The Landscape of Prostate Tumour Methylation

Jaron Arbet[1,2,3,4,*], Yu-Jia Shiah[5], Takafumi N. Yamaguchi[1,2,3,4], Rupert Hugh-White[1,2,3,4], Adriana Wiggins[6], Jieun Oh[1,2,3], Tsumugi Gebo[1,2,3], Adrien Foucal[5], Robert Lesurf[5], Chol-Hee Jung[7], Rachel M. A. Dang[1,3], Raag Agrawal[1,3,4], Julie Livingstone[1,2,3], Adriana Salcedo[3,4,8], Cindy Q. Yao[5], Shadrielle Melijah G. Espiritu[5], Kathleen E. Houlahan[1,2,3,5,8], Fouad Yousif[5], Lawrence E Heisler[5], Michael Fraser[5,9], Amar Kishan[1,10], Alejandro Berlin[9,11], Melvin LK Chua[11,12,13], Theodorus van der Kwast[9], Christopher M. Hovens[14,15], Robert G. Bristow[8,9,11,16], Paul C. Boutros[1,2,3,4,8*]

[1] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, California
[2] Institute for Precision Health, University of California, Los Angeles, California
[3] Department of Human Genetics, University of California, Los Angeles, California
[4] Department of Urology, David Geffen School of Medicine, University of California, California
[5] Ontario Institute for Cancer Research, Toronto, Canada
[6] Department of Mathematics and Statistics, Carleton College, Northfield, MN
[7] Melbourne Bioinformatics, University of Melbourne, Parkville, VIC, Australia
[8] Department of Medical Biophysics, University of Toronto, Toronto, Canada
[9] Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[10] Department of Radiation Oncology, University of California, Los Angeles
[11] Department of Radiation Oncology, University of Toronto, Toronto, Canada
[12] Divisions of Radiation Oncology and Medical Sciences, National Cancer Centre, Singapore
[13] Duke-NUS Medical School, Singapore
[14] Department of Surgery, University of Melbourne, Australia
[15] Victorian Comprehensive Cancer Centre, Parkville, Australia
[16] CRUK Manchester Institute and Manchester Cancer Research Centre, Division of Cancer Sciences-FBMH, University of Manchester, UK
[b*] Correspondence to pboutros@mednet.ucla.edu or JArbet@mednet.ucla.edu

Abstract

Prostate cancer is characterized by profound heterogeneity in its clinical trajectory. While genomic heterogeneity has been well-characterized, epigenomic heterogeneity remains less understood. To fill this gap, we compiled 2,149 multi-ancestric prostate methylomes spanning normal tissue through localized disease of varying grades to poly-metastatic disease, most with multi-omic DNA and/or RNA characterization. We identify four subtypes that varied by stage, grade and mutational subtype. We identify extensive interplay between DNA ploidy and DNA methylation, with transcriptional consequences that vary across driver-genes. Each major prostate cancer driver gene mutation triggers a specific epigenetic dysregulation, and we define a set of 14 reusable models that accurately predict clinico-molecular features of a prostate cancer from DNA methylation data. Specific epigenetic features predict disease aggression, including metastasis, with epigenomic and genomic features synergizing to optimize predictions. These results define a complex interplay between tumour genetics and epigenetics that converges to modify gene-expression programs, and subsequent clinical presentation.

Introduction

Over 280,000 men were diagnosed with prostate cancer in the United States in 2023, and it was the second leading cause of cancer mortality amongst men (1). Prostate cancer is highly heritable, with ~60% of variability in incidence being explained by rare and common germline genetics (2). While the majority of new diagnoses occur when the disease is localized and thus potentially curable, treatment outcomes vary widely (3). Current clinical prognostic factors only partially explain the heterogeneity in clinical outcomes to first-line local and systemic therapies (4). Once prostate cancer becomes hormone-therapy resistant and metastatic, it is uncurable.

This clinical heterogeneity is matched by extensive genomic heterogeneity. Localized disease in particular exhibits extensive intra- and inter-tumoural variability, which is associated with differential clinical outcomes (5-9). Somatic genetic heterogeneity is observed alongside extensive epigenomic and gene-expression variability (6,10-14). For example, aberrant methylation of CpG islands (CGI) in gene promoter regions is associated with transcriptional inactivation of tumour suppressor genes, oncogene activation and genome instability (15-18). A hypermethylated subtype has been identified in numerous cancers including prostate cancer (19,20). We and others have identified individual methylation loci that are independently associated with adverse clinical outcomes (7,21-24) when compared to current clinical biomarkers (25). However, these studies have been limited by small sample sizes, limited clinical data, limited sampling of the full natural history of the disease, and have not determined how DNA methylation interacts with other types of somatic mutations.

To fill this gap, we created a compendium of DNA methylation from 2,149 prostates: 1,824 prostate cancers and 325 non-cancerous prostate tissue. Our tumours represent a range of risk groups within both localized and metastatic disease, and arise in individuals from a range of ancestries. Using this resource, we identify distinct DNA methylation subtypes associated with adverse clinical outcomes and demonstrate that tumour-associated RNA abundance is jointly regulated by DNA methylation and gene copy number. We comprehensively define the epigenomic and transcriptomic consequences of somatic driver mutations in prostate cancer, and demonstrate how these can predict clinical outcomes. These data significantly extend our understanding of the interplay between genetic and epigenetic variability, and how these jointly rationalize the clinical heterogeneity that characterizes prostate cancer.

Results

Methylation subtypes of prostate cancer

We created a cohort of 2,149 prostate tissue methylomes from seven cohorts, one previously unpublished (6,7,14,22,26-28). This compendium represents the full spectrum from normal tissue (15%; n = 325) to localized cancer (80%; n = 1,713) and castrate-resistant metastatic cancer (5%; n = 111). The localized tumours represent all five ISUP grade groups (86 to 532 patients in each group; **Supplementary Figure 1a-e**, **Supplementary Table 1**).

Comprehensive methylation subtypes were derived in the largest cohort (TCGA: 498 tumours and 50 normal prostates) and verified in the remaining seven cohorts. Expanding on previous, smaller subtyping studies (6,28-30), we define four methylation subtypes, termed MS-1 through MS-4 (**Figure 1a**; **Supplementary Figure 1f**). To allow these subtypes to be applied to new samples we identified 5,486 subtype-defining CpGs and created a nearest shrunken centroid classifier for single-sample subtype-assignment (**Supplementary Figure 1g, Supplementary Figure 2**); this is available in an open-source R package associated with this submission. Subtypes were not driven by tumour purity (**Supplementary Figure 3a**).

The four methylation subtypes are tightly associated with clinical features of the disease. For example, subtype MS-2 overwhelmingly occurs in non-malignant prostatic tissue (96%; **Figure 1b**) and is the least common subtype in both localized (15%) and metastatic (2%) disease. In localized disease, it preferentially occurred in low-grade tumours (80% ISUP grades 1 or 2; Gleason Grade Groups 3+3 or 3+4). By contrast, metastatic prostate tumours preferentially exhibited subtype MS-4 (79% of metastatic compared to 32% of localized and <1% of normal prostate). Subtype distributions varied between normal prostate and localized disease ($p = 5.6 \times 10^{-197}$) and between localized and metastatic disease ($p = 1.3 \times 10^{-22}$; **Supplementary Figure 3b**), suggestive of subtype evolution during disease progression and response to treatment.

For patients with localized disease, subtypes MS-1 and MS-2 had the best outcomes and MS-4 the worst (HR = 2.8; $p = 5.3 \times 10^{-6}$; **Figure 1c**). MS-4 remained independently prognostic after adjusting for cohort, age, pathologic T-category, ISUP grade group, pre-treatment serum PSA and total CNA burden (adjusted HR = 2.04, p = 0.0068). These associations held in each dataset independently, and for a composite outcome of metastasis or death (**Supplementary Figure 3c-g**). Methylation subtypes were also significantly associated with pre-treatment prognostic features including tumor size and extent (T-category), ISUP grade group and serum

PSA abundance, with consistent effect-sizes across cohorts and in the full compendium (q-values < 0.05, **Figure 1d**; **Supplementary Figure 3h-k**).

The vast majority of subtype-defining CpGs (5,400/5,486; 98.4%) showed significant tumour-normal differences (U-test q-value < 0.05; **Supplementary Figure 4a**). Subtype-defining CpGs were modestly more likely to differ between tumour and normal samples than non-subtype-defining CpGs (relative risk = 1.10; $p = 2.6 \times 10^{-162}$). Similarly, 82.2% of subtype-defining CpGs showed significant differences between localized and metastatic samples, compared to 75.8% of non-subtype-defining CpGs (relative risk = 1.08; $p = 9.6 \times 10^{-30}$; **Supplementary Figure 4b**). Subtype-defining CpGs were significantly enriched on five chromosomes and depleted on seven (**Figure 1e**), and preferentially occurred in open sea rather than islands or shores (**Figure 1f**) and in intergenic regions rather than promoters or gene bodies (**Figure 1g**).

Tumour ploidy drives regional methylation & gene-expression

These positional biases in the genomic locations of subtype-defining CpGs suggested an association between copy number state and methylation status (7,31-33). We therefore considered the 870 tumours in our compendium with matched copy number alteration (CNA) profiling. Methylation was predominantly inversely correlated with ploidy, with 86% of genes showing negative correlations (**Figures 2a-b**). Given the inability of array-based technologies to quantify allele-specific methylation status, it remains unclear if the average methylation per allele differs as a function of the number of copies. Methylation of promoter CpG islands was strongly associated with decreased transcript abundance, with negative correlations for 92% of genes (**Figure 2c**). In contrast, CNA and mRNA were positively correlated in 86% of genes (**Figure 2d**).

To better delineate how the epigenome with copy number alteration jointly regulate mRNA abundance, we considered 725 tumours with all three data-types available. Given that CNAs are common and early driver events in prostate cancer, we hypothesized that DNA copy number may interact with methylation to regulate RNA abundance. Linear regression was used to test the interaction between CNA and methylation on RNA abundance (z-score transformed $\log_2$ TPM). Fully 45.3% of genes (4,950 / 10,919) exhibited significant CNA-methylation interactions ($q < 0.05$; **Figure 2e,f**). These associations were an even mix of synergistic and antagonistic effects: that is, gains were equally likely to be hypo- *vs.* hyper-methylated (34). Six examples of notable interactions are given in **Figure 2f**, including the cancer driver gene *PTEN*. In each

of these cases, the difference in RNA abundance between hyper- and hypo-methylated groups is much greater in patients with CNA gain or loss compared to patients with neutral copy number. Further, the *PTEN* interaction was independent of both T-category and ISUP grade-group (**Supplementary Figure 4c,d**). Taken together, these data support the hypothesis that DNA copy number changes lead to subsequent epigenomic changes that are evolutionarily selected and jointly dysregulate prostate cancer driver genes.

