tumor classification based on gene expression data has revealed distinct PCa tumor subtypes and uncovered expression patterns that were associated with clinical outcomes for localized and metastatic disease (Zhao *et al*, 2017; Zhao *et al*, 2020). Thus, DEGs can serve as biomarkers for tumor diagnosis, prognosis, response prediction and/or as direct targets for drug development.

Our method extends transcriptomic analyses by using ML and WGCNA together with different pathway analysis tools to identify DEG, which allow for robust group prediction (norm/prim/met) and functional biological interpretation of target genes. By integrating datasets from the TCGA and dbGAP databases, we focused the analyses on genes that were up-regulated in primary localized PCa compared to benign tissue, or overexpressed in mCRPC compared to primary PCa. This unbiased setting provided the opportunity to include disease stages from localized to lethal PCa, thereby highlighting potential general biomarkers for PCa diagnosis (e.g gene up-regulated in prim) or disease progression (genes up-regulated in met) irrespective of their heterogeneity or molecular subgroup.

Based on recent studies, cell surface proteins, which connect intracellular and extracellular signaling networks and largely determine a cell's capacity to communicate and interact with its environment, are highly suitable biomarker candidates and targets for pharmacological intervention. However, information on the cell surface proteins of PCa is largely missing (Bausch-Fluck *et al.*, 2015). In our analysis, we identified cell-surface proteins, including receptors, transporters, membrane-related proteins, cell-adhesion proteins, and extracellular enzymes as a rich source for biomarker development and as potential therapeutic targets. Several of the top deregulated genes in metastatic tumors were associated with cell-cell or cell-matrix interactions, which are highly relevant for metastatic processes. These included *SERPINA1, SERPINE2* and *SERPINI1*, three serine proteinase inhibitors that are part of the

large family of proteins, which have been implicated in diverse cancer driving pathways including angiogenesis and metastasis (Eatemadi *et al*, 2017; El-Akawi *et al*, 2008; Hasumi *et al*, 2005), although their specific role in PCa has not been fully elucidated. Previous reports rather pointed towards a role of *SERPINE2* for decreased tumor cell invasion and increased apoptosis in PCa (Xu *et al*, 2010). Thus, it will be important to study the roles of selected SERPIN proteins for PCa progression in future studies.

Of note, we found several genes implicated in neuronal processes among the top deregulated genes, including *NTN3* and *NTSR1*. NTN3 is part of the nexin family of proteins, and has been shown to be regulated by the AR-EZH2 axis in PCa (Hao *et al*, 2020). NTSRs and their ligand neurotensin (NTS) critically affect the progression of PCa. Immunohistochemical analysis revealed high or moderate expression of NTSR1 in 91,8% of PCa tissue including PSMA negative samples (He *et al*, 2019) and in 9,1% in primary PCa and 33% of lymph node metastases (Morgat *et al*, 2019), in two recent studies. In addition, molecular imaging tracers are available and might represent suitable tracers for PET imaging and radionuclide therapy, especially for PSMA negative tumors or tumors with neuroendocrine differentiation (Wu *et al.*, 2021a). Together, these data imply important functions for several neurodevelopmental genes and pathways for PCa progression also in non-neuroendocrine differentiated tumors.

Using WGCNA, we were able to identify key co-expression modules and hub genes, which play important roles in several cancer related pathways. Among several relevant genes, *TPX2* was identified as a hub gene for both pairwise comparisons, highlighting its relevance for PCa development and progression. TPX2 is an activator of AURKA signaling and is important for microtubule nucleation and stability during mitosis (Bayliss *et al*, 2003). AURKA is regulated by AR and represents a therapeutic target for PCa (Kivinummi *et al*, 2017). Furthermore, TPX2 is involved in homology-directed repair of DNA double strand breaks during replication stress and is a direct interaction partner of PARP1 (Mosler *et al*, 2022). This might also have

important implications for PARP inhibitor therapy for advanced PCa. A recent WGCNA analysis of the TCGA PRAD dataset also identified TPX2 as a hub gene and confirmed its prognostic value (Wang *et al*, 2022a).

Our integrative approach, combining ML and WGCNA data revealed a rather uniform trajectory towards disease progression from normal prostate, to primary, and metastatic PCa. The trajectory was characterized by a gradual up-regulation of genes such as *EZH2* and *TROAP* highlighting the central role of these two genes for PCa progression. Interestingly, *TROAP* was shown to be downstream of *EZH2* and is important for PCa progression via the TWIST/c-MYC pathway (Jin *et al*, 2021). *EZH2* is currently assessed as a therapeutic target for mCRPC in clinical studies, using the EZH2 inhibitor tazemetostat in combination with androgen signaling inhibitors (NCT04179864) or the PARPi talazoparib (NCT04846478).

In addition, the integration of ML and WGCNA yielded a selection of candidate biomarkers for localized PCa, including *SBK1*, *ARHGEF26*, *DLX2* and *NKX2*-3, or *HROB*, *GEN1* and *E2F2* for mCRPC, whose clear role for PCa has only been partially elucidated. This, might warrant further functional studies, especially for genes up-regulated in metastatic tumors including *HROB* or *GEN1*, which are implicated in DNA damage repair pathways or *SBK1*, *DLX2* and *E2F2* which are associated with worse prognosis of patients.

This work emphasizes a gradual alteration on the transcriptome level from primary to metastatic PCa, and highlights several relevant disease pathways and candidate biomarker genes. Future work that integrates multi omics data such as mutation data, epigenomics and RNA-seq data in large cohorts will enable further insights into the identified biomarkers. In parallel, complementary experimental approaches to study the functional relevance of the identified candidate markers will provide further information on potential targets for therapeutic intervention for PCa.

Methods

RNAseq data preparation

Data used in this study were extracted from two data cohorts: the cancer genome atlas (TCGA) (Weinstein *et al.*, 2013) and the database of Genotypes and Phenotypes (dbGap) (Mailman *et al.*, 2007). The first set of RNA-seq data belongs to the TCGA Prostate Adenocarcinoma project (TCGA-PRAD, https://portal.gdc.cancer.gov/repository, Last accessed in 09/2022). We downloaded the TCGA raw STAR read counts of prostate cancer tumor samples from the GDC data portal using the R/Bioconductor package TCGAbiolinks (Colaprico *et al*, 2016). The PRAD dataset includes the entire collection of 500 primary solid tumors and 52 solid tissue normal samples.

For the second set of RNAseq data, we obtained authorized access to the metastatic castration resistant prostate cancer (mCRPC) tissue samples of the SU2C-PCF dataset (Abida *et al.*, 2019; Robinson *et al.*, 2015) via dbGAP, with accession number phs000915.v2.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000915.v2.p2). SU2C-PCF comprises two studies, where a total of 429 patients were enrolled at an international consortium, all of whom underwent biopsy for the collection of mCRPC tissue. RNA sequencing was performed on a subset of 224 tumors.

We obtained the .sra files of all samples, which we then processed to get the final gene counts. We also downloaded the sample annotation information coming from different sources, such as basic annotation from dbGAP, clinical annotation from cBioPortal (Cerami *et al*, 2012) (https://cbioportal-datahub.s3.amazonaws.com/prad_su2c_2019.tar.gz) and SRA metadata from NCBI (https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/Traces/study/?acc=phs000915&o=acc_s%3Aa).

Data exploratory analysis clearly showed that metastatic samples were perfectly clustered into two groups after applying K-means to the PCa results. An in-depth analysis of the available

data annotation showed that samples separated based on the annotation for "sample use" as either "RNA_Seq" or "RNA_Seq_Exome". For this reason, and due to its comparability to the TCGA samples, we only included samples labeled as "RNA_Seq" in our study. Additionally, we removed all sample replicates per group, keeping only those samples that were sequenced the latest (samples sequenced in 2018 and 2019 at the University of Michigan, Broad Institute). Samples with neuroendocrine features (n=14) were also excluded from our study. Therefore, of the 224 RNA-seq samples available on dbGAP (Abida *et al.*, 2019; Robinson *et al.*, 2015), we finally included 57 samples with unique patient IDs to avoid technical bias (Metastatic_BB in Fig EV4). These samples were obtained from biopsies of lymph node (n=28; 50.90%), bone (n=13; 23.63%), liver (n=8; 13.54%), and other soft tissues (n=6; 7.27%).

Pre-processing

While the STAR raw counts were directly available for normal and primary tumor samples from TCGA, mCRPC samples from dbGAP were first processed from sra files. We extracted the fastq paired-end reads and used cutadapt (Martin, 2011) to remove unwanted sequences (e.g., adapters, poly-A tails, etc.) and low-quality reads. Then, the STAR aligner tool (Dobin *et al*, 2012) was used to map the reads to the GRCh38.p13 (release 105) human reference genome, obtained from ENSEMBL (https://www.ensembl.org/Homo_sapiens/Info/Index), as well as to obtain gene counts.

Only intersecting transcripts of TCGA and dbGAP count files were kept, resulting in 60,528 common transcripts. We also performed a naive pre-filtering step to remove low count genes by only keeping those transcripts with an average count across all samples bigger than 1. This yielded a total of 33,010 transcripts for further analysis.

Transcriptome analysis

We used the R/Bioconductor package DESeq2 (Love *et al*, 2014) to perform differential gene expression (DE) analysis of the un-stranded STAR transcript counts. Differentially expressed

genes were derived from likelihood ratio tests in the respective experimental subgroups: metastatic versus primary (met/prim) and primary versus normal (prim/norm). Genes with an adjusted P-value (Padj) < 0.05 and absolute log2 fold change (LFC) > 1 were considered significantly differentially expressed, representing a conservative and stringent approach. Finally, read counts were normalized by the DESeq2 normalization method of variance stabilizing transformation (VST), to be used in downstream analyses.

Functional gene set enrichment analysis for differentially expressed genes (DEGs)

To gain insight into the biological and/or clinical relevance of the significantly differentially expressed genes in each experimental subgroup, enrichment analysis was performed using several tools including GO enrichment analysis, cell surface protein identification, biomarker prediction and pathway analysis.

Gene ontology (GO) term enrichment analysis

The expression patterns of DEGs in each pair-wise comparison (prim/norm and met/prim) were explored by hierarchical clustering (hclust method) using the R package simplifyEnrichment (Gu & Hübschmann, 2022). The classification of samples was obtained by consensus partitioning with the R package cola (Gu *et al*, 2020). The signature genes were additionally clustered into two groups by k-means clustering (cluster 1, cluster 2). Furthermore, GO enrichment analysis was applied to the two groups of genes separately with the R package clusterProfiler (Wu *et al*, 2021b). To visualize whether the GO terms were significantly enriched in each subgroup, the binary cut was applied directly to the union of the two significant GO term lists, and a heatmap of Padj was put on the left side of the GO similarity heatmap (Padj < 0.05). This strategy keeps all significant GO terms without removing any. The summaries of the biological functions in clusters are visualized as word clouds and are attached

to the GO similarity heatmap, which gives a direct illustration of the common biological functions involved in each cluster.

Pathway analysis

With the help of the R package clusterProfiler (Wu *et al.*, 2021b), we mined enriched pathways for Reactome databases (Gillespie *et al.*, 2021). Additionally, we employed ingenuity pathway analysis (QIAGEN IPA (QIAGEN Inc., https://digitalinsights.qiagen.com/IPA) (Krämer *et al.*, 2014) to interpret the DEGs in the context of biological processes, pathways and networks using the "core analysis" function. The "biomarker analysis" function included in IPA® was used to identify targets, which were already reported as biomarkers in previous studies.

Cell surface protein (CSP) analysis

We performed CSP analysis using the cell surface protein atlas (CSPA), a public resource containing experimental evidence for cell-surface proteins identified in 41 human cell types (Bausch-Fluck *et al.*, 2015). The basis and reference for the presented human surface proteome analysis was the human proteome in UniProtKB/Swiss-Prot (Version 2015_01) (2015). We matched the differentially expressed genes to the human peptides from the CSPA, generating a list of cell surface proteins enriched in our data.

To gain insight into the functional aspects of up-regulated cell surface receptors (n=201) in the met/prim comparison, the respective genes were used as input for integrative pathway analysis (Paczkowska *et al.*, 2020). The following pathway tools for GO term analysis, Human phenotype ontology (HP), WikiPathways (WP) and molecular pathways of the Reactome database obtained as a GMT file from the g:Profiler web server (Raudvere *et al*, 2019).

Active pathways were used as input for the EnrichmentMap app (Merico *et al.*, 2010) of Cytoscape (Cline *et al.*, 2007) for network visualization of similar pathways. Then, we used the AutoAnnotate app (Kucera *et al*, 2016) to summarize the networks and clustering pathways. Auto-annotated enrichment maps of up-regulated CSP receptors were visualized with stringent

pathway similarity scores (Jaccard and overlap combined coefficient 0.6) and manually curated for the most representative groups of similar pathways and processes. Pathways that were redundant with larger groups of pathways were merged with the latter or discarded. Coloring of pathways was done according to their g:Profiler supporting resources.

DEGs filtering with machine learning

Explainable machine learning (ML) models can be a useful tool to rank DEGs based on their ability to predict different conditions. We trained ML models on our DEGs' normalized expression data and used them to produce "importance scores" for each DEG.

Input gene expression data was pre-processed in multiple steps. First, we filtered our DEGs of interest (Padj < 0.05, LFC > 1) by keeping only those transcripts that were uniquely mapped to ENTREZID/SYMBOL and were either protein-coding or non-coding RNA. We also removed transcripts that mapped to the same ENTREZID. For met/prim, we also removed genes whose value ranges did not overlap between classes. As a last step, we built the input matrix using the variance-stabilized transformed (VST) expression counts of our filtered DEGs and performed min-max feature scaling.