Next, we asked whether the associations of ploidy with methylation is primarily a feature of "passenger" or driver events. We considered 91 prostate cancer driver genes (7,35-37) to the remaining group of primarily non-driver genes. We stratified our analyses by CNA status and disease stage (localized *vs.* metastatic). Unexpectedly, in localized disease methylation changes in driver and passenger genes were indistinguishable on average: driver genes showed similar amounts of differential methylation as non-driver genes: this was true for both losses and gains (**Figure 2g**). By contrast, in metastatic disease CNA changes were associated with much larger methylation changes overall, and these were significantly enhanced in driver genes (**Figure 2g**).

To better understand which specific regulatory events vary in evolutionary importance between localized and metastatic prostate cancers, we compared methylation, CNA and RNA interactions between these disease states. Localized cancers were dramatically more likely than metastatic ones to show lower methylation as ploidy increases (**Figure 2h, Supplementary Figure 5a**). Similarly, localized cancers were dramatically more likely than metastatic ones to show lower methylation as mRNA abundance increases (**Figure 2i, Supplementary Figure 5b**). By contrast, the association of copy-number and mRNA abundance was largely indistinguishable between localized and metastatic disease (**Figure 2j, Supplementary Figure 5c**). Taken together, this suggests that methylation is more central to the regulation of gene-expression in localized than in metastatic prostate cancer (**Supplementary Figure 5d**).

To confirm this result, we assessed the mutual information between pairs of data types, which can capture nonlinear relationships. Consistent with our simple correlation analyses, mutual information showed substantial differences between localized and metastatic prostate cancer, particularly in methylation-RNA and methylation-CNA associations. By contrast, the influence of CNAs on RNA abundance was relatively consistent across both localized and metastatic disease (**Supplementary Figure 6**). These data suggest that copy number events in prostate

cancer have stage-specific influences on methylation, allowing a single somatic mutation to influence gene-expression in different ways during the natural history of the cancer.

Methylation subtypes with high CNA burden

To assess differences in CNA between the methylation subtypes, we compared the CNA frequency for each CpG using ordinal logistic regression, adjusting for total copy number burden and cohort (TCGA, ICGC PRAD-CA, Zhao *et al.* 2020). A fifth of genes containing subtype defining CpGs showed significant copy number alteration differences between the four subtypes (**Figure 3a**). Methylation subtypes MS-3 and MS-4 had noticeably higher percent of patients with copy number gains across the genome, while MS-4 had a higher percent of patients with copy number loss in chromosomes 2, 5, 6 and 13 (**Figure 3b, Supplementary Figure 7**). The subtypes differed significantly in their overall CNA burden (Kruskal-Wallis $p = 2.02 \times 10^{-55}$; **Figure 3c**). Tumours in MS-2 had the fewest CNAs, consistent with the observation that most normal samples (96%) had this subtype. Taken together, these data show that subtype-defining CpGs discriminate between tumour *vs.* normal tissue, localized *vs.* metastatic tissue, and have copy number alteration profiles that differ across subtypes. Subtypes MS-3 and MS-4 had significantly higher distributions of percent genome alteration (PGA), consistent with their higher risk of biochemical recurrence (BCR); while the methylation subtype most similar to normal prostate tissue (MS-2) had the least amount of copy number alteration events.

Having identified methylation subtypes associated with BCR and different CNA profiles, we hypothesized that genes with larger variation in methylation may be predictive of BCR and higher CNA burden. Of the top 500 genes with highest variation in methylation ("HVM genes"), 30 were significantly associated with BCR ($q < 0.05$; **Supplementary Figure 8a**). The HVM genes showed greater variation in size and direction of their correlation with PGA compared to non-HVM genes, with 30% of HVM genes being positively correlated with PGA, compared to only 9% of non-HVM genes (**Supplementary Figure 8b**). The HVM genes were not more likely to have copy number alterations compared to non-HVM genes (**Supplementary Figure 8c**) and did not show enrichment based on chromosomal location (**Supplementary Figure 8d**). The 30 HVM genes that were significantly associated with BCR showed strong correlation patterns within their CNA and methylation profiles (**Supplementary Figure 8ef**).

mRNA profiles differ by methylation subtype

To determine if the genome-epigenome interactions were associated with downstream gene-expression, we next considered 738 prostate tumours with matched methylation and RNA profiling. When assessing subtype differences, 40,496 of 53,078 (76%) mRNA transcripts significantly differed between subtypes ($q < 0.05$; adjusting for PGA and cohort differences). We chose to focus on a subset of 3,516 significant transcripts with large effect sizes. These transcripts showed unique expression profiles between the methylation subtypes (**Supplementary Figure 9a**) with subtypes MS-1 and MS-2 tending to have higher expression of them (**Figure 3d**). These differences reflected a broad range of cancer hallmark pathways (**Figure 3e**), with MYC targets showing the largest differences between subtypes (38,39).

Hypoxia, or low oxygen levels in the tumor microenvironment, is known to alter gene expression profiles, serving as a predictor of disease recurrence and metastasis (33,40-42). To assess the relationship between hypoxia and methylation, we compared a hypoxia RNA signature (43) between the four methylation subtypes. The hypoxia score significantly differed between subtypes ($p = 0.017$), with MS-2 having lower scores and little difference between the remaining subtypes (**Supplementary Figure 9b**). Subtype-defining CpGs had 1.6 times greater odds of being correlated with hypoxia compared to non-subtype-defining CpGs (95% CI: 1.4 to 1.8, $p = 1.48 \times 10^{-15}$; **Supplementary Figure 9c**). CpGs throughout the genome were associated with hypoxia score, with 69% of significant correlations ($q < 0.05$) being inversely related (**Supplementary Figure 9d**)**.** Methylation of CpG islands in gene promoters was significantly correlated with hypoxia score in 32 genes (**Supplementary Figure 9e**).

Somatic driver mutations shape the prostate cancer epigenome

Prostate cancer is a C-class tumour (7,44), with few recurrent driver SNVs, but an abundance of driver structural variants. It is unknown whether driver genomic aberrations interact with DNA methylation to promote prostate cancer growth and clinical aggression. To address this question, we evaluated the association of global methylation with 28 recurrent driver aberrations (7), including SNVs, mutation density, CNAs and GRs in 870 patients with localized prostate cancer. PGA showed the greatest effect on methylation levels (87.5% significant CpGs, $q < 0.05$) followed by global SNV burden (87.0% significant CpGs, $q < 0.05$) and *MYC* copy number gain (79.7% significant CpGs, $q < 0.05$; **Figure 4a**). For each driver mutation associated with one or more CpGs, methylated status (*i.e.* β-value > 0.61; **Supplementary Figure 10a**) was found in 48% of significant CpGs, with 69% of significant CpGs occurring in

in non-CGI regions (**Supplementary Figure 10b**). Moreover, three of the top four driver events with the greatest number of significant CpGs were related to mutation density, suggesting that mutation burden plays a crucial role in regulating tumour-associated methylation (**Figure 4a**).

To better understand the relationship between driver mutations, methylation and downstream mRNA abundance levels, we next evaluated the methylation profile of the top 100 significant CpGs for PGA, SNV burden, *MYC* copy number gain and number of genomic rearrangements (*i.e.* the top four drivers in **Figure 4a**). Distinct methylation patterns were identified among patients with and without *MYC* gain or patients who have higher or lower PGA, SNV burden, and genomic rearrangements (defined by a median threshold; **Figure 4b**). The distinct methylation profiles for PGA, SNV burden and MYC gain contribute to significant changes in mRNA abundance of the genes corresponding to the 100 significant CpGs from **Figure 4b** (Mann-Whitney U test, $p < 0.05$; **Supplementary Figure 10c**), further supporting an intimate interplay between methylation and mRNA abundance levels in prostate cancer. Lastly, the methylation subtypes were significantly associated with 12 of 28 driver mutations ($q<0.05$, logistic regression adjusting for cohort and PGA), with MS-3/4 tending to have greater odds of mutations relative to the normal subtype MS-2 (**Figure 4c**, **Supplementary Figure 10d,e**).

After demonstrating the impact of somatic driver mutations on the prostate cancer epigenome, we sought to compare the total somatic driver mutation burden between the four methylation subtypes. The number of driver events with at least 1 significant CpG (from **Figure 4a**) were quantified for each patient and compared between the four methylation subtypes (**Figure 4d, Supplementary Figure 10f**). Subtype MS-4 had a significantly higher distribution of driver events, while MS-2 had a significantly lower distribution ($q < 0.05$). The percentage of patients having at least 4 driver mutation events in descending order was subtype MS-4: 73%, MS-3: 58%, MS-1: 49% and MS-2: 10%. Taken together, these data support the hypothesis that tumour DNA methylation is influenced by driver genomic aberrations, and that these classes of aberration act jointly to regulate global mRNA abundance patterns.

Given the high mutational burden observed in MS-4, we hypothesize that similar epigenomic dysregulation may precede and contribute to the elevated mutational load characteristic of metastatic disease. To explore this, we analyzed differences in driver mutation rates between localized and metastatic samples. Metastatic samples had significantly higher genomic rearrangements, PGA and total SNV burden ($p < 0.05$) and significantly different odds of mutation for 11 driver events, adjusting for the corresponding total mutation burden ($q < 0.05$;

i.e. CNA drivers adjusted for PGA, SNVs adjusted for $\log_2$ SNVs per Mbps, and genomic rearrangements adjusted for $\log_2$ GR count) (**Supplementary Figure 11**).

Epigenetic signatures of clinical features and somatic driver mutations

Having demonstrated the association between the prostate cancer epigenome with clinical features and somatic driver mutations, we next sought to determine whether the epigenome can predict these features for new patients. A sample of 1,436 patients were randomly split into a 70% train (n=1,005) and 30% test (n=431) dataset. Random forest and glmnet models were compared for predicting each feature, with prediction accuracy assessed by Pearson's correlation for continuous features and balanced accuracy (BA = (sensitivity + specificity) / 2) for categorical features. The predictors consisted of prostate cancer gene-level methylation for 13,603 genes, defined as the median beta-value among CpG islands in the gene promoter region. Feature selection methods were performed in the training dataset in attempt to improve performance and identify a smaller subset of important predictors (**Methods**; **Supplementary Figure 12**). We did not find evidence of overfitting as there was little difference between training and test model performance (**Supplementary Figure 13**). To encourage a more parsimonious model, the final model for each feature was chosen as the model with the smallest number of predictors that was within 0.02 of the best Pearson's correlation and balanced accuracy for numeric and categorical features, respectively.

Performance of the final models on the test dataset is shown in **Figure 4e-g** and **Supplementary Figure 14**. For numeric features, performance on the test dataset resulted in Pearson's correlations between the true and predicted values of 0.80 ($\log_2$ PSA), 0.72 ($\log_2$ SNVs per Mbps), 0.61 (PGA) and 0.49 (Age). For binary features, the minimum test specificity was 67% with 6 out of 8 features exceeding 78% specificity. The minimum test sensitivity was 53.5%, with two features obtaining sensitivity greater than 76% (*MYC* gain and *PTEN* loss). Categorical features with three or more categories were challenging to predict, as at least one category always had accuracy below 50%. Among these multicategory features, tumour category had the highest balanced accuracy (0.71) with 71% of samples in tumour categories 2-4 correctly classified. Most models were robust to cohort-level differences, except for continuous PSA, while categorical PSA did not exhibit cohort bias (**Supplementary Figure 15**). Thus we decided to only make the categorical PSA model publicly available, rather than continuous PSA. Overall, these results show that the prostate cancer epigenome can predict clinical features and somatic driver mutations with moderate to high accuracy. The PrCaMethy

R package is made available for predicting these features in new patients using prostate cancer methylation data from the Illumina Infinium Human Methylation 450K BeadChip array. The workflow of the three main R functions is given in **Supplementary Figure 16a** and the overlap of gene predictors between all models is shown in **Supplementary Figure 16b**.

A multi-omic signature predicts disease recurrence

A total of 774 localized prostate cancer patients with time-until biochemical recurrence (BCR) were randomly split into a 70% train and 30% test dataset. Random forest and a regularized Cox proportional hazards model were compared for predicting time-until-BCR, with prediction accuracy assessed using the C-index. Several combinations of data types were considered for model training (**Supplementary Figure 17a-b**). Given the large number of features, feature selection was performed in the training dataset to identify a smaller subset of potential predictors (**Methods**; **Supplementary Figure 17c-d**).

A random forest model resulted in the highest C-index in both the training and test datasets (**Figure 4h**) and selected a multi-omic signature containing 20 features: 2 clinical (PSA and ISUP grade), 11 methylation features (methylation subtype and CpG island methylation in the promoter region of 10 genes: *CRYM*, *NXNL2*, *TP53INP2*, *ELOVL3*, *HIST1H2AI*, *HIST1H2BL*, *SRRM2*, *PPIE*, *C1orf92*, *SLITRK4*) and RNA abundance in 7 genes (*BCL2L14*, *MATN3*, *LEFTY2*, *RBM46*, *DKK2*, *TMEM130*, *MIR3616*). The random forest C-index in the test dataset was 0.86 (95% CI 0.77-0.91), and was significantly higher than the test C-index of a random forest that only considered pre-treatment clinical features (0.81, 95% CI for the difference: 0.0011 to 0.11). When median-dichotomizing the predicted risk score of the final model, the high-risk group had 8.2-fold greater risk of BCR (95% CI 1.9-36.1; p = 0.0055; **Figure 4j**) and the predicted risk score significantly differed between non-malignant, localized and metastatic prostate samples (**Figure 4k**). The random forest risk score remained prognostic even within the subset of patients classified as intermediate risk (HR = 16.23, p = 0.013; **Supplementary Figure 17e**). Overall, these results suggest the prostate cancer epigenome and its interactions with the transcriptome play a key role in disease evolution and recurrence.

Discussion

Herein, we created a compendium of 1,824 prostate tumour methylomes (1,713 localized, 111 metastatic) compared with 325 normal prostate tissue methylomes, and characterized the interactions between aberrant methylation and other classes of recurrent molecular aberrations. We identified four methylation subtypes that were prognostic of rapid biochemical relapse (in cohorts TCGA, ICGC PRAD-CA, Kirby *et al*. 2017) and metastasis or death (Zhao *et al.* 2017), providing important insights into the molecular subclasses that define aggressive localized prostate cancer. Subtype MS-4 had adverse outcomes in all cohorts of localized prostate cancer, with 2.77 times greater risk of BCR when pooling cohorts and retained statistical significance when adjusting for pre-treatment clinical features and PGA. Consistent with their higher risk of BCR, subtypes MS-3/4 had significantly higher distributions of PGA and MS-4 had a higher distribution of somatic driver mutation events. The 5,486 subtype-defining CpGs were shown to discriminate between tumour *vs.* normal tissue, localized *vs.* metastatic tissue, have copy number alteration profiles that significantly differed across subtypes and were more likely to be associated with hypoxia compared to non-subtype-defining CpGs.

We showed that mRNA gene expression profiles significantly differed between the four methylation subtypes, with subtypes MS-1/2 tending to have higher expression and MS-3/4 having lower expression among the most significant transcripts. We investigated how DNA methylation interacts with copy number status to influence mRNA abundance through *cis*-mechanisms, while further research is needed to assess *trans*-associations (*i.e.* methylation-CNA interaction within a gene that affects other genes). For example, the *PTEN* tumour suppressor gene was synergistically regulated such that both hypermethylation and hemizygous deletion resulted in further decreased mRNA abundance. Moreover, this *PTEN* loss has also been shown to lead to unique DNA methylation profiles (45). This suggests that *PTEN* haploinsufficiency, which has been widely reported in prostate and other cancers (46-48), may in part be explained by suppression of *PTEN* function due to aberrant methylation of the remaining allele. These data support a novel mechanism of somatic tumour suppressor loss in localized prostate cancer, involving hemizygous copy number loss and silencing of the remaining allele *via* methylation. Previous studies have established that key prostate cancer SNVs in genes such as *SPOP* and *TP53* can alter methylation profiles (49). Furthermore, recent evidence suggests that other forms of epigenomic regulation, including chromatin modifications such as H3K27 acetylation, are key drivers of prostate cancer phenotypes, *via* interaction with

recurrent somatic aberrations such *TMPRSS2:ERG* fusion and androgen receptor mutations (50-52).

Consistent with these findings, indices of global mutational burden, such as SNVs/Mbp, PGA and number of genomic rearrangements, are strongly associated with alterations in global methylation patterns. Overall somatic mutational burden was shown to be higher in MS-4 and in metastatic versus localized samples, indicating that epigenomic dysregulation similar to MS-4 may be a precursor to the increased mutational load seen in metastatic disease. Collectively, these suggest that the genomic architecture of prostate cancer is regulated at multiple levels and that deep multi-parametric analyses are required to ascertain the determinants of aggressive disease.

Prostate cancer epigenetic signatures were derived for 15 pre-treatment clinical features and somatic driver mutations. Although prior research has shown the epigenome can predict a patient's age (53), our results suggest the epigenome can be used to predict other clinical and molecular features. The models for predicting these features are made publicly available in the PrCaMethy R package, thus enabling researchers to predict these features in new patients. Further, the R package allows researchers to assign new patients to the four prostate cancer methylation subtypes.

Lastly, a multi-omic signature of 20 clinical, methylation and mRNA features was predictive of time until biochemical recurrence and significantly improved accuracy over a model that only considered clinical features. Although this suggests the prostate cancer epigenome and its interactions with the transcriptome play a key role in disease evolution, future studies with larger samples sizes are needed to definitively identify predictive biomarkers and characterize interactions between different 'omic data types. Additional studies for developing models to validate our findings and predict metastasis-free survival and prostate cancer specific survival are also needed.

A potential limitation of this study is the pooling of multiple cohorts that differ based on the technology used to measure methylation and genomic data, and data preprocessing procedures. Although we adjusted for differences between cohorts in our analyses where possible, future research with a large sample of diverse patients with standardized data collection and processing may improve results. In addition, although our prostate cancer subtypes were derived using methylation data alone and showed significant associations with

clinical and multi-omic data, future research can use clustering methods to derive subtypes based on the interactions between all data types (*e.g.* clinical, methylation, CNA, RNA, driver mutations) and assess whether these integrated subtypes have stronger associations with outcomes compared to subtypes derived from one data type. Lastly, future research is needed to build more ancestrally diverse cohorts with larger sample sizes for metastatic castration-resistant and non-castrate-resistant prostate cancers.

In summary, we identified novel methylation subtypes that stratify patients based on disease prognosis and have different CNA, RNA and somatic driver mutation profiles. We provide novel insights into the associations between genomic mutations and methylation patterns and their potential relation in dysregulation of the tumour transcriptome. Prostate cancer epigenetic signatures were derived for predicting 14 clinical features and somatic driver mutations, while a multi-omic methylation and mRNA signature improved prediction of disease recurrence over standard clinical features. A consistent narrative emerges that epigenomic dysregulation is associated with altered gene expression, higher rates of CNA and somatic driver mutations, potentially evolving to advanced metastatic disease. These findings highlight the intricate relationship between epigenetic, genomic and transcriptomic alterations, deepening our understanding of DNA methylation as a key factor in the clinical aggressiveness of localized prostate cancer.

Materials and Methods

Data processing

Patient cohorts

The TCGA cohort (n = 498 tumour samples) (6) was used to derive the methylation subtypes which were validated in seven external prostate cancer cohorts: Zhao *et al.* 2017 (n = 525) (22), ICGC PRAD-CA (n = 286) (7), Li *et al.* 2020 (n = 186) (26), ICGC PRAD-FR (n = 130) (28), Zhao *et al.* 2020 (n = 100) (14), Kirby *et al.* 2017 (n = 72) (27) and Melbourne (n = 27). To evaluate the relationship between the subtypes and methylation profiles in normal tissue, the subtypes were classified into four cohorts of normal tissue: Li *et al.* 2020 (n = 187), Kirby *et al.* 2017 (n = 63), TCGA (n = 50) and Melbourne (n = 25). TCGA, ICGC PRAD-CA and Zhao *et al.* 2020 prostate cancer cohorts were included in the methylation, mRNA, CNA, and somatic driver mutation analyses. See **Supplementary Figure 1a** for sample sizes per cohort and data type used in each analysis.

All International Cancer Genome Consortium (ICGC) samples were processed from prostate cancer tissue taken from patients with localized disease, following ethics approval from the University Health Network Research Ethics Board (REB)–approved study protocols (UHN 06-0822-CE, UHN 11-0024-CE, CHUQ 2012-913:H12-03-192) as previously described (7).

Pathology and Tissue Processing

ICGC PRAD-CA fresh-frozen tissue specimens were audited for tumour content, Gleason/International Society of Urological Pathologists (ISUP) grade, and tumour cellularity by two independent genitourinary pathologists, as previously reported (7). Specimens of at least 70% tumour cellularity were manually macro-dissected and digested for nucleic acid extraction, as previously reported (7,8).