We trained our models following two main steps. First, we obtained the optimal random forest hyper-parameters by applying grid search cross-validation on our data, as implemented in the Python package scikit-learn (Fabian Pedregosa, 2011). Second, we performed bootstrap training for 10.000 iterations. In other words, we trained each model multiple times, each under a different random seed, which impacts the stochasticity of the training, from model fitting to train/test (80/20) data splits. For each bootstrap training iteration, we calculated the shapely additive explanations (SHAP) (Scott M. Lundberg, 2017) values for all samples, a method based on cooperative game theory used for interpretability of ML models. SHAP values are then averaged among all iterations, resulting in the final gene importance scores for a given contrast.

Finally, we used the DEGs SHAP values in each experimental subgroup to rank the genes as well as to keep only the most relevant ones (e.g., those with SHAP values above 0.001). The mean test balanced accuracy among all bootstrap iterations was 0.95 (±0.05std) for met/prim tumors and 0.86 (±0.07std) for prim/norm samples.

WGCNA

In order to obtain additional insights from our DEGs, we explored the co-expression relationships among them with the R/Bioconductor package WGCNA (Langfelder & Horvath, 2008).

We first loaded the VST expression counts and pre-processed our data the same way we did for the ML analysis, except that we did not remove non-overlapping genes and avoided any post-processing scaling.

We then detected modules of genes as well as the hub genes, based on node degree, within each module. For this, we first did a power estimate to weight the co-expression network to approximate a scale free topology, using a signed network type and biweight midcorrelation as correlation type. Finally, we used Pearson correlation to assess the interrelationships between each module and each condition (i.e. sample type) that was tested.

Final evaluation of potential biomarkers

We obtained sets of potential biomarkers by intersecting the up-regulated DEGs obtained from different methods (i.e. DE, ML and WGCNA) for both of our experimental subgroups.

We estimated the predictive capability of these biomarkers through a 10-fold cross-validation of a logistic regression model trained to classify between sample types in each contrast (i.e. prim/norm and met/prim). We trained a different model for each individual biomarker and obtained the mean and standard deviation of weighted performance metrics (i.e. F1 score, precision, recall, balanced accuracy and AUROC) for each gene.

Finally, as additional validation, we repeated the steps above for an external dataset, the Prostate Cancer Transcriptome Atlas (PCTA) (Li *et al.*, 2021). This dataset contains 174 normal, 714 primary and 316 metastatic samples from 11 datasets. We removed all samples from TCGA-PRAD and SU2C-PCF and kept only the common transcripts. This step yielded a dataset with 122 normal, 209 primary and 180 metastatic samples. We also ran our DE pipeline on this data to obtain the log2 fold change and P-value adjusted metrics. Then we examined whether our biomarkers behaved similarly for each experimental subgroup when comparing PCTA to our data. Lastly, we performed recursive partitioning-based survival analysis (Hothorn *et al.*, 2006) of the 58 individual biomarkers using the MSKCC dataset (Taylor *et al.*, 2010) via the web-based camcAPP tool (Dunning *et al.*, 2017).

Acknowledgements

The results shown in our research are in whole based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and and the database of Genotypes and Phenotypes (dbGap) (https://www.ncbi.nlm.nih.gov/gap/). The authors would like to thank for the availability of TCGA data and also getting authorized access to dbGAP database in this study. This work was supported by grants provided by the FWF (National Science Foundation in Austria, grant no. P 32771) and DOC Fellowship of the Austrian Academy of Sciences (25276).

Authors and Affiliations

1. Ludwig Boltzmann Institute Applied Diagnostics, Vienna, Austria

Raheleh Sheibani-Tezerji, Carlos Uziel Pérez Malla, Anna Malzer, Katarina Misura, Loan Tran & Gerda Egger

2. Department of Pathology, Medical University of Vienna, Vienna, Austria

Jessica Kalla, Anna Malzer, Katarina Misura, Loan Tran, Gabriel Wasinger & Gerda Egger

3. Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria

Gerda Egger

Authors' contributions

GE and RST conceived and designed the study and drafted the manuscript. CUPM and RST performed all computational analyses. GE, RST and CUPM were major contributors in writing the manuscript. AM collected the patient data from public databases and commented on the manuscript. KM, LT, JK and GW contributed in background and literature search and editing the manuscript. All authors read and approved the final manuscript.

Consent for publication

All authors agree to publication of the article.

Conflict of interest

The authors declare no competing interests.

Availability of data

Data used in this study were extracted from the TCGA Prostate Adenocarcinoma project (TCGA-PRAD, https://portal.gdc.cancer.gov/repository) and the SU2C-PCF dataset via dbGAP, with accession number phs000915.v2.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000915.v2.p2).

References

(2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204-212

Abdulghani J, Gu L, Dagvadorj A, Lutz J, Leiby B, Bonuccelli G, Lisanti MP, Zellweger T, Alanen K, Mirtti T *et al* (2008) Stat3 Promotes Metastatic Progression of Prostate Cancer. *The American Journal of Pathology* 172: 1717-1728

Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Christopher, Annala M, Aprikian A, Armenia J, Arora A *et al* (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163: 1011-1025

Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I, Cieslik M, Benelli M, Robinson D, Van Allen EM *et al* (2019) Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci U S A* 116: 11428-11436

Adamo P, Ladomery MR (2016) The oncogene ERG: a key factor in prostate cancer. *Oncogene* 35: 403-414

Adnan A, Basu S (2023) PSMA Receptor-Based PET-CT: The Basics and Current Status in Clinical and Research Applications. *Diagnostics* 13: 158

Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, Chatila WK, Chakravarty D, Han GC, Coleman I *et al* (2018) The long tail of oncogenic drivers in prostate cancer. *Nat Genet* 50: 645-651

Arthur R, Williams R, Garmo H, Holmberg L, Stattin P, Malmström H, Lambe M, Hammar N, Walldius G, Robinsson D *et al* (2018) Serum inflammatory markers in relation to prostate cancer severity and death in the Swedish AMORIS study. *Int J Cancer* 142: 2254-2262

Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, Moest H, Omasits U, Gundry RL, Yoon C *et al* (2015) A Mass Spectrometric-Derived Cell Surface Protein Atlas. *PLOS ONE* 10: e0121314

Bayliss R, Sardon T, Vernos I, Conti E (2003) Structural Basis of Aurora-A Activation by TPX2 at the Mitotic Spindle. *Molecular Cell* 12: 851-862

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al* (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2: 401-404

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B *et al* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2: 2366-2382

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I *et al* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 44: e71

Cui Z, Sun G, Bhandari R, Lu J, Zhang M, Bhandari R, Sun F, Liu Z, Zhao S (2021) Comprehensive Analysis of Glycolysis-Related Genes for Prognosis, Immune Features, and Candidate Drug Development in Colon Cancer. *Front Cell Dev Biol* 9: 684322

Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, Mesher D, Speights V, Stankiewicz E, Foster CS *et al* (2011) Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *The Lancet Oncology* 12: 245-255

Davies A, Nouruzi S, Ganguli D, Namekawa T, Thaper D, Linder S, Karaoğlanoğlu F, Omur ME, Kim S, Kobelev M *et al* (2021) An androgen receptor switch underlies lineage infidelity in treatment-resistant prostate cancer. *Nat Cell Biol* 23: 1023-1034

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21

Dunning MJ, Vowler SL, Lalonde E, Ross-Adams H, Boutros P, Mills IG, Lynch AG, Lamb AD (2017) Mining Human Prostate Cancer Datasets: The "camcAPP" Shiny App. *EBioMedicine* 17: 5-6

Eatemadi A, Aiyelabegan HT, Negahdari B, Mazlomi MA, Daraee H, Daraee N, Eatemadi R, Sadroddiny E (2017) Role of protease and protease inhibitors in cancer pathogenesis and treatment. *Biomedicine & Pharmacotherapy* 86: 221-231

Ecke TH, Schlechte HH, Schiemenz K, Sachs MD, Lenk SV, Rudolph BD, Loening SA (2010) TP53 gene mutations in prostate cancer progression. *Anticancer Res* 30: 1579-1586

El-Akawi ZJ, Al-Hindawi FK, Bashir NA (2008) Alpha-1 antitrypsin (alpha1-AT) plasma levels in lung, prostate and breast cancer patients. *Neuro Endocrinol Lett* 29: 482-484

Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, Aldubayan SH, Salari K, Kregel S, Richter C *et al* (2021) Biologically informed deep neural network for prostate cancer discovery. *Nature* 598: 348-352

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

Fabian Pedregosa GV, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825-2830

Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A *et al* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* 38: D652-657

Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C *et al* (2021) The reactome pathway knowledgebase 2022. *Nucleic Acids Research* 50: D687-D692

Giovannini S, Weller MC, Hanzlíková H, Shiota T, Takeda S, Jiricny J (2020) ATAD5 deficiency alters DNA damage metabolism and sensitizes cells to PARP inhibition. *Nucleic Acids Res* 48: 4928-4939

Goldenberg SL, Nir G, Salcudean SE (2019) A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology* 16: 391-403

Green WJ, Ball G, Hulman G, Johnson C, Van Schalwyk G, Ratan HL, Soria D, Garibaldi JM, Parkinson R, Hulman J *et al* (2016) KI67 and DLX2 predict increased risk of metastasis formation in prostate cancer-a targeted molecular approach. *Br J Cancer* 115: 236-242

Gu Z, Hübschmann D (2022) Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results. *Genomics, Proteomics & Bioinformatics*

Gu Z, Schlesner M, Hübschmann D (2020) cola: an R/Bioconductor package for consensus partitioning through a general framework. *Nucleic Acids Research* 49: e15-e15

Hao W, Yu M, Lin J, Liu B, Xing H, Yang J, Sun D, Chen F, Jiang M, Tang C *et al* (2020) The pan-cancer landscape of netrin family reveals potential oncogenic biomarkers. *Scientific Reports* 10

Hasumi H, Ishiguro H, Nakamura M, Sugiura S, Osada Y, Miyoshi Y, Fujinami K, Yao M, Hamada K, Yamada-Okabe H *et al* (2005) Neuroserpin (PI-12) is upregulated in high-grade prostate cancer and is associated with survival. *International Journal of Cancer* 115: 911-916

He T, Wang M, Wang H, Tan H, Tang Y, Smith E, Wu Z, Liao W, Hu S, Li Z (2019) Evaluation of neurotensin receptor 1 as potential biomarker for prostate cancer theranostic use. *European Journal of Nuclear Medicine and Molecular Imaging* 46: 2199-2207

Ho HN, West SC (2022) Generation of double Holliday junction DNAs and their dissolution/resolution within a chromatin context. *Proc Natl Acad Sci U S A* 119: e2123420119

Hothorn T, Hornik K, Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15: 651-674

Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA (2010) Machine Learning: An Indispensable Tool in Bioinformatics. In: *Bioinformatics Methods in Clinical Research*, Matthiesen R. (ed.) pp. 25-48. Humana Press: Totowa, NJ

Jin L, Zhou Y, Chen G, Dai G, Fu K, Yang D, Zhu J (2021) EZH2-TROAP Pathway Promotes Prostate Cancer Progression Via TWIST Signals. *Frontiers in Oncology* 10

Kamoun A, Cancel-Tassin G, Fromont G, Elarouci N, Armenoult L, Ayadi M, Irani J, Leroy X, Villers A, Fournier G *et al* (2018) Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease. *Annals of Oncology* 29: 1814-1821

Kivinummi K, Urbanucci A, Leinonen K, Tammela TLJ, Annala M, Isaacs WB, Bova GS, Nykter M, Visakorpi T (2017) The expression of AURKA is androgen regulated in castration-resistant prostate cancer. *Scientific Reports* 7

Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, Chan JM, Li J, Cowan JE, Tsiatis AC *et al* (2014) A 17-gene Assay to Predict Prostate Cancer Aggressiveness in the Context of Gleason Grade Heterogeneity, Tumor Multifocality, and Biopsy Undersampling. *European Urology* 66: 550-560

Kotova ES, Savochkina YA, Doludin YV, Vasilyev AO, Prilepskay EA, Potoldykova NV, Babalyan KA, Kanygina AV, Morozov AO, Govorov AV *et al* (2020) Identification of Clinically Significant Prostate Cancer by Combined PCA3 and AMACR mRNA Detection in Urine Samples. *Res Rep Urol* 12: 403-413

Krämer A, Green J, Pollard J, Jr., Tugendreich S (2014) Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 30: 523-530

Kucera M, Isserlin R, Arkhangorodsky A, Bader GD (2016) AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Res* 5: 1717

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A *et al* (2006) Machine learning in bioinformatics. *Briefings in Bioinformatics* 7: 86-112

Li R, Zhu J, Zhong W-D, Jia Z, 2021. PCaDB - a comprehensive and interactive database for transcriptomes from prostate cancer population cohorts. Cold Spring Harbor Laboratory.