Microarray data generation

ICGC PRAD-CA methylation, SNP and mRNA microarray data were generated as previously described (7). Briefly, Illumina Infinium HumanMethylation450 BeadChip kits were used to assess global methylation levels with 500 ng of input genomic DNA at the McGill University and Genome Quebec Innovation Centre (Montreal, QC). SNP microarrays were performed with 200 ng of DNA on Affymetrix OncoScan FFPE Express 2.0 and 3.0 arrays. Where DNA quantities were limiting (17 samples), we used whole-genome amplification (WGA; WGA2, Sigma-Aldrich,

St. Louis, MO), and validated that WGA of genomic DNA does not significantly alter CNA profiles (8). Total RNA was extracted from alternating adjacent sections (to reduce potential heterogeneity relative to the DNA specimen), using the mirVana miRNA Isolation Kit (Life Technologies). Extracted total RNA (100-150 ng) was assayed on Affymetrix Human Gene 2.0 ST (HuGene 2.0 ST) array and Human Transcriptome Array 2.0 (HTA 2.0) platforms at The Centre for Applied Genomics (The Hospital for Sick Children, Ontario, Canada) and the London Regional Genomics Centre (Robarts Research Institute, London, Ontario, Canada).

Methylation

Methylation processing was performed in R statistical environment (v3.4.0). Methylation data for the ICGC PRAD-CA cohort is in the Gene Expression Omnibus (GEO): GSE107298, and was processed using the Dasen method (54,55) as previously described (7). Probes that have detectability below background noise, non-CpG methylation, SNPs with an annotated minor allele frequency of greater than 5% with a maximum distance of two nucleotides to the nearest CpG site, and those known to cross-hybridize to multiple genomic locations were excluded. Probe average intensity levels were calculated across technical replicates. Probes were annotated to gene symbols *via* the IlluminaHumanMethylation450kanno.ilmn12.hg19 annotation package (v0.6.0). Unless stated otherwise, all methylation analyses were performed using the β-value, which represents the ratio of methylated intensity over combined methylated and unmethylated intensity and reflects the methylation level of a specific CpG loci. Tumour purities were estimated using LUMP as previously described (56).

Illumina Infinium Human Methylation 450K BeadChip CpG processed beta-values were obtained for the following cohorts: Kirby *et al*. 2017 (NCBI GEO GSE76938) (27), Zhao *et al.* 2017 (dbGaP study accession phs001921.v1.p1) (22), TCGA (57,58) and ICGC PRAD-FR (28). Additional details about data processing can be found in the corresponding citations.

For Melbourne methylation data, Illumina's Infinium Human Methylation 450K array was used to profile the methylation status of all Australian prostate cancer samples. Raw signal data were processed and normalised for both within-array and between-array variations using the minfi package (59). Minfi's implementation of the Illumina's reference-based normalisation and the SWAN method within the minfi package were used for between-array and within-array (correcting for the probe-type bias) normalisations, respectively. Any methylation measurement with detection p-value greater than 0.01 were omitted from further analysis.

Whole genome bisulfite sequencing methylation data was used in Li *et al.* 2020 and Zhao *et al.* 2020 as detailed in (14,26). The “illumina450k_hg38” dataset from the Rockermeth R package (v.0.1.1) (60) was used to identify locations of the CpGs corresponding to the HumanMethylation450 BeadChip, thus allowing us to obtain the same CpGs used in the other cohorts. In Zhao *et al.* 2020, CpGs with a median number of reads < 10 and patient-level CpG calls with < 10 reads were omitted from analyses; while a cutoff of 5 was used in Li *et al.* 2020 due to lower coverage. For each CpG, the β-value was calculated as the proportion of methylated reads at that CpG site: # methylated reads / (# methylated reads + # unmethylated reads).

Copy number aberrations (CNAs)

ICGC PRAD-CA copy number alteration data was obtained as previously described (7). The output from the Affymetrix OncoScan FFPE Express 3.0 SNP probe assay was aligned to a custom reference (7) using the OncoScan Console v1.1. Copy number segments were subsequently called using the BioDiscovery's Nexus Express for OncoScan software using the SNP-FASST2 algorithm. Parameters were identical to the default one except for the minimum number of probes per segment which was set to 20. Diploid re-centering was performed when needed using the Nexus Express software. Using BEDTools the gene level copy number alterations were identified with the RefGene annotation (2014-07-15). Percent genome alteration was defined as PGA = (base-pair length of all genome regions with gain or loss) / 3.2 billion bases * 100. Zhao *et al.* 2020 CNA data was obtained from (61) and TCGA CNA data was downloaded from the TCGAbiolinks R package v2.25.3 (57,62), with copy number calls of 2 treated as “neutral”, less than 2 as “loss” and greater than 2 as “gain”.

mRNA

ICGC PRAD-CA pre-processed mRNA data was downloaded from GSE107299 and processed in R (v3.4.3). Briefly, the RMA algorithm implemented in the affy (v1.48.0) package(63) was used with modified Entrez Gene CDFs to pre-process the raw intensity data. Probes were mapped to Entrez gene ID using custom CDF files (v18) for HTA 2.0 and HuGene 2.0 ST array (64). The sva package (v3.24.4) was used to correct for batch effects between different arrays. mRNA abundance levels from HuGene 2.0 ST and HTA 2.0 were combined based on Entrez Gene IDs. The mRNA abundance levels were averaged amongst duplicated Entrez Gene IDs. Entrez gene IDs were then converted into gene symbols and chromosome locations based on

the human reference genome GRCh37 from UCSC table browser (download date: 2018/01/30). TCGA processed mRNA abundance levels were downloaded from the TCGAbiolinks R package v2.25.3 (57,65). Zhao *et al.* 2020 FASTQ files were aligned to GRCh38 with GENCODE v36 using STAR (v2.7.6a) and mRNA abundance was calculated using RSEM (v1.3.3). All mRNA abundance levels were analyzed as $\log_2$ transcripts per million ($\log_2$ TPM), unless specified otherwise.

Whole-genome sequencing

ICGC PRAD-CA raw sequencing reads were processed as previously reported except BWA-mem (v0.7.12+) was used for alignment and the Genome Analysis Toolkit (GATK v3.4.0+) was used for local realignment and base quality recalibration on tumour/normal pairs (7). SNVs, genomic rearrangements (GRs), fusion events, chromothripsis and kataegis score were also generated as previously described (7). Shortly, SNVs were detected using SomaticSniper (v1.0.5) with the same filters as described in Fraser *et al.* (7) with the updated version of COSMIC (v81). GRs were identified using Delly (v0.7.2-0.7.7) and filtered based on a mapping quality of 20, split-read and pair-end read evidence with a cut-off threshold of five and small indels (50 bp). Fusion events involving the *ERG*, *ETV1*, *ETV4* and *FLI1* genes with *TMPRSS2*, *SLC45A3* and *NDRG1* genes were considered as *ETS* events. We have also taken into account ERG immunohistochemistry results and deletion calls between *TMPRSS2* and *ERG* loci in OncoScan SNP array data to support *ETS* status. Chromothripsis scores were generated using ShatterProof (v0.14, default settings) (66). Kataegis was quantified using a fixed-width sliding window to test deviation of observed SNV trinucleotide content and inter-mutational distance from expected using an exact binomial test (67).

Corresponding whole-genome sequencing data for the Zhao *et al.* 2020 cohort (14,61) was downloaded as BAM files (See Data Availability Section). BAM files were converted to FASTQs in two batches. The first batch was converted using the commands RevertSam and SamToFastq from Picard (v2.18.23). The second batch was processed using RevertSam from Picard (v2.27.1-0) then the commands collate followed by fastq from Samtools (v1.12)(68). The FASTQs were realigned using BWA-MEM2 (v.2.2.1)(69) alt-aware mode to the human reference genome GRCh38 including decoy contigs. Read group-level SAMs were converted to BAM files using SAMtools (v1.12) and sorted by queryname using SAMtools (v.1.15.1). Duplicates were marked using MarkDuplicatesSpark from GATK (v4.2.4.1)(70). GATK was used to perform indel realignment (v3.7.0) and base-quality score recalibration (v4.2.4.1).

Somatic SNVs were called using Mutect2 (v.4.2.4.1 and v4.4.0.0)(71), SomaticSniper (1.0.5.0)(72), MuSE (2.0, 2.0.3 and 2.0.4)(73) and Strelka2 (v2.9.10)(74). A consensus SNV call set was created by retaining SNVs that were present in two or more callers using BCFtools (v1.17)(75). Consensus SNVs were then filtered using allow- and deny-lists. Deny-lists that were used included AccuSNP, Complete Genomics, dbSNP (v156)(76), duplicate gene (v68)(77), Encode DAC and Duke(78), gnomAD (v3.1.2, variants with 'PASS' flag)(79) and 1000 Genome project (80,81). Allow-lists involved Catalogue of Somatic Mutations in Cancer (COSMIC) (v92)(82) coding and non-coding variants. Somatic structural variants were called using Manta (v1.6.0)(83) and annotated using the StructuralVariantAnnotation R package (v1.18.0)(84).

Statistical Analysis

General statistical considerations

The Benjamini-Hochberg method was used to calculate multiple testing adjusted p-values to control the false discovery rate (FDR) at level 0.05 (referred to as 'q' or 'q-value' throughout) (85). When using Fisher's Exact Test on larger than 2x2 contingency tables, the stats R package fisher.test function was used with 100,000 simulation replicates.

Methylation subtypes

Methylation subtypes were identified in the TCGA cohort (n = 498) using Interpretable MiniPatch Adaptive Consensus Clustering (IMPACC, v0.1.0) (86) with default settings, 10,000 clustering partitions and CpG beta-values. CpGs were filtered to keep the top 50,000 highest variance CpGs for clustering. The silhouette method (87) was used to choose the number of subtypes that maximized the median silhouette score, resulting in 4 subtypes. Specifically, the total number of subtypes was varied from 2 to 6, and each time the median silhouette score was calculated for each cluster, with the median of the cluster-wise scores used as an overall measure of clustering accuracy. A nearest shrunken centroid classifier (88) was trained on the TCGA subtypes to classify samples in seven external prostate cancer cohorts (n = 1,326) and four cohorts of normal tissue (n = 325), as detailed in section "Patient cohorts." When classifying samples within a given cohort, all CpGs in common between that cohort and TCGA were used to train the classifier, with missing values imputed by k-nearest neighbors imputation (impute R package v1.74.1, default settings) (89).

Subtypes *vs.* clinical features

The subtypes were tested for association with pre-treatment clinical features of age, tumour category, ISUP grade group and PSA. Each of the clinical variables was categorized as in **Supplementary Table 1** and tested for an association with the methylation subtypes using Fisher's Exact Test and the bias-corrected Cramer's V effect size (51). The subtypes were then tested for associations with the following outcomes: time until biochemical recurrence (logrank test; TCGA, ICGC, Kirby *et al*. 2017), time until death (logrank test; Zhao *et al*. 2020) and a binary metastasize or death during follow-up (Fisher's Exact Test; Zhao *et al.* 2017). A permutation logrank test was used in cohorts Kirby *et al*. 2017 and Zhao *et al*. 2020, due to small and uneven group sample sizes (10,000 permutations) (90). Furthermore, the association between subtypes and BCR was assessed when pooling all patients with available BCR data (n = 774) where hazard ratios with 95% confidence intervals and Wald-test p-values were estimated using a Cox proportional hazard model treating MS-1 as the reference group. BCR was defined as two consecutive post radical prostatectomy (RP) PSA measurements of > 0.2 ng/mL, within 6 months. Patients who underwent successful salvage radiotherapy following RP (*i.e.* post-RP PSA below 0.2 ng/ml) were classified as non-failures; those that underwent unsuccessful salvage radiotherapy (*i.e.* two post-RP PSA readings above 0.2 ng/ml) were classified as failures, backdated to the original date of post-RP PSA rise. Patients who underwent other salvage treatments (*e.g.* hormones, chemotherapy) to control rising PSA were considered as having BCR.

Subtype-defining CpGs

IMPACC intrinsically performs variable selection and identified 5,486 CpGs that were important for deriving the subtypes (referred to as 'subtype-defining CpGs' throughout). The subtype-defining CpGs were compared between subtypes using Kruskal-Wallis rank sum test and the epsilon-squared effect size (91), while further compared between tumour and normal samples, and localized *vs.* metastatic tumour samples using the Mann-Whitney U-test. Chromosome position, CpG type (island, open sea, shelf, shore (92)), and gene region (gene body, intergenic, or gene promoter) were compared between the subtype-defining CpGs and non-subtype-defining CpGs using Fisher's Exact Test and odds ratio. The gene regions were defined as follows: CpGs were annotated using the IlluminaHumanMethylation450kanno.ilmn12.hg19 bioconductor R package (v0.6.0) where CpGs in gene promoter regions were defined as within 200 or 1500 base pairs from a transcription start site (TSS). CpGs that were not mapped to any

gene symbol were defined as “intergenic regions”, and remaining CpGs were defined as occurring within the gene body.

Methylation interaction with mRNA and CNAs

Adjusted mRNA abundance was calculated by fitting a linear regression model to each mRNA transcript ($\log_2$ TPM) with cohort (TCGA, ICGC PRAD-CA, Zhao *et al.* 2020) and PGA as predictors, and using the standardized residuals (z-scores) in subsequent analyses. The adjusted mRNA abundance was compared between methylation subtypes using Kruskal-Wallis rank-sum test. Pathway analysis was performed using 50 hallmark pathways (39) by first calculating patient-specific pathway scores using adjusted mRNA abundance with the GSVA R package (v.1.50.0, default settings) (93) and comparing the pathway scores between methylation subtypes using Kruskal-Wallis rank-sum test.

An mRNA signature of hypoxia was calculated separately for the TCGA and ICGC PRAD-CA cohorts using the a panel of 51 genes (43). The hypoxia score was calculated by median-dichotomizing each feature, then a patient’s score would receive +1 for each feature above the median, otherwise -1 if less than or equal to the median. The hypoxia score was then z-score transformed (separately by cohort), and compared between methylation subtypes using Kruskal Wallis rank sum test and pairwise Mann-Whitney U-tests. Spearman’s correlation was used to assess the association between each CpG and hypoxia score, with Fisher’s exact test and odds ratio to assess whether subtype-defining CpGs were more likely to be correlated with hypoxia compared to non-subtype-defining CpGs. Lastly, Spearman’s correlation was used to assess the association of hypoxia score with gene-level methylation, where the latter was calculated for each patient as the median beta-value of CpG islands in the gene promoter region.

To test the association of prostate tumour methylation levels with CNAs and mRNA abundance level, we first examined copy number effect on CpG methylation β-values. For each patient, we generated segmented CNA profiles and called each segment as: homozygous deletion, hemizygous deletion, neutral, gain or high gain. The β-values for each CpG from all patients within any given segmented CNA state were then used to examine CNA effect on methylation levels.

Spearman’s correlation was used to test the within-gene associations of CNA and methylation of CpG island in gene promoters (CGIp) and CGIp methylation with mRNA abundance levels, when pooling TCGA, ICGC PRAD-CA and Zhao *et al.* 2020 cohorts. Furthermore, linear

regression of z-score transformed $\log_2$ TPM was used to test the interaction between CNA and methylation. This model considered binary methylation (hyper *vs.* hypomethylation; median dichotomized), binary CNA status (loss *vs.* neutral if loss was the more common mutation, otherwise gain *vs.* neutral), adjusted for cohort (TCGA, ICGC PRAD-CA, Zhao *et al.* 2020) as a covariate, and removed genes with < 5 CNA events in either methylation group.

The copy number alteration status of genes containing the subtype-defining CpGs were compared between the four methylation subtypes using ordinal logistic regression from the polr function in the MASS R Package (94). A given gene's copy number status (loss, neutral, gain) was used as the dependent variable, with subtype, PGA and cohort as the independent variables, and a likelihood ratio test was calculated for subtype. Percent genome alteration was defined as PGA = (length of all genome regions with gain or loss) / 3.2 billion bases * 100 and was compared between the four subtypes using Kruskal-Wallis rank sum test and pairwise Mann-Whitney U-tests.

The top 500 genes with highest variation in methylation ("HVM genes") were tested for associations with time-until-BCR using a Cox proportional hazards model Wald test. Spearman's correlation was used to test the association between each HVM gene with PGA and compared with the distribution of PGA correlations for non-HVM genes. The percent of patients with CNA loss or gain was calculated separately for HVM and non-HVM genes and compared using Mann-Whitney U-test.

Lastly, we investigated whether the associations between methylation, CNA and RNA within each gene differed between localized and metastatic prostate cancers. Two approaches were taken: 1) visually comparing the distribution of Spearman's correlations for each pair of data types, and 2) comparing the distribution of mutual information for each pair of data types. Unlike Spearman or Pearson's correlation, mutual information can capture arbitrary linear or nonlinear relationships. The mutual information was calculated as follows: for each pairwise data type comparison, data from each gene were discretized into $N^{1/3}$ bins, where N is the intersection of the samples with data available for each pair of datatypes (95). Mutual information (MI) between two data types was defined as:

$$MI_{xy} = H_x + H_y - H_{xy}$$

where H_x and H_y are the marginal entropies of data types x and y and H_{xy} is the joint entropy, calculated using the R package Entropy (v1.2.1) (96). To reduce bias in MI estimates, the

James-Stein shrinkage estimator (97) was applied, as implemented via the entropy.shrinkage function of the entropy R package (96). MI was normalized over the mean entropy of the two input vectors. To assess statistical significance of normalized MI values, permutation testing was performed. Gene-level data were permuted 1000 times to generate a null MI distribution, and p-values were calculated as the proportion of null MI values greater than or equal to the true observed MI. Multiple testing adjusted $q < 0.05$ was regarded as statistically significant.

Driver mutations

Within the TCGA, ICGC PRAD-CA and Zhao *et al.* 2020 cohorts, CpG methylation levels were compared with each of 28 previously established somatic events (7). Spearman's correlation test was used for continuous events and Kruskal-Wallis rank-sum test for discrete events. Somatic features were divided into six categories: mutational burden (chromothripsis, kataegis, number of GRs, number of SNVs and PGA). SNVs, CNAs (losses of *CDH1*, *CDKN1B*, *CHD1*, *NKX3-1*, *PTEN*, *RB1, TP53* and gain of *MYC)*, *ETS* fusion, CTXs (translocations with break-ends within chr4:148, chr6:97, chr7:61, chr17:25 and chr21:42 Mbp blocks) and INVs (inversions with break-ends within chr3:125, chr3:129, chr10:89 and chr21:42 Mbp blocks). CpG β-values were tested for an association with the 28 somatic events and clinical features, using Spearman's ρ and Kruskal-Wallis tests for continuous and discrete features, respectively. Driver events that were significantly associated with at least one CpG ($q < 0.05$), were summed to calculate the total number of significant driver events per patient, and then compared between subtypes using the Kruskal-Wallis test and pairwise Mann-Whitney U-tests.

Lastly, the distribution of each driver mutation was compared between localized and metastatic prostate cancers. Mann-Whitney U-test p-value was calculated for continuous drivers. Fisher's exact test q-value and odds ratio were calculated for categorical drivers, as well as logistic regression odds ratio adjusting for total mutation burden (CNA drivers adjusted for PGA, SNVs adjusted for $\log_2$ SNVs per Mbp, and genomic rearrangements adjusted for $\log_2$ GR count).

Epigenetic predictors of clinical features and driver mutations

A sample of 1,436 patients were randomly split into a 70% train (n=1,005) and 30% test (n=431) dataset. Two machine learning models were compared for predicting time-until-BCR: random forest and a regularized general linear model (glmnet) (98,99). These two models were considered because both random forests and regularized linear models have been shown to perform well across a wide variety of applications (100,101). Random forest can automatically

capture potential interactions or non-linear effects (102), but is harder to interpret compared to glmnet. In contrast, glmnet assumes linear and additive effects (no interactions) by default, and has a similar interpretation as traditional logistic or multinomial classification models (estimates log odds ratios for each predictor) but performs automatic feature selection by shrinking coefficients towards zero, attempting to remove unimportant features. Random forest was fit using the randomForest R package v4.7.1.1 (103) with the mtry parameter tuned by the caret R package v6.0.94 (104) with 5-fold cross-validation to select the best value. Glmnet was fit with the glmnet R package v4.1.8 (105) setting the alpha parameter to 0.5 (*i.e.* “elastic net” penalty that is half way between ridge and LASSO (106,107)) with the main penalty parameter lambda chosen by 5-fold cross validation. Missing values were imputed by k-nearest neighbors imputation (impute R package v1.74.1, default settings) (89).