Li R, Zhu J, Zhong WD, Jia Z (2022) Comprehensive Evaluation of Machine Learning Models and Gene Expression Signatures for Prostate Cancer Prognosis Using Large Population Cohorts. *Cancer Res* 82: 1832-1843

Liu X, Wang C (2022) Pan-Cancer Analysis Identified Homologous Recombination Factor With OB-Fold (HROB) as a Potential Biomarker for Various Tumor Types. *Front Genet* 13: 904060

Loeb S, Bjurlin MA, Nicholson J, Tammela TL, Penson DF, Carter HB, Carroll P, Etzioni R (2014) Overdiagnosis and Overtreatment of Prostate Cancer. *European Urology* 65: 1046-1055

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15

Luo Z, Farnham PJ (2020) Genome-wide analysis of HOXC4 and HOXC6 regulated genes and binding sites in prostate cancer cells. *PLoS One* 15: e0228590

Ma X, Guo J, Liu K, Chen L, Liu D, Dong S, Xia J, Long Q, Yue Y, Zhao P *et al* (2020) Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell RNA sequencing. *Mol Cancer* 19: 147

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L *et al* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* 39: 1181-1186

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17: 10

Meng M, Wu YC (2022) LMX1B Activated Circular RNA GFRA1 Modulates the Tumorigenic Properties and Immune Escape of Prostate Cancer. *J Immunol Res* 2022: 7375879

Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE* 5: e13984

Morgat C, Chastel A, Molinie V, Schollhammer R, Macgrogan G, Vélasco V, Malavaud B, Fernandez P, Hindié E (2019) Neurotensin Receptor-1 Expression in Human Prostate Cancer: A Pilot Study on Primary Tumors and Lymph Node Metastases. *International Journal of Molecular Sciences* 20: 1721

Mosler T, Baymaz HI, Gräf JF, Mikicic I, Blattner G, Bartlett E, Ostermaier M, Piccinno R, Yang J, Voigt A *et al* (2022) PARP1 proximity proteomics reveals interaction partners at stressed replication forks. *Nucleic Acids Res* 50: 11600-11618

Nagano K, Yamashita T, Inoue M, Higashisaka K, Yoshioka Y, Abe Y, Mukai Y, Kamada H, Tsutsumi Y, Tsunoda S (2014) Eph receptor A10 has a potential as a target for a prostate cancer therapy. *Biochem Biophys Res Commun* 450: 545-549

Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, Mee MW, Boutros PC, Reimand J (2020) Integrative pathway enrichment analysis of multivariate omics data. *Nature Communications* 11

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* 47: W191-W198

Robinson D, Eliezer, Wu Y-M, Schultz N, Robert, Mosquera J-M, Montgomery B, Taplin M-E, Colin, Attard G *et al* (2015) Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* 161: 1215-1228

Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, Barberan CJ, Dannenfelser R, Dun C, Edrisi M *et al* (2022) Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications* 13

Scott M. Lundberg S-IL (2017) A Unified Approach to Interpreting Model Predictions. *Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)*

Shen MM, Abate-Shen C (2010) Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev* 24: 1967-2000

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* 71: 209-249

Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B *et al* (2010) Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* 18: 11-22

Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RGAB, Otte AP *et al* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 419: 624-629

Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ *et al* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* 8: 393-406

Wang FM, Xing NZ (2021) Systemic Coagulation Markers Especially Fibrinogen Are Closely Associated with the Aggressiveness of Prostate Cancer in Patients Who Underwent Transrectal Ultrasound-Guided Prostate Biopsy. *Dis Markers* 2021: 8899994

Wang H, Wu R, Yu L, Wu F, Li S, Zhao Y, Li H, Luo G, Wang J, Zhou J (2012) SGEF is overexpressed in prostate cancer and contributes to prostate cancer progression. *Oncol Rep* 28: 1468-1474

Wang J, Zhang Z, Li R, Mao F, Sun W, Chen J, Zhang H, Bartsch J-W, Shu K, Lei T (2018) ADAM12 induces EMT and promotes cell migration, invasion and proliferation in pituitary adenomas via EGFR/ERK signaling pathway. *Biomedicine & Pharmacotherapy* 97: 1066-1077

Wang L, Liu X, Liu Z, Wang Y, Fan M, Yin J, Zhang Y, Ma Y, Luo J, Li R *et al* (2022a) Network models of prostate cancer immune microenvironments identify ROMO1 as heterogeneity and prognostic marker. *Scientific Reports* 12

Wang W, Yuan D, Jiang K, Li R, Qu H, Jiang FN, Zhong WD, Sun F, Jia Z, Zhu J (2022b) Genome-wide CRISPR-Cas9 screening and identification of potential genes promoting prostate cancer growth and metastasis. *Curr Cancer Drug Targets*

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45: 1113-1120

Wu L, Quan W, Yue G, Luo Q, Peng D, Pan Y, Zhang G (2021a) Identification of a novel six autophagy-related genes signature for the prognostic and a miRNA-related autophagy predictor for anti-PD-1 therapy responses in prostate cancer. *BMC Cancer* 21

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L *et al* (2021b) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* 2: 100141

Xu D, McKee CM, Cao Y, Ding Y, Kessler BM, Muschel RJ (2010) Matrix Metalloproteinase-9 Regulates Tumor Cell Invasion through Cleavage of Protease Nexin-1. *Cancer Research* 70: 6988-6998

Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, Wu X, Stack EC, Loda M, Liu T *et al* (2012) EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* 338: 1465-1469

Ye J, Chu C, Chen M, Shi Z, Gan S, Qu F, Pan X, Yang Q, Tian Y, Wang L *et al* (2018) TROAP regulates prostate cancer progression via the WNT3/survivin signalling pathways. *Oncology Reports*

Zhang B, Zhang M, Li Q, Yang Y, Shang Z, Luo J (2021) TPX2 mediates prostate cancer epithelial-mesenchymal transition through CDK1 regulated phosphorylation of ERK/GSK3β/SNAIL pathway. *Biochem Biophys Res Commun* 546: 1-6

Zhao SG, Chang SL, Erho N, Yu M, Lehrer J, Alshalalfa M, Speers C, Cooperberg MR, Kim W, Ryan CJ *et al* (2017) Associations of Luminal and Basal Subtyping of Prostate Cancer With Prognosis and Response to Androgen Deprivation Therapy. *JAMA Oncology* 3: 1663

Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, Aggarwal R, Playdle D, Liao A, Alumkal JJ *et al* (2020) The DNA methylation landscape of advanced prostate cancer. *Nature Genetics* 52: 778-789

Zhao XM, Li YB, Sun P, Pu YD, Shan MJ, Zhang YM (2021) Bioinformatics analysis of key biomarkers for retinoblastoma. *J Int Med Res* 49: 3000605211022210

Figure legends

Figure 1. **Analysis setup.**

(A) Overview of sample origin and grouping for metastatic (n=57), primary tumor (n=500) and normal adjacent (n=52) samples from patients with PCa and mCRPC obtained from dbGAP and TCGA databases.

(B) Number and biopsy sites of mCRPC samples. (met, metastatic; norm, normal adjacent tissue; prim, primary).

Figure 2. **Global scaling of RNAseq data.**

(A) Principal component analysis (PCA) of RNA expression levels. The first two components shown explain the largest part of the variation in RNA expression (21%). Individual samples (circles) are colour-coded by sample type (met, metastatic; norm, normal adjacent; prim, primary).

(B) Heatmap of the unsupervised hierarchical clustering analysis on the variance-stabilized transformed (VST) expression of the top 500 most variable RNA genes across all samples within each sample type. Each row corresponds to a single gene, whereas each column corresponds to a single sample.

Figure 3. **Expression patterns and biological function of differentially expressed genes.**

(A) Venn diagrams illustrate overlaps and unique features identified in up- and down- regulated differentially expressed genes (DEGs) in each pair-wise comparison (prim/norm; met/prim). **(B)** and **(C)** Hierarchical clustering of signature genes differentially expressed between prim/norm and met/prim samples (B1 and C1). Rows correspond to genes and columns correspond to samples. Sample class was obtained by consensus partitioning methods to separate samples into subclasses. The two gene clusters (cluster1 and cluster2) were generated by applying k-means clustering on rows of the expression matrix. The z-score standardization was applied on matrix rows. **(B)** and **(C)** Significant GO terms enriched in either of the

corresponding gene clusters were clustered based on their similarities and visualized as a heatmap (B2 and C2). The columns on the left of the GO heatmap indicate the significance levels of GO terms in the individual clusters (Padj < 0.05). The summaries of the biological functions in clusters are visualized as word clouds and are attached to the GO similarity heatmap. In word clouds, enrichment of keywords is assessed by Fisher's exact test, and the significance is mapped to the font size of keywords.

Figure 4. **Integrative pathway enrichment analysis of cell surface protein receptors.**

(A) Top cell surface proteins (based on significant differential gene expression in met/prim) identified among the significantly up-regulated DEGs in met/prim samples, in which Padj are represented by colors and log2fold changes are represented by bar size.

(B) Enrichment map of Gene ontology (GO) terms (GOBP, GO biological process; GOMF, GO molecular function; GOCC, Cellular Component), Human phenotype ontology (HP), WikiPathways (WP) and Reactome (REAC) pathways identified from up-regulated cell surface proteins in met/prim. Number of enriched terms in each database is shown on the enrichment map labels.

Figure 5. Beeswarm plots of explainable machine learning binary classifiers. Top differentially expressed genes (model features) influencing the output of explainable machine learning binary classifiers, as ranked by SHAP value, are shown for prim/norm **(A)** and met/prim **(B)**. Each dot represents a predicted sample color-coded by VST gene expression values (from blue, low expression to red, high expression). The further out a dot is from the separating vertical line, the higher the absolute SHAP value, as indicated in the X axis, and hence the higher the impact on the model output. Negative and positive SHAP values steer the model towards a classification output of 0 or 1, respectively.

Figure 6. Weighted gene co-expression network analysis comparing primary versus normal samples.

(A) Number of genes in each module.

(B) Module hub genes based on node degree.

(C) Pearson correlation between modules and sample types. Darker colours show higher correlation.

(D) Visualization of the summarized co-expression networks for modules M2 and M3. Edges with a weighted co-expression coefficient < 0.1 were removed, as well as isolated nodes. Nodes are color-coded based on the module they belong to and their size is proportional to their SHAP value. Word clouds below the network graphs summarize the Reactome functional enrichment results for each module.

Figure 7. Weighted gene co-expression network analysis comparing metastatic versus primary tumors.

(A) Number of genes in each module.

(B) Module hub genes based on node degree.

(C) Pearson correlation between modules and sample types.

(D) Visualization of selected co-expression networks. Edges with a weighted co-expression coefficient < 0.1 were removed, as well as isolated nodes. Nodes are color-coded based on the module they belong to and their size is proportional to their SHAP value. Word clouds summarize the Reactome functional enrichment results for individual modules.

Figure 8. Candidate biomarker gene expression.

(A) Violin plots showing intersecting DEGs (SHAP > 0.001) between prim/norm and met/prim filtered gene sets (WGCNA M2 and M4, respectively). LFC, Padj, and F1-score (F1S) metrics between comparisons are also shown.

(B) Heatmap of DEGs (SHAP > 0.001) unique to the prim/norm filtered gene set (WGCNA M2), showing each DEG variance-stabilized transformed (VST) expression among the samples of each sample type.

(C) Heatmap of DEGs (SHAP > 0.001) unique to the met/prim filtered set (WGCNA M4). In (B) and (C), for each contrast, samples are annotated on the left side of the heatmap as being biologically significant (LFC > 1) in black, or not, in gray.

Expanded View Figure Legends

Figure EV1: Principal component analysis (PCA) of RNA expression levels. PCA graphs including PC1-4 show sample variance. Individual samples (circles) are color-coded by sample type.

Figure EV2. Validation of candidate biomarkers on Prostate Cancer Transcriptome Atlas (PCTA). Only up-regulated DEGs (Padj < 0.05, LFC > 1) were considered.

(**A)** Violin plots show intersecting DEGs (SHAP > 0.001) between prim/norm and met/prim filtered gene sets (WGCNA M2 and M4, respectively). LFC, Padj, and F1-score (F1S) metrics between comparisons are also shown.

(**B)** Heatmap of DEGs (SHAP > 0.001) unique to the prim/norm filtered gene set (WGCNA M2), showing each DEG variance-stabilized transformed (VST) expression among the samples of each sample type.

(**C)** Heatmap of DEGs (SHAP > 0.001) unique to the met/prim filtered set (WGCNA M4). In (B) and (C), for each contrast, samples are annotated on the left side of the heatmap as being biologically significant (LFC > 1) in black, or not, in gray.

Figure EV3. Kaplan Meyer survival curves for high and low expression levels of the 22 biomarkers were significantly associated with shorter time to biochemical recurrence.

(A) Survival curves of intersecting DEGs (SHAP > 0.001) between prim/norm and met/prim (WGCNA M2 and M4, respectively).

(B) Survival curves of unique DEGs (SHAP > 0.001) to the prim/norm filtered gene set (WGCNA M2).

(C) Survival curves of unique DEGs (SHAP > 0.001) to the met/prim filtered set (WGCNA M4).

Figure EV4. Principal component analysis (PCA) for all samples from TCGA and dbGAP databases. PCA graph of RNA expression levels for metastatic samples with unique patient Ids (Metastatic_AA: n=67, Metastatic_BB: n=57, Endocrine: n=14), primary tumors (n=500) and normal (n=52) samples from patients with PCa and mCRPC obtained from dbGAP and TCGA databases.

Figures

Figure 1

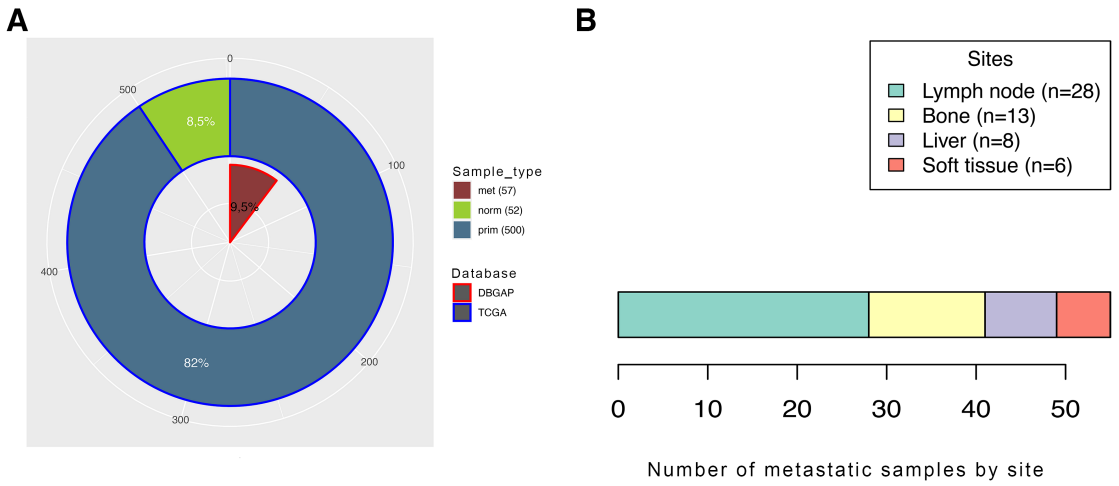
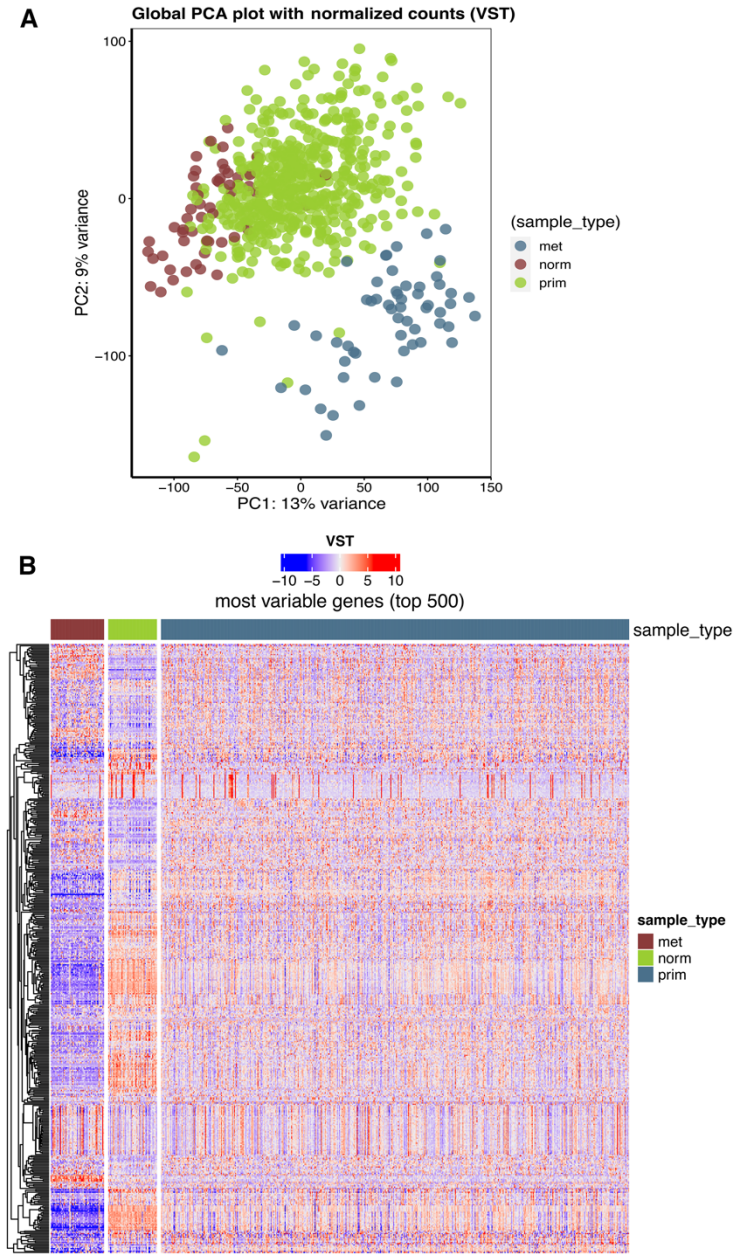


Figure 1. **Analysis setup. (A)** Overview of sample origin and grouping for metastatic (n=57), primary tumor (n=500) and normal adjacent (n=52) samples from patients with PCa and mCRPC obtained from dbGAP and TCGA databases. **(B)** Number and biopsy sites of mCRPC samples. (met, metastatic; norm, normal adjacent tissue; prim, primary).

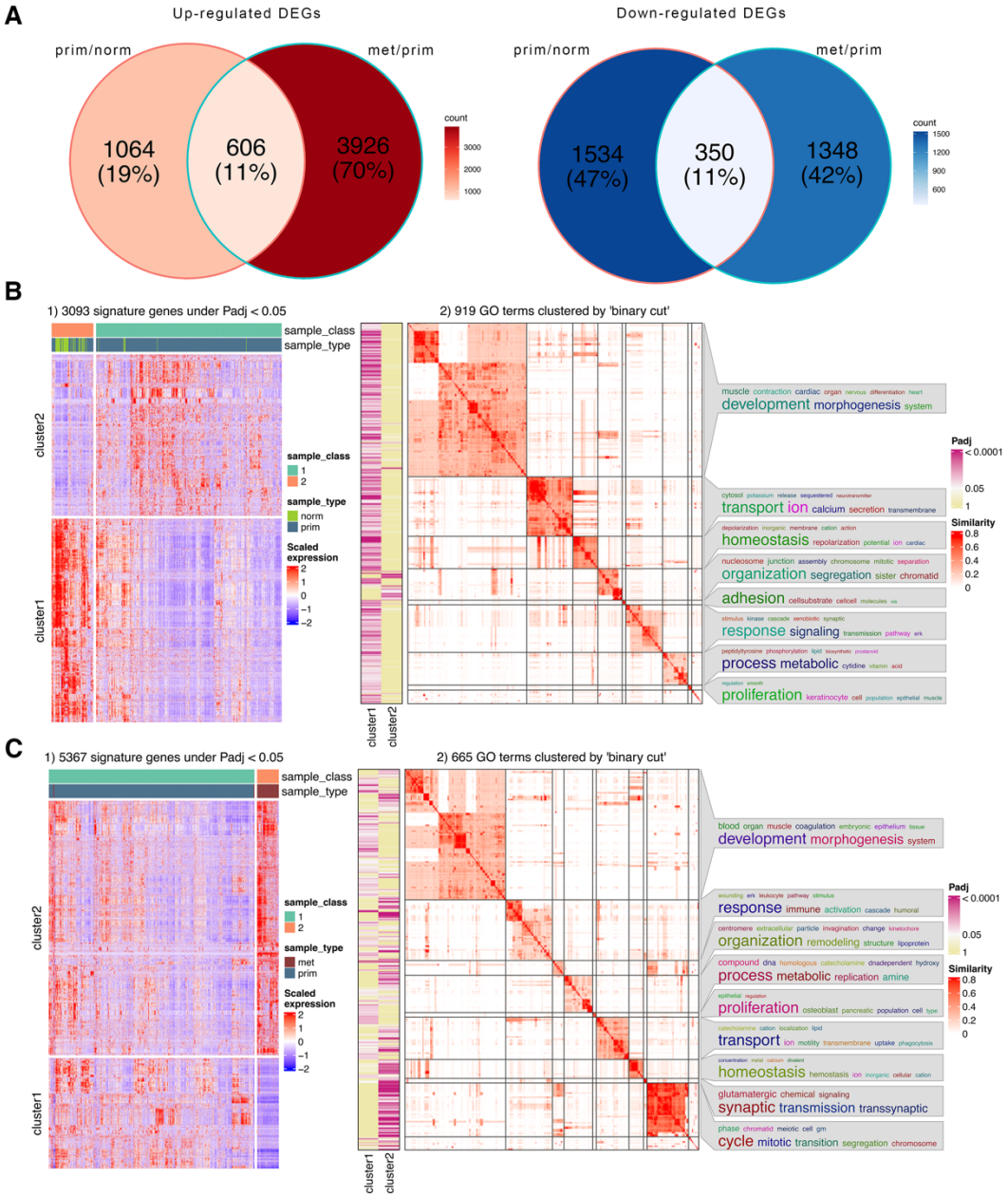
1 **Figure 2**



2

3 **Figure 2**. **Global scaling of RNAseq data. (A)** Principal component analysis (PCA) of RNA
4 expression levels. The first two components shown explain the largest part of the variation in
5 RNA expression (21%). Individual samples (circles) are colour-coded by sample type (met,
6 metastatic; norm, normal adjacent; prim, primary). **(B)** Heatmap of the unsupervised
7 hierarchical clustering analysis on the variance-stabilized transformed (VST) expression of the
8 top 500 most variable RNA genes across all samples within each sample type. Each row
9 corresponds to a single gene, whereas each column corresponds to a single sample.
10
11
12
13
14

1 **Figure 3**



2

3 **Figure 3**. **Expression patterns and biological function of differentially expressed genes.**
4 **(A)** Venn diagrams illustrate overlaps and unique features identified in up- and down- regulated
5 differentially expressed genes (DEGs) in each pair-wise comparison (prim/norm; met/prim).
6 **(B)** and **(C)** Hierarchical clustering of signature genes differentially expressed between
7 prim/norm and met/prim samples (B1 and C1). Rows correspond to genes and columns
8 correspond to samples. Sample class was obtained by consensus partitioning methods to
9 separate samples into subclasses. The two gene clusters (cluster1 and cluster2) were generated
10 by applying k-means clustering on rows of the expression matrix. The z-score standardization
11 was applied on matrix rows. **(B)** and **(C)** Significant GO terms enriched in either of the
12 corresponding gene clusters were clustered based on their similarities and visualized as a
13 heatmap (B2 and C2). The columns on the left of the GO heatmap indicate the significance
14 levels of GO terms in the individual clusters (Padj < 0.05). The summaries of the biological
15 functions in clusters are visualized as word clouds and are attached to the GO similarity

heatmap. In word clouds, enrichment of keywords is assessed by Fisher's exact test, and the significance is mapped to the font size of keywords.

Figure 4

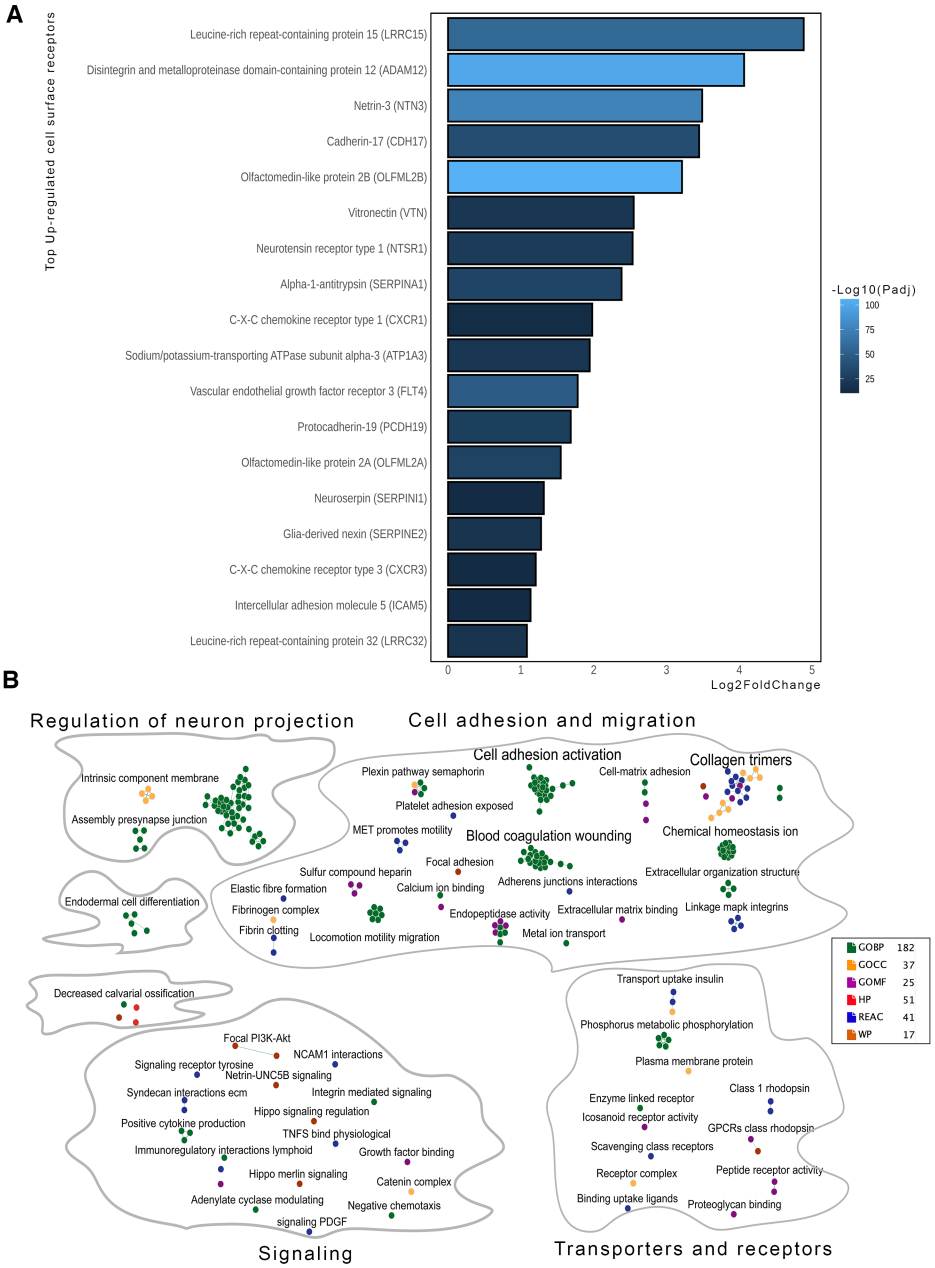


Figure 4. **Integrative pathway enrichment analysis of cell surface protein receptors. (A)** Top cell surface proteins (based on significant differential gene expression in met/prim) identified among the significantly up-regulated DEGs in met/prim samples, in which Padj are represented by colors and log2fold changes are represented by bar size. **(B)** Enrichment map of Gene ontology (GO) terms (GOBP, GO biological process; GOMF, GO molecular function; GOCC, Cellular Component), Human phenotype ontology (HP), WikiPathways (WP) and Reactome (REAC) pathways identified from up-regulated cell surface proteins in met/prim. Number of enriched terms in each database is shown on the enrichment map labels.