Prediction accuracy was assessed by Pearson’s correlation for continuous features and balanced accuracy (BA = (sensitivity + specificity) / 2) for categorical features. For categorical features with three or more categories, BA was calculated separately for each category using a “one-versus-all” approach (*i.e.* a given category compared to all other categories combined) and then averaged. The predictors consisted of prostate cancer gene-level methylation for 13,603 genes, defined as the median beta-value among CpG islands in the gene promoter region. Feature selection methods were performed in the training dataset in attempt to improve performance and identify a smaller subset of important predictors. Specifically, the univariate association of each gene methylation predictor with each clinical feature and driver mutation was assessed using Kruskal-Wallis rank-sum test for categorical features and Spearman’s correlation for continuous features. To account for cohort differences, linear regression was used to regress each gene against cohort, and the residuals were used in the aforementioned univariate test. Then models were fit using only the top 10, 50, 100, 1000 or 5000 most significant genes, as well as a model that considered all 13603 genes. To encourage a more parsimonious model, the final model for predicting a given feature was chosen as the model with the smallest number of predictors that was within 0.02 of the best Pearson’s correlation or balanced accuracy for numeric and categorical features, respectively. Lastly, using the test-dataset, patients were resampled with replacement and 95% bootstrap BCa confidence intervals were calculated for performance metrics using the boot R package v1.3.28.1 (108,109).

Predicting time-until-BCR using multi-omic data

A total of 774 localized prostate cancer patients with time-until biochemical recurrence (BCR) were randomly split into a 70% train and 30% test dataset. Two machine learning models were compared for predicting time-until-BCR: random survival forest (110) and a regularized Cox proportional hazards model (105), with prediction accuracy assessed using the C-index (111). The random forest was fit using the randomForestSRC R package v3.2.3 (110) with 1000 trees and default settings, while the regularized Cox model was fit with the glmnet R package v4.1.8 (105) setting the alpha parameter to 0.5 (*i.e.* "elastic net" penalty that is half way between ridge and LASSO (106,107)) with the main penalty parameter lambda chosen by using 10-fold cross validation to maximize the C-index in the training data. Missing values were imputed by k-nearest neighbors imputation (impute R package v1.74.1, default settings) (89).

Several combinations of data types were considered for model training (**Supplementary Figure 17a**). Given the large number of potential features, feature selection was performed in the training dataset to identify a smaller subset of potential predictors. To perform feature selection, a separate Cox model was fit to each feature treating it as a predictor of time-until-BCR, while adjusting for the pre-treatment clinical features. Then only features with Wald-test p-value < α were retained. Several values for α were considered (0.001, 0.005, 0.01, 0.05, and 1) with the final value of α=0.005 chosen that maximized the random forest C-index in the training and test datasets (**Supplementary Figure 17d**).

The best machine learning model (highest C-index on test-dataset) was compared to the test C-index from a model that only considered pre-treatment clinical features. Using the test-dataset, patients were resampled with replacement and a 95% bootstrap BCa confidence interval was calculated for the difference in C-indices between the two models, using 50,000 bootstrap replications with the boot R package v1.3.28.1 (108,109). Similarly, a 95% bootstrap BCa CI was calculated for the test C-index of the best performing model. Lastly, the predicted BCR risk score from the best model was compared between 52 non-malignant TCGA samples, 161 localized prostate cancer samples from the test dataset, and 99 metastatic samples. Given that the best model included PSA and ISUP grade as predictors, and these were not measured for the non-malignant samples (ISUP grade was not measured for metastatic samples), their values were imputed using the randomForestSRC impute.rfsrc R function (default settings) (112).

Data Availability

ICGC methylation, mRNA abundance and sequencing data are available through NCBI GEO (GSE84043) and the European Genome-Phenome Archive (EGAS00001000900). TCGA clinical, methylation, mRNA and CNA data were downloaded using the TCGAbiolinks R package v2.25.3 (57). Kirby *et al*. 2017 clinical data was obtained through correspondence with the authors, and methylation data was from NCBI GEO (GSE76938). Zhao *et al.* 2017 clinical and methylation data was obtained from dbGaP study accession phs001921.v1.p1. Zhao *et al.* 2020 clinical data was obtained from https://github.com/DavidQuigley/WCDT/blob/master/clinical_metadata/2021_05_13_european_urology_2019.txt, methylation data was obtained through correspondence with the authors, CNA data from https://quigleylab.ucsf.edu/data, BAM files for whole-genome sequencing and RNA data from https://portal.gdc.cancer.gov/projects/WCDT-MCRPC. Li *et al*. 2020 data was downloaded from https://ngdc.cncb.ac.cn/methbank/samples?q=HRA000099. Melbourne methylation data was obtained through correspondence with co-authors Drs. Hovens and Jung. ICGC PRAD-FR methylation data was downloaded from https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-6131.

Software Availability

The PrCaMethy R package offers tools for predicting methylation subtype, clinical features and somatic driver mutations using prostate cancer DNA methylation data generated from the Illumina 450K array. The R package will be made publicly available upon manuscript publication.

Data Visualization

Data visualization was performed in the R statistical environment (v4.1.3) using the BPG (v7.0.8) package (26).

Acknowledgments

The authors thank all members of the Boutros lab, especially Veronica Y. Sabelnykova, for helpful suggestions. The results described here are in part based on data generated by the TCGA Research Network. We greatly appreciate the West Coast Dream Team consortium for sharing processed methylation data from metastatic prostate tumor biopsies, and the authors of Kirby *et al.* 2017 for sharing clinical data.

Author's Contributions

Pathology analyses: TvdK

Data processing: JA, JO, TG, YJS, AF, RL, TNY, RH-W, RMAD, RA, AS, CQY, SMGE, KEH, LEH

Statistical and machine-learning analyses: JA, AW, YJS, TNY, RMAD, KEH, FY

Wrote the first draft of the manuscript: JA, YJS, MF, PCB

Initiated the project: TvdK, RGB, MF, PCB

Supervised research: AK, AB, MLKC, MF, TvdK, RGB, PCB

Approved the manuscript: all authors

Funding

This study was conducted with the support of Movember funds through Prostate Cancer Canada and with the additional support of the Ontario Institute for Cancer Research, funded by the Government of Ontario. This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation - Grant #RS2014-01 to PCB. PCB was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. This research is funded by the Canadian Cancer Society (grant #705649) and by a CIHR Project Grant to PCB. RGB was supported by a Canadian Cancer Society Research Scientist Award and a Cancer Research UK core investigator grant (C5759). The research is funded in part by Canadian Epigenetics, Environment and Health Research Consortium grants to PCB. This work was supported by the NIH through awards P30CA016042, U01CA214194, U2CCA271894, R01CA270108; and NIH NIGMS grants T32GM008042 and T32GM152342. This work was supported by a Prostate Cancer Foundation Special Challenge Award to PCB (Award ID #: 20CHAS01) made possible by the generosity of Mr. Larry Ruvo. This work was supported by the DOD PCRP through awards W81XWH2210247 and W81XWH2210751. JO was supported by the UCLA Jonsson Comprehensive Cancer Center Fellowship.

Competing Interests

MLKC declared the following potential conflict of interest unrelated to the submitted work: personal fees from Astellas, Axiom, Bicara Therapeutics, IQVIA, Janssen, MSD, Seagen, Varian; non-financial support from Veracyte Inc; personal fees and grants from Bayer; personal

fees and grants from BeiGene; consults for Bicara Therapeutics; consults for ImmunoScape; consults for PVMed; consults for Varian Thought Leadership Council; and is a co-inventor of the patent of a High Sensitivity Lateral Flow Immunoassay For Detection of Analyte in Sample (10202107837T), Singapore, and serves on the Board of Directors of Digital Life Line Pte Ltd that owns the licensing agreement of the patent. PCB sits on the Scientific Advisory Boards of BioSymetrics Inc. and of Intersect Diagnostics Inc., and formerly sat on that of Sage Bionetworks.

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: a cancer journal for clinicians **2023**;73(1).
2. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T*, et al.* Familial risk and heritability of cancer among twins in Nordic countries. Jama **2016**;315(1):68-76.
3. Penney KL, Stampfer MJ, Jahn JL, Sinnott JA, Flavin R, Rider JR*, et al.* Gleason grade progression is uncommon. Cancer Res **2013**;73(16):5163-8 doi 10.1158/0008-5472.CAN-13-0427.
4. Fraser M, Berlin A, Bristow RG, Van der Kwast T. Genomic, pathological, and clinical heterogeneity as drivers of personalized medicine in prostate cancer. 2015. Elsevier. p 85-94.
5. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY*, et al.* The genomic complexity of primary human prostate cancer. Nature **2011**;470(7333):214-20.
6. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD*, et al.* The molecular taxonomy of primary prostate cancer. Cell **2015**;163(4):1011-25.
7. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V*, et al.* Genomic hallmarks of localized, non-indolent prostate cancer. Nature **2017** doi 10.1038/nature20788.
8. Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E*, et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat Genet **2015**;47(7):736-45 doi 10.1038/ng.3315.
9. Espiritu SMG, Liu LY, Rubanova Y, Bhandari V, Holgersen EM, Szyca LM*, et al.* The evolutionary landscape of localized prostate cancers drives clinical aggression. Cell **2018**;173(4):1003-13.
10. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C*, et al.* Gene expression correlates of clinical prostate cancer behavior. Cancer cell **2002**;1(2):203-9.
11. Chen S, Huang V, Xu X, Livingstone J, Soares F, Jeon J*, et al.* Widespread and functional RNA circularization in localized prostate cancer. Cell **2019**;176(4):831-43.
12. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C*, et al.* Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell reports **2014**;8(3):798-806.
13. Houlahan KE, Shiah Y-J, Gusev A, Yuan J, Ahmed M, Shetty A*, et al.* Genome-wide germline correlates of the epigenetic landscape of prostate cancer. Nature medicine **2019**;25(10):1615-26.
14. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M*, et al.* The DNA methylation landscape of advanced prostate cancer. Nature genetics **2020**;52(8):778-89.
15. Sheaffer KL, Elliott EN, Kaestner KH. DNA Hypomethylation Contributes to Genomic Instability and Intestinal Cancer Initiation. Cancer Prev Res (Phila) **2016**;9(7):534-46 doi 10.1158/1940-6207.CAPR-15-0349.
16. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet **2010**;70:27-56 doi 10.1016/B978-0-12-380866-0.60002-2.
17. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene **2002**;21(35):5427-40 doi 10.1038/sj.onc.1205600.
18. Massie CE, Mills IG, Lynch AG. The importance of DNA methylation in prostate cancer development. J Steroid Biochem Mol Biol **2017**;166:1-15 doi 10.1016/j.jsbmb.2016.04.009.
19. Miller BF, Sánchez-Vega F, Elnitski L. The emergence of pan-cancer CIMP and its elusive interpretation. Biomolecules **2016**;6(4):45.