Figure 5

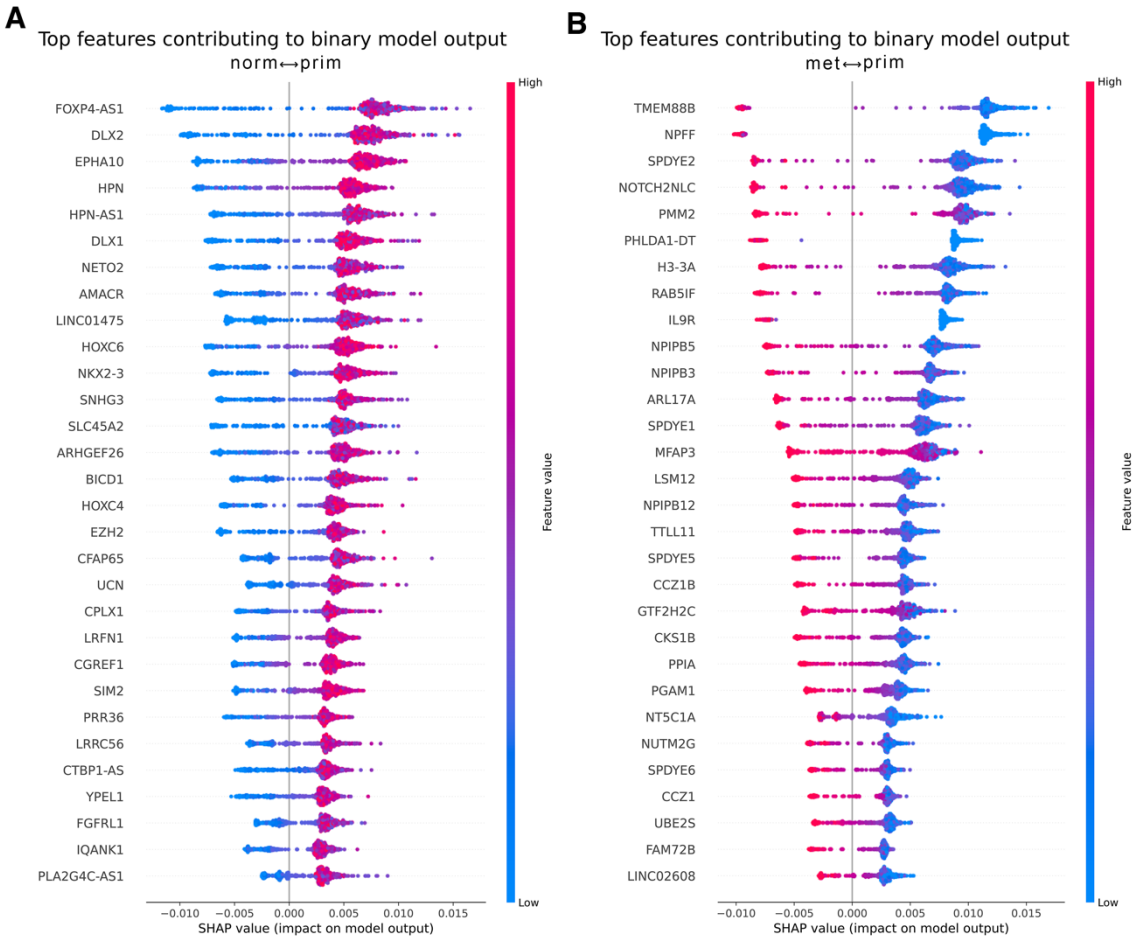
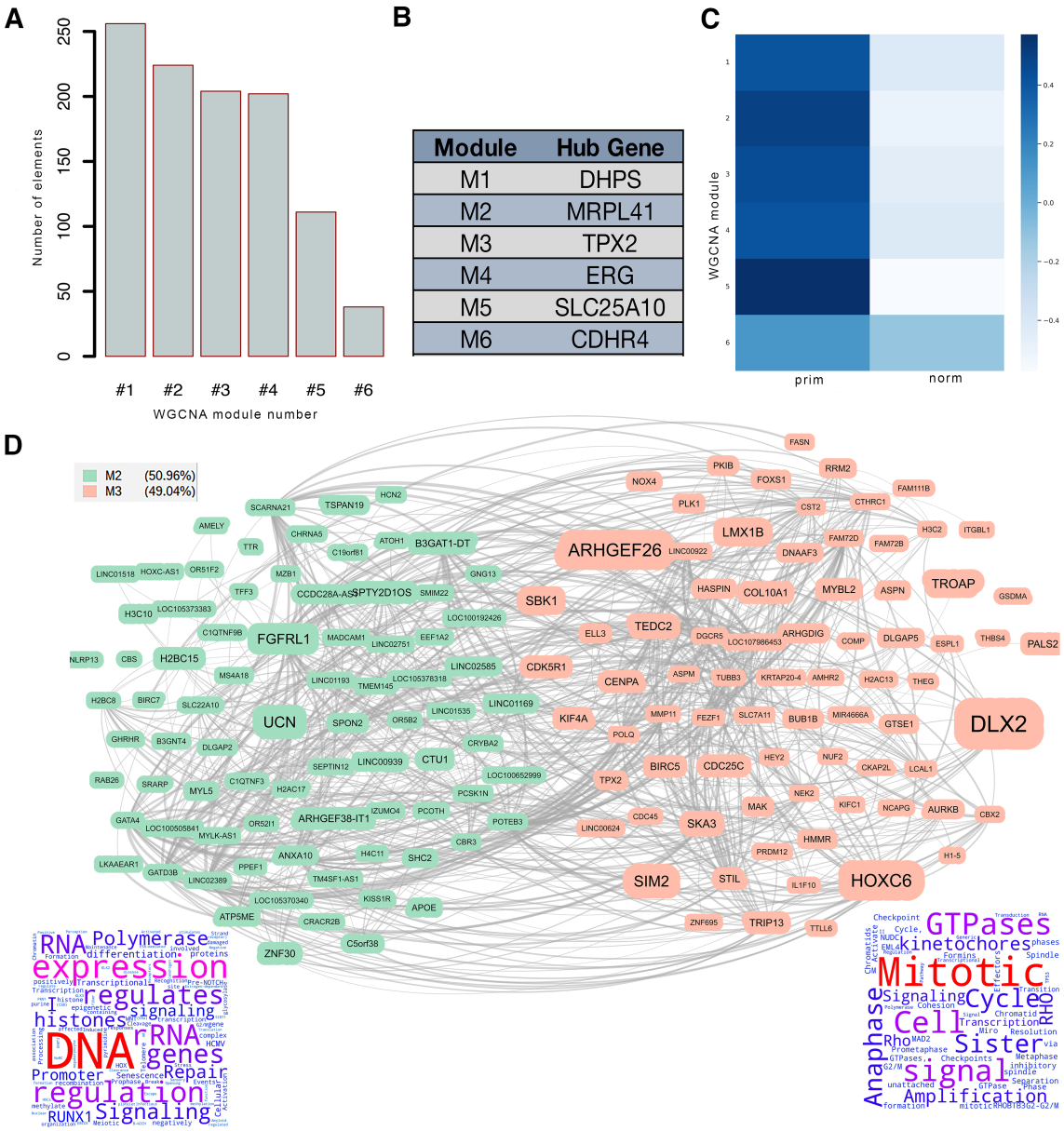


Figure 5. Beeswarm plots of explainable machine learning binary classifiers. Top differentially expressed genes (model features) influencing the output of explainable machine learning binary classifiers, as ranked by SHAP value, are shown for prim/norm **(A)** and met/prim **(B)**. Each dot represents a predicted sample color-coded by VST gene expression values (from blue, low expression to red, high expression). The further out a dot is from the separating vertical line, the higher the absolute SHAP value, as indicated in the X axis, and hence the higher the impact on the model output. Negative and positive SHAP values steer the model towards a classification output of 0 or 1, respectively.

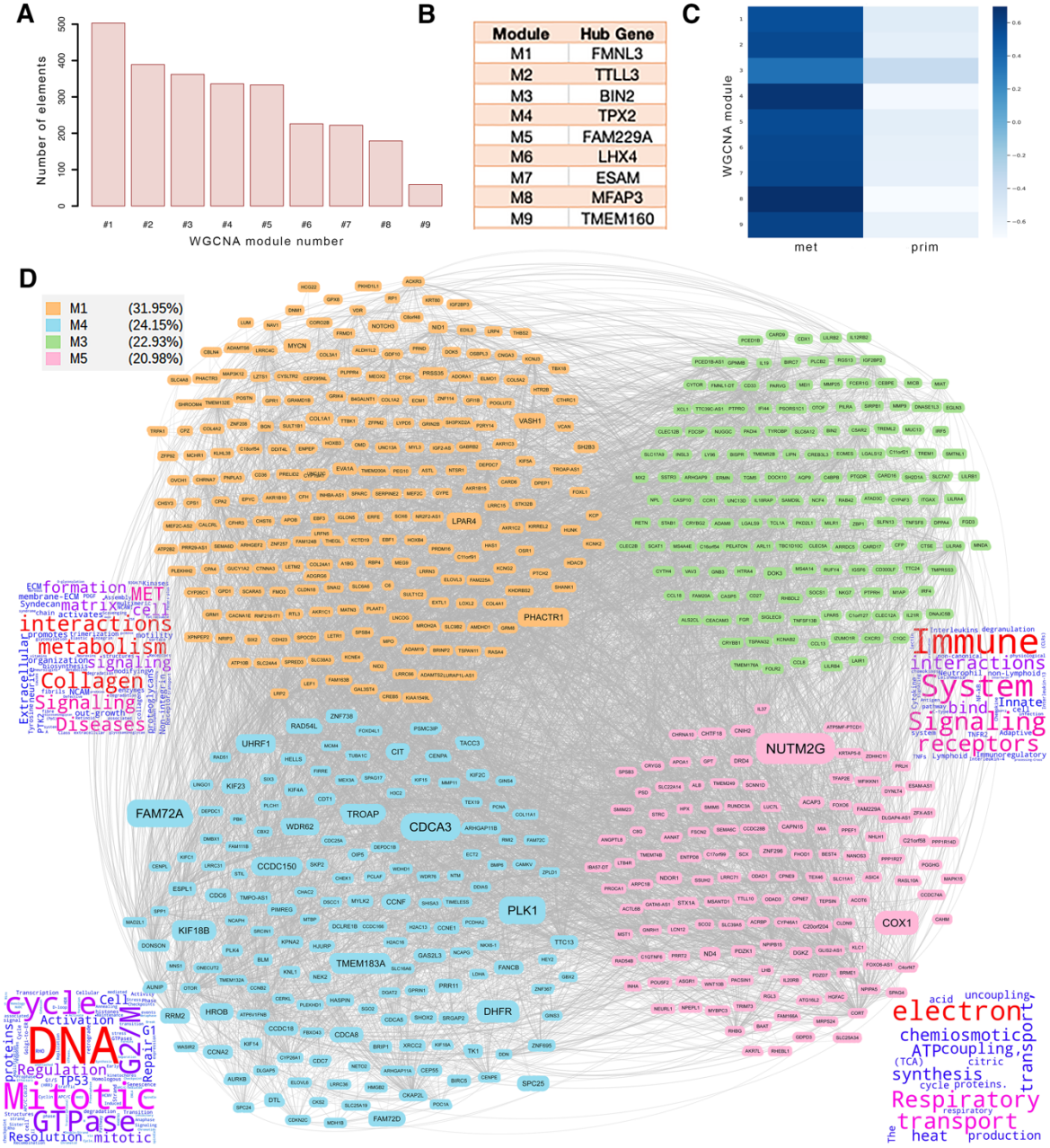
Figure 6



Module	Hub Gene
M1	DHPS
M2	MRPL41
M3	TPX2
M4	ERG
M5	SLC25A10
M6	CDHR4

Figure 6. Weighted gene co-expression network analysis comparing primary versus normal samples. (A) Number of genes in each module. **(B)** Module hub genes based on node degree. **(C)** Pearson correlation between modules and sample types. Darker colours show higher correlation. **(D)** Visualization of the summarized co-expression networks for modules M2 and M3. Edges with a weighted co-expression coefficient < 0.1 were removed, as well as isolated nodes. Nodes are color-coded based on the module they belong to and their size is proportional to their SHAP value. Word clouds below the network graphs summarize the Reactome functional enrichment results for each module.

Figure 7



Module	Hub Gene
M1	FMNL3
M2	TTLL3
M3	BIN2
M4	TPX2
M5	FAM229A
M6	LHX4
M7	ESAM
M8	MFAP3
M9	TMEM160

Figure 7. Weighted gene co-expression network analysis comparing metastatic versus primary tumors. (A) Number of genes in each module. **(B)** Module hub genes based on node degree. **(C)** Pearson correlation between modules and sample types. **(D)** Visualization of selected co-expression networks. Edges with a weighted co-expression coefficient < 0.1 were removed, as well as isolated nodes. Nodes are color-coded based on the module they belong to and their size is proportional to their SHAP value. Word clouds summarize the Reactome functional enrichment results for individual modules.

Figure 8

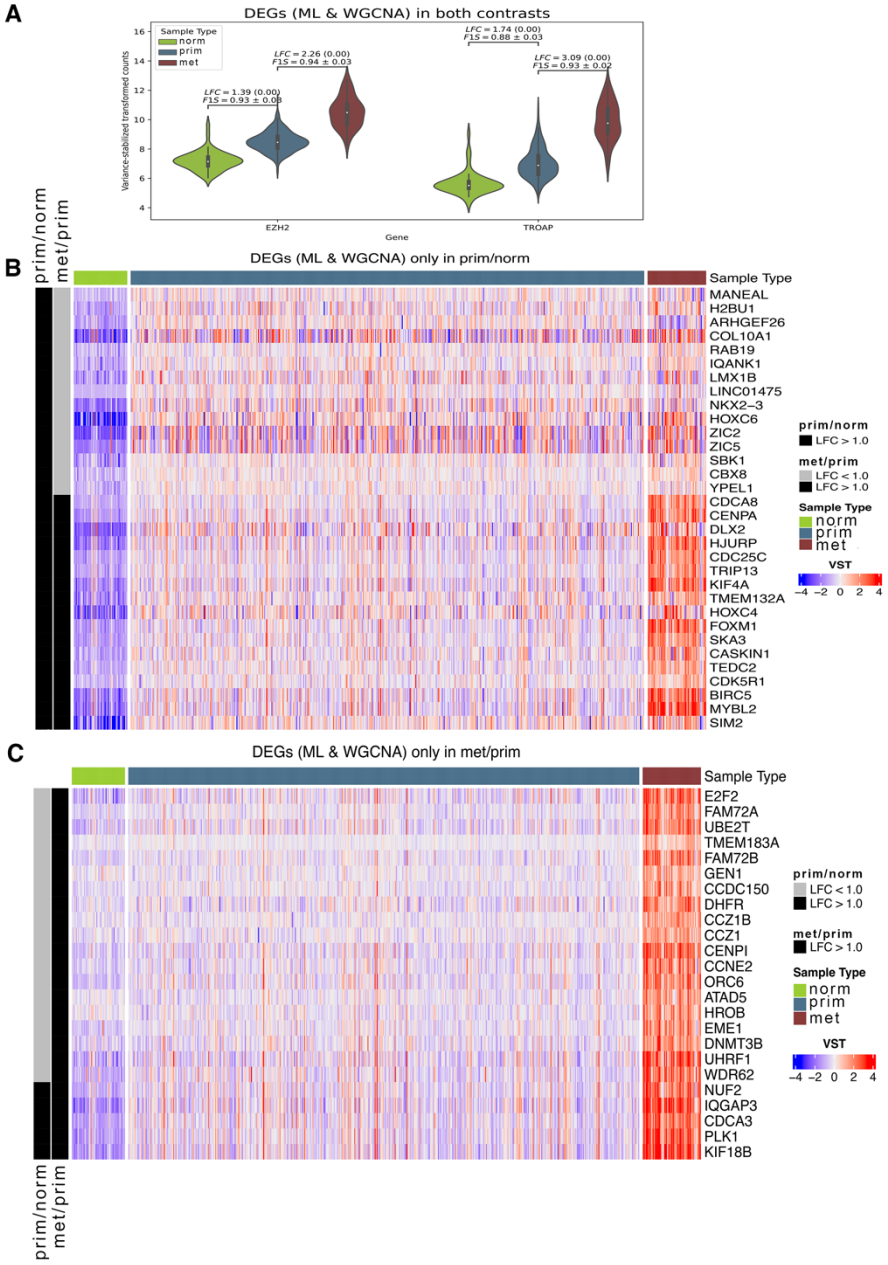
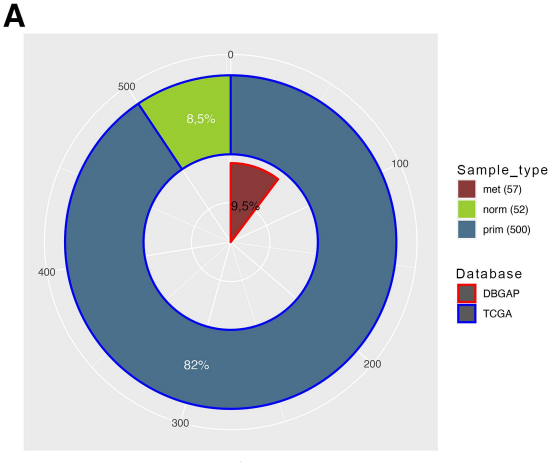
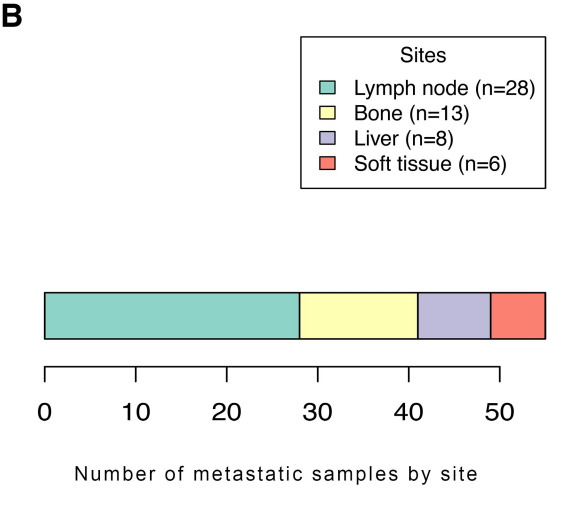
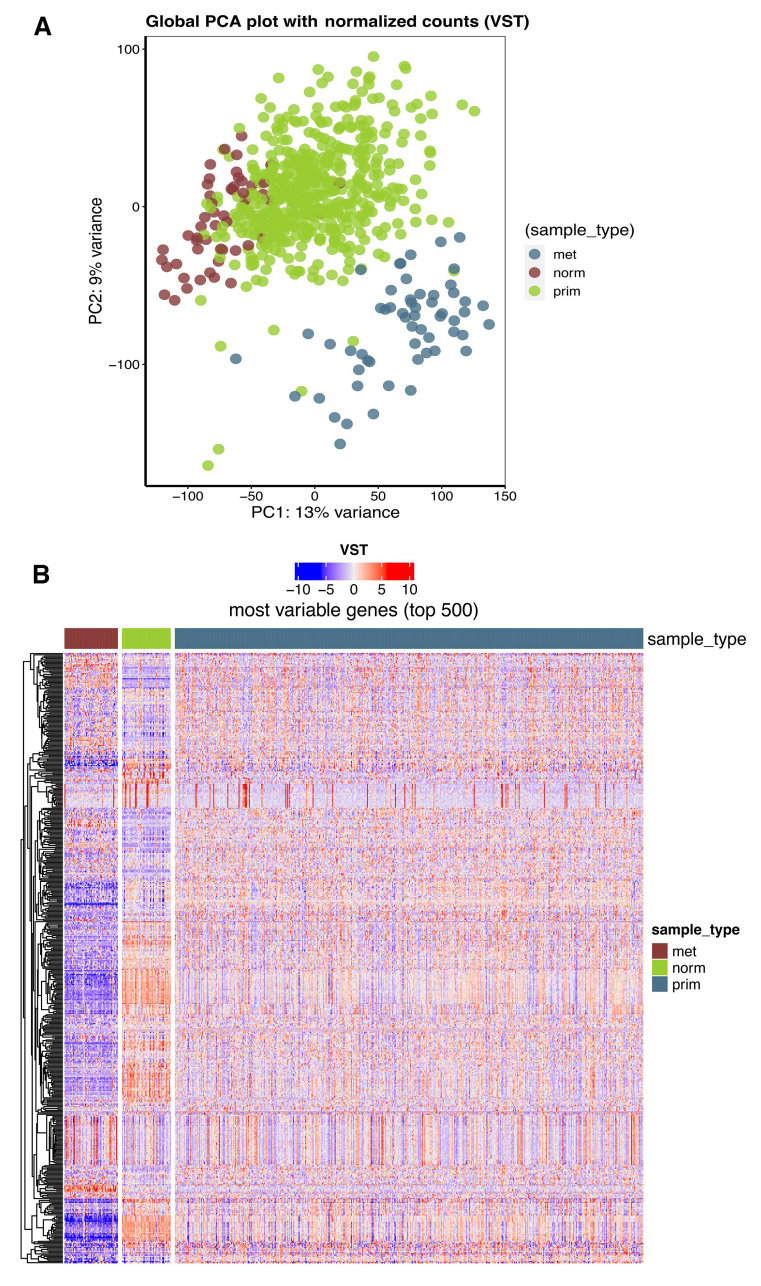
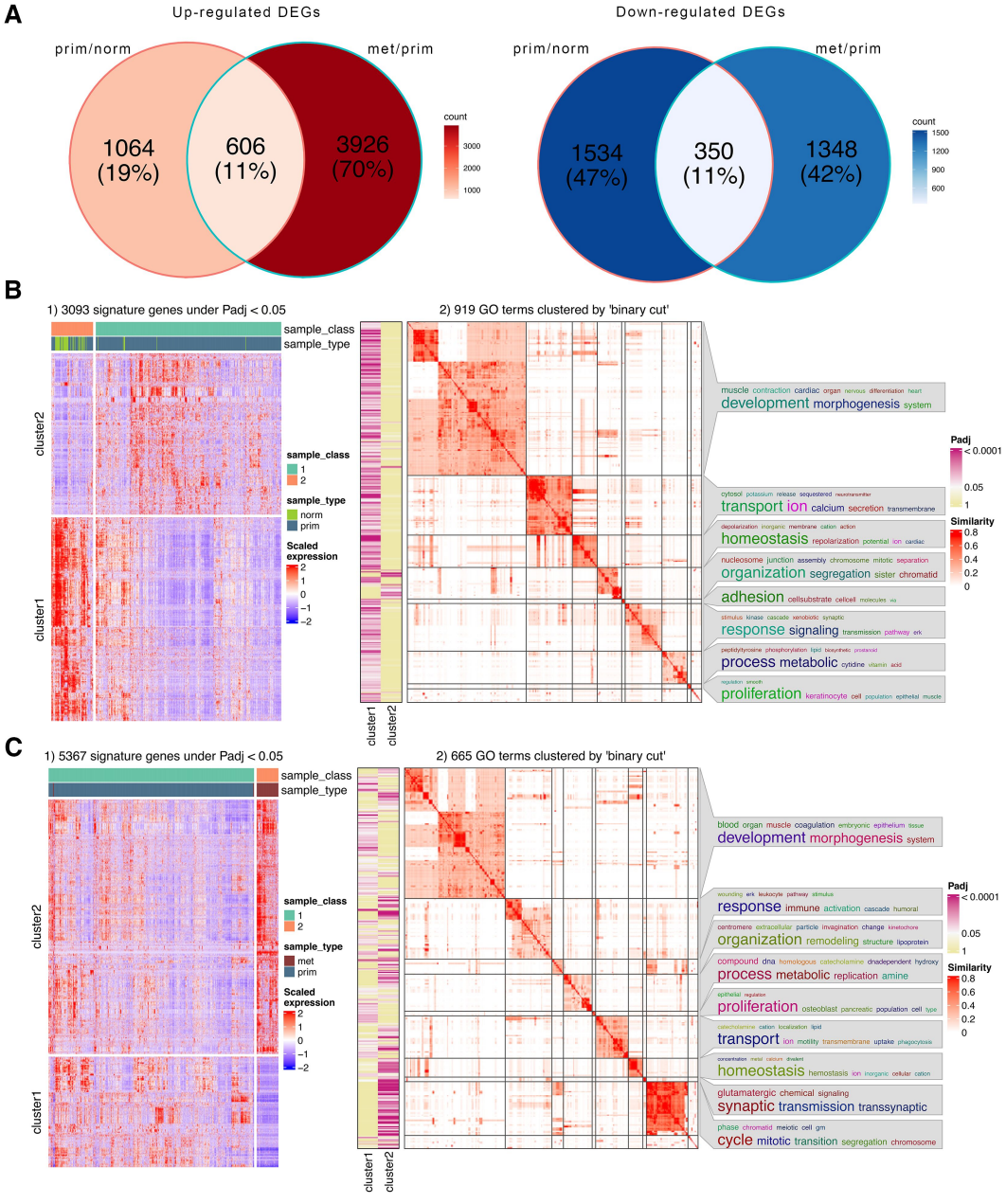


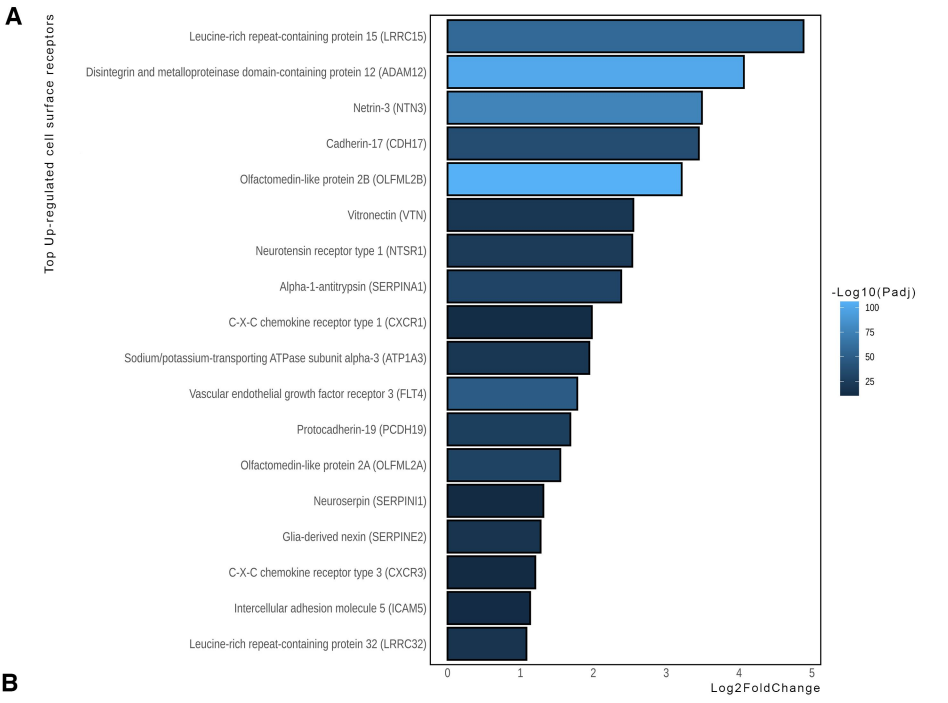
Figure 8. Candidate biomarker gene expression. (A) Violin plots showing intersecting DEGs (SHAP > 0.001) between prim/norm and met/prim filtered gene sets (WGCNA M2 and M4, respectively). LFC, Padj, and F1-score (F1S) metrics between comparisons are also shown. **(B)** Heatmap of DEGs (SHAP > 0.001) unique to the prim/norm filtered gene set (WGCNA M2), showing each DEG variance-stabilized transformed (VST) expression among the samples of each sample type. **(C)** Heatmap of DEGs (SHAP > 0.001) unique to the met/prim filtered set (WGCNA M4). In (B) and (C), for each contrast, samples are annotated on the left side of the heatmap as being biologically significant (LFC > 1) in black, or not, in gray.

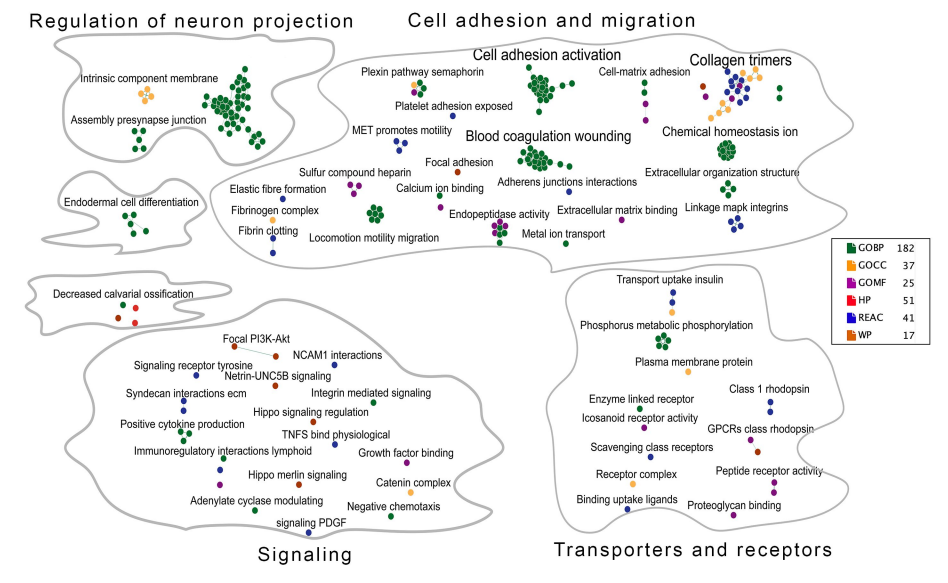

A
0
100
200
300
400
500
8,5%
9,5%
82%
Sample_type
met (57)
norm (52)
prim (500)
Database
DBGAP
TCGA

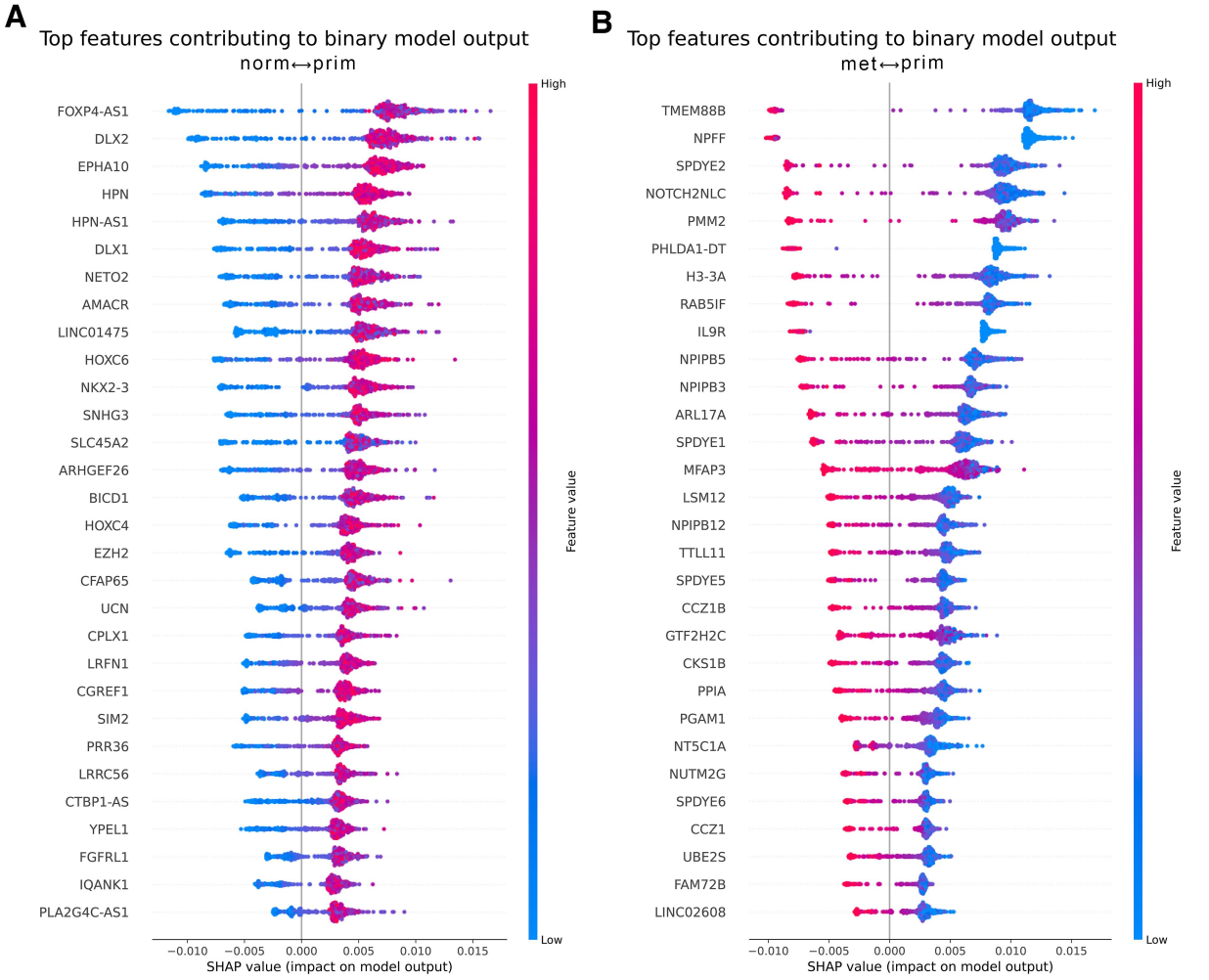

B
Sites
Lymph node (n=28)
Bone (n=13)
Liver (n=8)
Soft tissue (n=6)
0
10
20
30
40
50
Number of metastatic samples by site

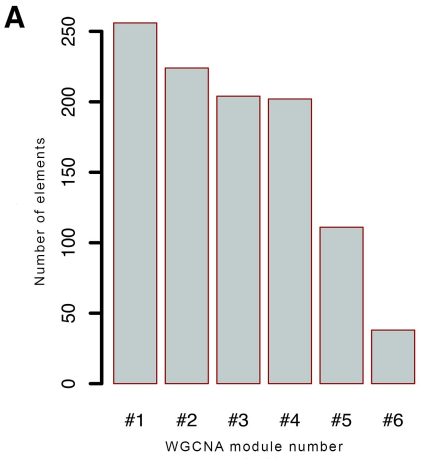

A
Global PCA plot with normalized counts (VST)
PC2: 9% variance
PC1: 13% variance
100
0
−100
−100
−50
0
50
100
150
(sample_type)
met
norm
prim
B
VST
−10 −5 0 5 10
most variable genes (top 500)
sample_type
sample_type
met
norm
prim


A
Up-regulated DEGs
prim/norm
met/prim
1064 (19%)
606 (11%)
3926 (70%)
count
3000
2000
1000
Down-regulated DEGs
prim/norm
met/prim
1534 (47%)
350 (11%)
1348 (42%)
count
1500
1200
900
600
B
1) 3093 signature genes under Padj < 0.05
2) 919 GO terms clustered by 'binary cut'
sample_class
sample_type
cluster2
cluster1
sample_class
1
2
sample_type
norm
prim
Scaled expression
2
1
0
−1
−2
cluster1
cluster2
muscle contraction cardiac organ nervous differentiation heart
development morphogenesis system
cytosol potassium release sequestered neurotransmitter
transport ion calcium secretion transmembrane
depolarization inorganic membrane cation action
homeostasis repolarization potential ion cardiac
nucleosome junction assembly chromosome mitotic separation
organization segregation sister chromatid
adhesion cellsubstrate cellcell molecules wa
stimulus kinase cascade xenobiotic synaptic
response signaling transmission pathway erk
peptidyltyrosine phosphorylation lipid biosynthetic prostanoid
process metabolic cytidine vitamin acid
regulation smooth
proliferation keratinocyte cell population epithelial muscle
Padj
< 0.0001
0.05
1
Similarity
0.8
0.6
0.4
0.2
0
C
1) 5367 signature genes under Padj < 0.05
2) 665 GO terms clustered by 'binary cut'
sample_class
sample_type
cluster2
cluster1
sample_class
1
2
sample_type
met
prim
Scaled expression
2
1
0
−1
−2
cluster1
cluster2
blood organ muscle coagulation embryonic epithelium tissue
development morphogenesis system
wounding erk leukocyte pathway stimulus
response immune activation cascade humoral
centromere extracellular particle invagination change kinetochore
organization remodeling structure lipoprotein
compound dna homologous catecholamine dnadependent hydroxy
process metabolic replication amine
epithelial regulation
proliferation osteoblast pancreatic population cell type
catecholamine cation localization lipid
transport ion motility transmembrane uptake phagocytosis
concentration metal calcium divalent
homeostasis hemostasis ion inorganic cellular cation
glutamatergic chemical signaling
synaptic transmission transsynaptic
phase chromatid meiotic cell gm
cycle mitotic transition segregation chromosome
Padj
< 0.0001
0.05
1
Similarity
0.8
0.6
0.4
0.2
0


A
Top Up-regulated cell surface receptors
Leucine-rich repeat-containing protein 15 (LRRC15)
Disintegrin and metalloproteinase domain-containing protein 12 (ADAM12)
Netrin-3 (NTN3)
Cadherin-17 (CDH17)
Olfactomedin-like protein 2B (OLFML2B)
Vitronectin (VTN)
Neurotensin receptor type 1 (NTSR1)
Alpha-1-antitrypsin (SERPINA1)
C-X-C chemokine receptor type 1 (CXCR1)
Sodium/potassium-transporting ATPase subunit alpha-3 (ATP1A3)
Vascular endothelial growth factor receptor 3 (FLT4)
Protocadherin-19 (PCDH19)
Olfactomedin-like protein 2A (OLFML2A)
Neuroserpin (SERPINI1)
Glia-derived nexin (SERPINE2)
C-X-C chemokine receptor type 3 (CXCR3)
Intercellular adhesion molecule 5 (ICAM5)
Leucine-rich repeat-containing protein 32 (LRRC32)
0
1
2
3
4
5
Log2FoldChange
-Log10(Padj)
100
75
50
25

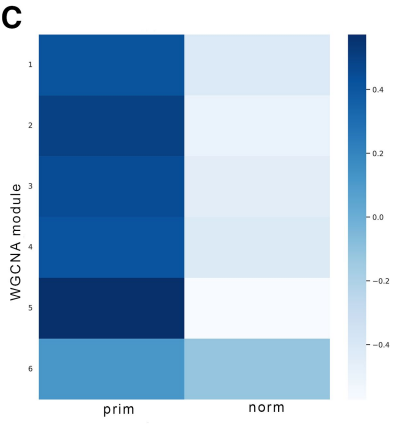

B
Regulation of neuron projection
Intrinsic component membrane
Assembly presynapse junction
Cell adhesion and migration
Cell adhesion activation
Collagen trimers
Plexin pathway semaphorin
Cell-matrix adhesion
Platelet adhesion exposed
MET promotes motility
Blood coagulation wounding
Chemical homeostasis ion
Focal adhesion
Sulfur compound heparin
Extracellular organization structure
Elastic fibre formation
Calcium ion binding
Adherens junctions interactions
Endodermal cell differentiation
Fibrinogen complex
Endopeptidase activity
Extracellular matrix binding
Linkage mapk integrins
Fibrin clotting
Locomotion motility migration
Metal ion transport
GOBP 182
GOCC 37
GOMF 25
HP 51
REAC 41
WP 17
Decreased calvarial ossification
Transport uptake insulin
Phosphorus metabolic phosphorylation
Plasma membrane protein
Focal PI3K-Akt
NCAM1 interactions
Signaling receptor tyrosine
Netrin-UNC5B signaling
Syndecan interactions ecm
Integrin mediated signaling
Hippo signaling regulation
Positive cytokine production
TNFS bind physiological
Immunoregulatory interactions lymphoid
Growth factor binding
Hippo merlin signaling
Catenin complex
Adenylate cyclase modulating
Negative chemotaxis
signaling PDGF
Signaling
Class 1 rhodopsin
Enzyme linked receptor
Icosanoid receptor activity
GPCRs class rhodopsin
Scavenging class receptors
Receptor complex
Peptide receptor activity
Binding uptake ligands
Proteoglycan binding
Transporters and receptors

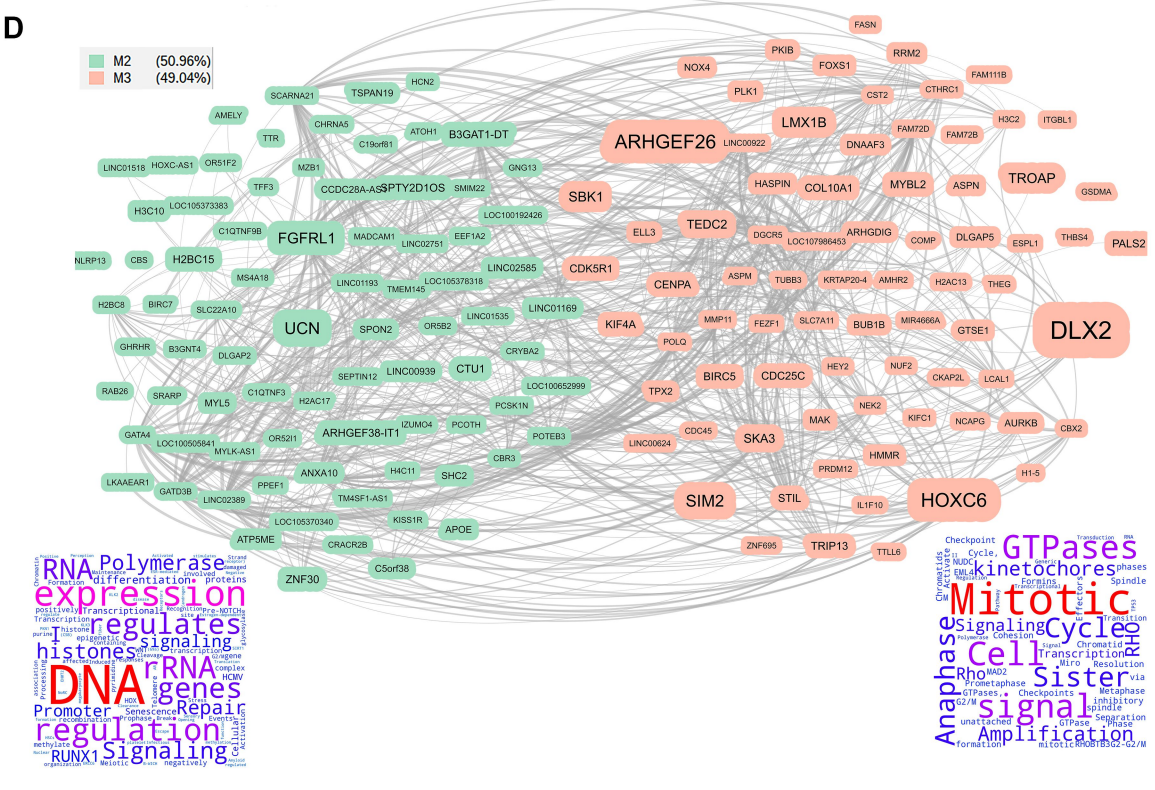

A
Top features contributing to binary model output
norm↔prim
FOXP4-AS1
DLX2
EPHA10
HPN
HPN-AS1
DLX1
NETO2
AMACR
LINC01475
HOXC6
NKX2-3
SNHG3
SLC45A2
ARHGEF26
BICD1
HOXC4
EZH2
CFAP65
UCN
CPLX1
LRFN1
CGREF1
SIM2
PRR36
LRRC56
CTBP1-AS
YPEL1
FGFRL1
IQANK1
PLA2G4C-AS1
−0.010
−0.005
0.000
0.005
0.010
0.015
SHAP value (impact on model output)
High
Feature value
Low
B
Top features contributing to binary model output
met↔prim
TMEM88B
NPFF
SPDYE2
NOTCH2NLC
PMM2
PHLDA1-DT
H3-3A
RAB5IF
IL9R
NPIPB5
NPIPB3
ARL17A
SPDYE1
MFAP3
LSM12
NPIPB12
TTLL11
SPDYE5
CCZ1B
GTF2H2C
CKS1B
PPIA
PGAM1
NT5C1A
NUTM2G
SPDYE6
CCZ1
UBE2S
FAM72B
LINC02608
−0.010
−0.005
0.000
0.005
0.010
0.015
SHAP value (impact on model output)
High
Feature value
Low

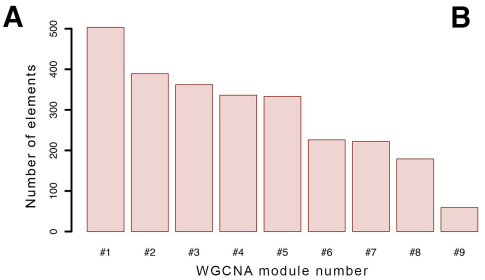


B

Module	Hub Gene
M1	DHPS
M2	MRPL41
M3	TPX2
M4	ERG
M5	SLC25A10
M6	CDHR4

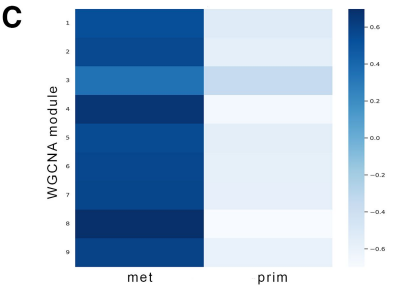


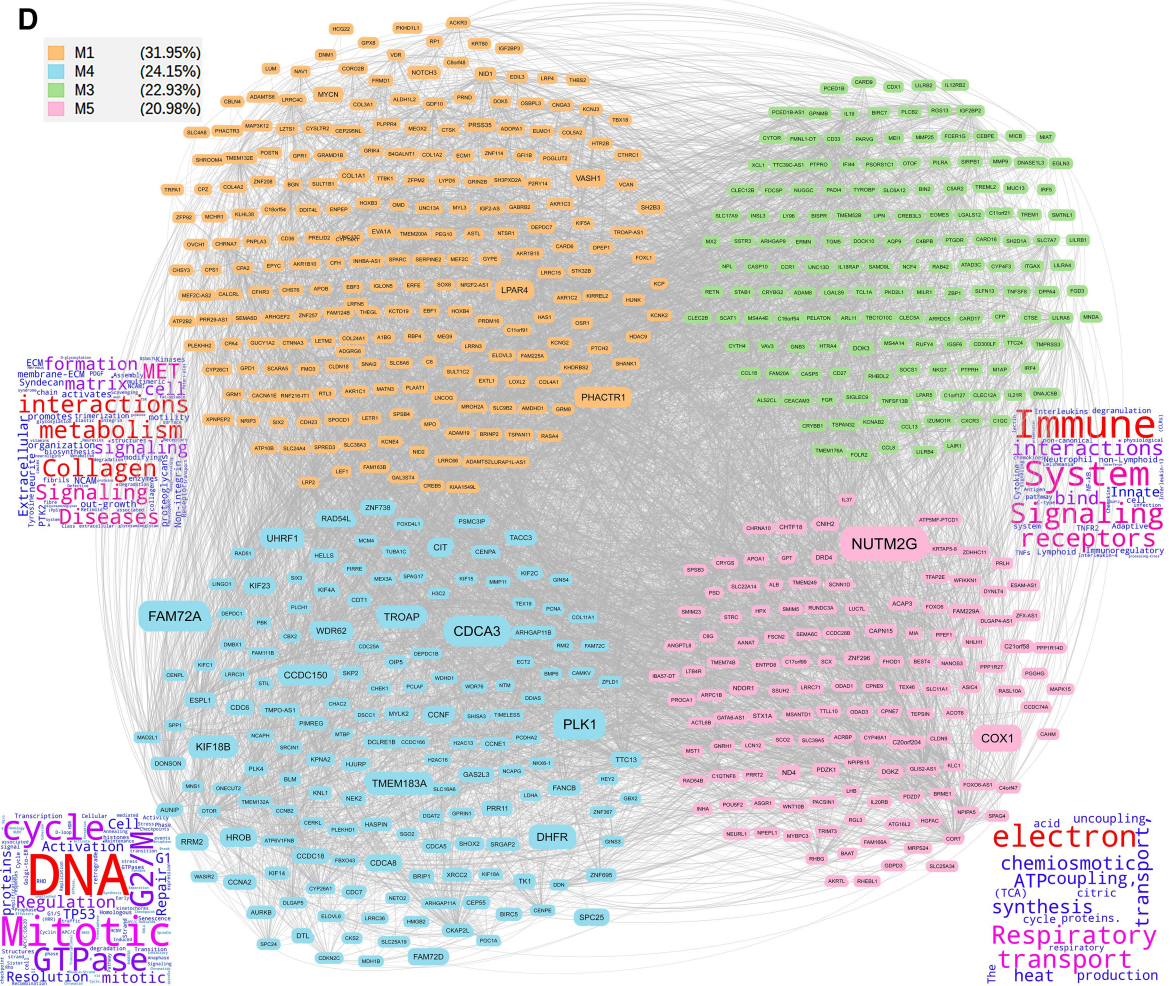


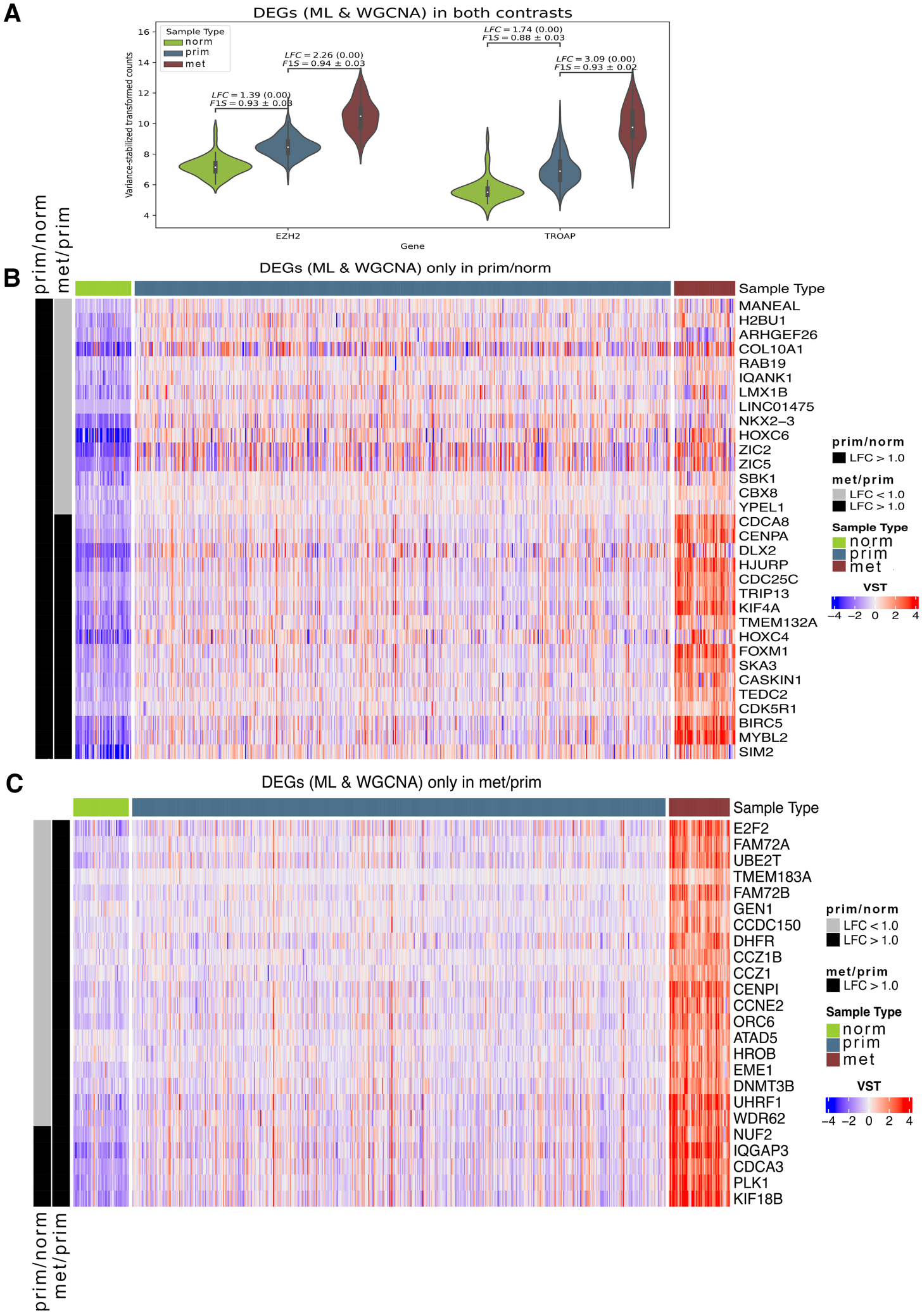


B

Module	Hub Gene
M1	FMNL3
M2	TTLL3
M3	BIN2
M4	TPX2
M5	FAM229A
M6	LHX4
M7	ESAM
M8	MFAP3
M9	TMEM160






A
DEGs (ML & WGCNA) in both contrasts
Sample Type
norm
prim
met
LFC = 1.39 (0.00)
F1S = 0.93 ± 0.03
LFC = 2.26 (0.00)
F1S = 0.94 ± 0.03
LFC = 1.74 (0.00)
F1S = 0.88 ± 0.03
LFC = 3.09 (0.00)
F1S = 0.93 ± 0.02
Variance-stabilized transformed counts
EZH2
TROAP
Gene
B
DEGs (ML & WGCNA) only in prim/norm
prim/norm
met/prim
Sample Type
MANEAL
H2BU1
ARHGEF26
COL10A1
RAB19
IQANK1
LMX1B
LINC01475
NKX2–3
HOXC6
ZIC2
ZIC5
SBK1
CBX8
YPEL1
CDCA8
CENPA
DLX2
HJURP
CDC25C
TRIP13
KIF4A
TMEM132A
HOXC4
FOXM1
SKA3
CASKIN1
TEDC2
CDK5R1
BIRC5
MYBL2
SIM2
prim/norm
LFC > 1.0
met/prim
LFC < 1.0
LFC > 1.0
Sample Type
norm
prim
met
VST
−4 −2 0 2 4
C
DEGs (ML & WGCNA) only in met/prim
Sample Type
E2F2
FAM72A
UBE2T
TMEM183A
FAM72B
GEN1
CCDC150
DHFR
CCZ1B
CCZ1
CENPI
CCNE2
ORC6
ATAD5
HROB
EME1
DNMT3B
UHRF1
WDR62
NUF2
IQGAP3
CDCA3
PLK1
KIF18B
prim/norm
met/prim
prim/norm
LFC < 1.0
LFC > 1.0
met/prim
LFC > 1.0
Sample Type
norm
prim
met
VST
−4 −2 0 2 4