20. Shin HJ, Hua JT, Li H. Recent advances in understanding DNA methylation of prostate cancer. Frontiers in Oncology **2023**;13:1182727.
21. Strand SH, Switnicki M, Moller M, Haldrup C, Storebjerg TM, Hedegaard J*, et al.* RHCG and TCAF1 promoter hypermethylation predicts biochemical recurrence in prostate cancer patients treated by radical prostatectomy. Oncotarget **2017**;8(4):5774-88 doi 10.18632/oncotarget.14391.
22. Zhao S, Geybels MS, Leonardson A, Rubicz R, Kolb S, Yan Q*, et al.* Epigenome-Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized Prostate Cancer. Clin Cancer Res **2017**;23(1):311-9 doi 10.1158/1078-0432.CCR-16-0549.
23. Vanaja DK, Ehrich M, Van den Boom D, Cheville JC, Karnes RJ, Tindall DJ*, et al.* Hypermethylation of genes for diagnosis and risk stratification of prostate cancer. Cancer investigation **2009**;27(5):549-60.
24. Bañez LL, Sun L, Van Leenders GJ, Wheeler TM, Bangma CH, Freedland SJ*, et al.* Multicenter clinical validation of PITX2 methylation as a prostate specific antigen recurrence predictor in patients with post-radical prostatectomy prostate cancer. The Journal of urology **2010**;184(1):149-56.
25. Ahmad AS, Vasiljević N, Carter P, Berney DM, Møller H, Foster CS*, et al.* A novel DNA methylation score accurately predicts death from prostate cancer in men with low to intermediate clinical risk factors. Oncotarget **2016**;7(44):71833.
26. Li J, Xu C, Lee HJ, Ren S, Zi X, Zhang Z*, et al.* A genomic and epigenomic atlas of prostate cancer in Asian populations. Nature **2020**;580(7801):93-9.
27. Kirby MK, Ramaker RC, Roberts BS, Lasseigne BN, Gunther DS, Burwell TC*, et al.* Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns. BMC cancer **2017**;17:1-10.
28. Kamoun A, Cancel-Tassin G, Fromont G, Elarouci N, Armenoult L, Ayadi M*, et al.* Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease. Annals of Oncology **2018**;29(8):1814-21.
29. Wang X, Jordahl KM, Zhu C, Livingstone J, Rhie SK, Wright JL*, et al.* Methylation subtypes of primary prostate cancer predict poor prognosis. Cancer Epidemiology, Biomarkers & Prevention **2022**;31(7):1473-82.
30. Zhang E, Shiori F, Mu OY, He J, Ge Y, Wu H*, et al.* Establishment of novel DNA methylation-based prostate cancer subtypes and a risk-predicting eight-gene signature. Frontiers in Cell and Developmental Biology **2021**;9:639615.
31. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS*, et al.* Integrative genomic profiling of human prostate cancer. Cancer cell **2010**;18(1):11-22.
32. Hieronymus H, Schultz N, Gopalan A, Carver BS, Chang MT, Xiao Y*, et al.* Copy number alteration burden predicts prostate cancer relapse. Proceedings of the National Academy of Sciences **2014**;111(30):11139-44.
33. Lalonde E, Ishkanian AS, Sykes J, Fraser M, Ross-Adams H, Erho N*, et al.* Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. The lancet oncology **2014**;15(13):1521-32.
34. Piggott JJ, Townsend CR, Matthaei CD. Reconceptualizing synergism and antagonism among multiple stressors. Ecology and evolution **2015**;5(7):1538-47.
35. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G*, et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature **2020**;578(7793):102-11.

36. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A*, et al.* Comprehensive characterization of cancer driver genes and mutations. Cell **2018**;173(2):371-85.
37. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E*, et al.* The long tail of oncogenic drivers in prostate cancer. Nature genetics **2018**;50(5):645-51.
38. Koh CM, Bieberich CJ, Dang CV, Nelson WG, Yegnasubramanian S, De Marzo AM. MYC and prostate cancer. Genes & cancer **2010**;1(6):617-28.
39. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell systems **2015**;1(6):417-25.
40. Bhandari V, Hoey C, Liu LY, Lalonde E, Ray J, Livingstone J*, et al.* Molecular landmarks of tumor hypoxia across cancer types. Nature genetics **2019**;51(2):308-18.
41. Yang L, Roberts D, Takhar M, Erho N, Bibby BAS, Thiruthaneeswaran N*, et al.* Development and validation of a 28-gene hypoxia-related prognostic signature for localized prostate cancer. EBioMedicine **2018**;31:182-9.
42. Cameron S, Deblois G, Hawley JR, Qamra A, Zhou S, Tonekaboni SAM*, et al.* Chronic hypoxia favours adoption to a castration-resistant cell state in prostate cancer. Oncogene **2023**;42(21):1693-703.
43. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. British journal of cancer **2010**;102(2):428-35.
44. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet **2013**;45(10):1127-33 doi 10.1038/ng.2762.
45. Geybels MS, Fang M, Wright JL, Qu X, Bibikova M, Klotzle B*, et al.* PTEN loss is associated with prostate cancer recurrence and alterations in tumor DNA methylation profiles. Oncotarget **2017**;8(48):84338-48 doi 10.18632/oncotarget.20940.
46. Kwon CH, Zhao D, Chen J, Alcantara S, Li Y, Burns DK*, et al.* Pten haploinsufficiency accelerates formation of high-grade astrocytomas. Cancer Res **2008**;68(9):3286-94 doi 10.1158/0008-5472.CAN-07-6867.
47. Hill R, Calvopina JH, Kim C, Wang Y, Dawson DW, Donahue TR*, et al.* PTEN loss accelerates KrasG12D-induced pancreatic cancer development. Cancer Res **2010**;70(18):7114-24 doi 10.1158/0008-5472.CAN-10-1649.
48. Kwabi-Addo B, Giri D, Schmidt K, Podsypanina K, Parsons R, Greenberg N*, et al.* Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. Proc Natl Acad Sci U S A **2001**;98(20):11563-8 doi 10.1073/pnas.201167798.
49. Chen YC, Gotea V, Margolin G, Elnitski L. Significant associations between driver gene mutations and DNA methylation alterations across many cancer types. PLoS Comput Biol **2017**;13(11):e1005840 doi 10.1371/journal.pcbi.1005840.
50. Kron KJ, Murison A, Zhou S, Huang V, Yamaguchi TN, Shiah YJ*, et al.* TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nat Genet **2017** doi 10.1038/ng.3930.
51. Valdes-Mora F, Gould CM, Colino-Sanguino Y, Qu W, Song JZ, Taylor KM*, et al.* Acetylated histone variant H2A.Z is involved in the activation of neo-enhancers in prostate cancer. Nat Commun **2017**;8(1):1346 doi 10.1038/s41467-017-01393-8.
52. Urbanucci A, Barfeld SJ, Kytola V, Itkonen HM, Coleman IM, Vodak D*, et al.* Androgen Receptor Deregulation Drives Bromodomain-Mediated Chromatin Alterations in Prostate Cancer. Cell Rep **2017**;19(10):2045-59 doi 10.1016/j.celrep.2017.05.049.
53. Teschendorff AE, Horvath S. Epigenetic ageing clocks: statistical methods and emerging

computational challenges. Nature Reviews Genetics **2025**:1-19.
54. Shiah YJ, Fraser M, Bristow RG, Boutros PC. Comparison of Pre-processing Methods for Infinium HumanMethylation450 BeadChip Array. Bioinformatics **2017** doi 10.1093/bioinformatics/btx372.
55. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics **2013**;14:293 doi 10.1186/1471-2164-14-293.
56. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun **2015**;6:8971 doi 10.1038/ncomms9971.
57. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D*, et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research **2016**;44(8):e71-e.
58. TCGA. 10/26/2023. Methylation analysis pipeline. <https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_Pipeline/>. 10/26/2023.
59. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD*, et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics **2014**;30(10):1363-9.
60. Benelli M, Franceschini GM, Magi A, Romagnoli D, Biagioni C, Migliaccio I*, et al.* Charting differentially methylated regions in cancer with Rocker-meth. Communications Biology **2021**;4(1):1249.
61. Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ*, et al.* Genomic hallmarks and structural variation in metastatic prostate cancer. Cell **2018**;174(3):758-69.
62. TCGA. 10/26/2023. Copy Number Variation Analysis Pipeline. <https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/>. 10/26/2023.
63. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **2004**;20(3):307-15 doi 10.1093/bioinformatics/btg405.
64. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG*, et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res **2005**;33(20):e175 doi 10.1093/nar/gni179.
65. TCGA. 10/26/2023. mRNA Analysis Pipeline. <https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/>. 10/26/2023.
66. Govind SK, Zia A, Hennings-Yeomans PH, Watson JD, Fraser M, Anghel C*, et al.* ShatterProof: operational detection and quantification of chromothripsis. BMC Bioinformatics **2014**;15:78 doi 10.1186/1471-2105-15-78.
67. Yousif F, Prokopec S, Sun RX, Fan F, Lalansingh CM, Park DH*, et al.* The Origins and Consequences of Localized and Global Somatic Hypermutation. bioRxiv **2018**.
68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N*, et al.* The sequence alignment/map format and SAMtools. bioinformatics **2009**;25(16):2078-9.
69. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 **2013**.
70. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **2010**;20(9):1297-303.
71. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and indels with Mutect2. BioRxiv **2019**:861054.

72. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ*, et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics **2012**;28(3):311-7.
73. Ji S, Montierth MD, Wang W. MuSE: A novel approach to mutation calling with sample-specific error modeling. Variant Calling: Methods and Protocols: Springer; 2022. p 21-7.
74. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M*, et al.* Strelka2: fast and accurate calling of germline and somatic variants. Nature methods **2018**;15(8):591-4.
75. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO*, et al.* Twelve years of SAMtools and BCFtools. Gigascience 10: giab008. 2021.
76. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM*, et al.* dbSNP: the NCBI database of genetic variation. Nucleic acids research **2001**;29(1):308-11.
77. Ouedraogo M, Bettembourg C, Bretaudeau A, Sallou O, Diot C, Demeure O*, et al.* The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. PloS one **2012**;7(11):e50653.
78. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C*, et al.* Architecture of the human regulatory network derived from ENCODE data. Nature **2012**;489(7414):91-100.
79. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q*, et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature **2020**;581(7809):434-43.
80. Fuentes Fajardo KV, Adams D, Program NCS, Mason CE, Sincan M, Tifft C*, et al.* Detecting false-positive signals in exome sequencing. Human mutation **2012**;33(4):609-13.
81. Genomes Project C. A global reference for human genetic variation. Nature **2015**;526(7571):68.
82. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H*, et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research **2015**;43(D1):D805-D11.
83. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M*, et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics **2016**;32(8):1220-2.
84. Cameron DL, Dong R, Papenfuss AT. StructuralVariantAnnotation: a R/Bioconductor foundation for a caller-agnostic structural variant software ecosystem. Bioinformatics **2022**;38(7):2046-8.
85. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **1995**;57(1):289-300.
86. Gan L, Allen GI. Fast and interpretable consensus clustering via minipatch learning. PLoS Computational Biology **2022**;18(10):e1010577.
87. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **1987**;20:53-65.
88. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A **2002**;99(10):6567-72 doi 10.1073/pnas.082099299.
89. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R*, et al.* Missing value estimation methods for DNA microarrays. Bioinformatics **2001**;17(6):520-5.
90. Rappoport N, Shamir R. Inaccuracy of the log-rank approximation in cancer data analysis. Molecular systems biology **2019**;15(8):e8754.

91. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. **2014**.
92. Illumina. 2016 Field Guide to Methylation Methods. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/field_guide_methylation.pdf>.
93. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC bioinformatics **2013**;14:1-15.
94. Venables WN, Ripley BD. Modern applied statistics with S-PLUS. Springer Science & Business Media; 2013.
95. Ish-Horowicz J, Reid J. Mutual information estimation for transcriptional regulatory network inference. BioRxiv **2017**:132647.
96. Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research **2009**;10(7).
97. James W, Stein C. Estimation with quadratic loss. Breakthroughs in statistics: Foundations and basic theory: Springer; 1992. p 443-60.
98. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **2010**;33(1):1.
99. Breiman L. Random forests. Machine learning **2001**;45:5-32.
100. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research **2014**;15(1):3133-81.
101. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology **2019**;110:12-22.
102. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods **2009**;14(4):323.
103. Liaw A. Classification and regression by randomForest. R news **2002**.
104. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software **2008**;28:1-26.
105. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software **2011**;39(5):1.
106. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology **1996**;58(1):267-88.
107. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology **2005**;67(2):301-20.
108. Canty AJ. Resampling methods in R: the boot package. The Newsletter of the R Project Volume **2002**;2(3):2-7.
109. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge university press; 1997.
110. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. **2008**.
111. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine **1996**;15(4):361-87.
112. Tang F, Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal **2017**;10(6):363-77.

Figure Legends

Figure 1: The epigenomic landscape of prostate cancer

a. Subtype methylation profiles of 1,824 prostate tumours and 325 normal samples from seven cohorts, including clinical pre-treatment variables. The heatmap includes 5,486 CpGs which were selected by IMPACC as important for deriving the subtypes. CpGs were sorted using hierarchical clustering with complete linkage and Euclidean distance, then split into five equally sized groups for visualization. Missing values are colored in white and clinical variables are not shown for normal samples (other than age).

b. The percent of patients within each subtype is shown for seven cohorts of localized tumour samples (left), two cohorts of metastatic samples (middle-left), four cohorts of normal samples (middle-right), and when pooling all samples together (right). The total sample size within each cohort is given above the bars. L: localized, M: metastatic, N: normal sample.

c. Kaplan–Meier curves were used to estimate BCR-free probabilities separately in each subtype, pooling all patients with available BCR data (n = 774). The p-value in the top-right corner is from a log-rank test. Hazard ratios with 95% confidence intervals and Wald-test p-values were estimated using a Cox proportional hazard model treating MS-1 as the reference group. The table below the plot gives the number of patients at risk from 0-5 years.

d. The dot-map summarizes the overall association of subtypes with four clinical features: age, tumour category (T), ISUP grade group and PSA when pooling all patients together (Overall) and separately by cohort. Each of the clinical variables was categorized as in **Supplementary Table 1** and tested for an association with the four methylation subtypes using Fisher's Exact Test. The size of the dots reflects the bias-corrected Cramer's V effect size and the background color represents the q-value, which adjusts for multiple testing separately for each clinical variable. An x within a cell means that feature was unavailable for that cohort.

e-g. Fisher's exact test and odds ratio were used to compare the rates of subtype-defining and non-subtype-defining "background" CpGs within each chromosome (**e**), within different types of CpGs (island, open sea, shelf, and shore (92)) (**f**) and within gene body, intergenic and gene promoter regions (as defined in Methods) (**g**).

Figure 2: The regulatory landscape of prostate cancer

a. The y-axis indicates the percent of patients with copy number gain (red bar) or loss (blue bar) within a gene at that location of the genome (x-axis). The vertical magenta bars indicate genes later shown in (**f**).

b-d. Spearman's correlation of methylation and CNA (**b**), methylation and mRNA (**c**), and CNA and mRNA (**d**) within each gene. Positive correlations are shown in blue and negative correlations in orange. Gene-level methylation was measured for each patient by the median beta-value among CpG islands in the gene promoter region.

e. Linear regression interaction coefficient of binary methylation (hyper *vs.* hypomethylation; median dichotomized) and binary CNA status (loss *vs.* neutral if loss was the more common mutation, otherwise gain *vs.* neutral). The dependent variable is RNA abundance (z-score transformed $\log_2$ TPM). Genes with significant interaction effects ($q < 0.05$) are shown as well as gene BCL2L10 ($q = 0.055$).

f. Examples of methylation-CNA interactions on mRNA abundance for six genes. The q-value for testing the interaction effect is indicated by q_{int}.

g. For each gene, the more common "major" mutation was identified (CNA loss or gain). Then the median methylation in the gene promoter region was compared between the major mutation and neutral copy number ($\log_2$ fold change; y-axis). Results were compared by tumour type and for known cancer driver genes *vs.* other genes (x-axis).

h-j. The distributions of Spearman's correlation for gene-level methylation, RNA and CNA in localized and metastatic samples.

Figure 3: Methylation subtypes have different CNA and RNA profiles

a. Copy number alteration of 1653 genes containing the subtype-defining CpGs. Each gene (row of heatmap) was summarized by the percent of patients with copy number gain and loss within each subtype (columns), which was discretized to >30% for visualization. Rows were sorted by hierarchical clustering with Euclidean distance and complete linkage. Ordinal logistic regression was used to model a gene's copy number status (loss, neutral, gain) as the dependent variable, with subtype, PGA and cohort as the independent variables. A likelihood ratio test p-value was calculated for subtype, and multiple testing-adjusted q-values are displayed in the right barplot.

b. The y-axis indicates the percent of patients with copy number gain (red bar) or loss (blue bar) at that location of the genome (x-axis). Results are stratified by methylation subtype (row).

c. Percent genome alteration (PGA) was compared between the methylation subtypes with Kruskal-Wallis rank sum test p-value. Groups sharing the same letter do not significantly differ (Mann-Whitney U-test $q > 0.05$), while groups with different letters have $q < 0.05$.

d. For 3,516 mRNA transcripts (adjusted for PGA and cohort, see Methods) that significantly differed between subtypes (Kruskal Wallis rank-sum test $q<0.05$ and standard deviation (SD) of subtype-medians >0.4), the subtype medians were ranked from 1-4 (smallest to largest). Row-wise percentages show the percent of time a given subtype had each rank.

e. Patient-specific pathway scores were calculated for 50 hallmark pathways using GSVA (39,93) adjusted for PGA and cohort (Methods). For each pathway, the subtype median adjusted mRNA abundance z-score is displayed in the heatmap. Kruskal Wallis test q-values are shown in the right barplot, and pathways were sorted in descending order by the standard deviation of the subtype medians. The top 20 most significant pathways are shown.

Figure 4: Somatic driver mutations impact the prostate cancer epigenome; and epigenetic signatures of clinico-molecular features

a. CpG β-values were tested for an association with 28 somatic events and clinical features. Somatic features were divided into five categories: mutational burden (mutation density), CNAs, INVs, SNVs, and CTXs including ETS fusions. Barplot on the left shows different q-value category percentages, resulting from Spearman's ρ and Kruskal-Wallis tests for continuous and discrete features, respectively. Barplot on the right shows the number of significantly differentially methylated CpGs associated with each feature ($q < 0.05$), with methylated *vs.* unmethylated status (median β-value threshold = 0.61).

b. Methylation profiles of the top 4 driver events with the most number of significant CpG associations. For each driver event, rows were sorted using hierarchical clustering with the top 100 significant CpGs ($q < 0.05$), complete linkage and Euclidean distance. PGA, SNV count and number of genomic rearrangements were median-dichotomized into high *vs.* low groups.

c. Logistic regression was used to test the association between driver events and subtypes while adjusting for cohort and PGA. The odds of having a given mutation (row) are shown for each subtype (column) when treating MS-2 as the reference group. PGA and SNVs/Mbp were median-dichotomized; 12 drivers with $q<0.05$ are shown.

d. Boxplots of the total number of driver events by methylation subtype, with Kruskal-Wallis p-value. Pairwise comparisons between groups were conducted using the Mann-Whitney U-test, such that groups with different letters significantly differ ($q < 0.05$).

e. Performance of the final selected models for predicting 15 features (y-axis labels). Dots indicate the estimated performance on the test dataset with adjacent lines giving 95% bootstrap confidence intervals. Pearson's correlation (r) and balanced accuracy (BA) were used for numeric and categorical features, respectively.

f. Test dataset performance for predicting $\log_2$ SNVs/Mbp. Pearson's r is reported with 95% bootstrap confidence interval in parentheses.

g. Test dataset confusion matrix for predicting MYC copy number gain. Balanced accuracy (BA) is reported with 95% bootstrap confidence interval in parentheses. Each cell is shaded by the percentage of the row total.

h. A random forest (RF, blue) and regularized Cox proportional hazards model (rCoxPH, red) were trained for predicting time-until-BCR using different combinations of data as predictors (x-axis labels; see **i** for definitions). The height of the bars indicates the estimated C-index in the training and test datasets. Higher C-index indicates better performance.

i. The combinations of predictors used in figure **h**. For a given row, the y-axis label names the data combination and the boxes indicate which data types are included.

j. The best model from **h**. was used to predict BCR risk score in the test dataset of 161 patients and median-dichotomized into High *vs.* Low risk. The Kaplan-Meir plot shows the BCR-free probabilities for the High and Low risk groups, with hazard ratio, 95%

confidence interval, and Wald-test p-value estimated from a Cox proportional hazards model. The table below the plot gives the number of patients at risk from 0-5 years.

k. The best model from **h**. was used to predict the BCR risk score for 52 normal samples, 161 localized prostate cancer samples from the test dataset, and 99 metastatic samples. The risk score was compared between the three tissue groups using Kruskal-Wallis rank sum test p-value. Groups with different letters have significantly different risk score profiles (pairwise Wilcoxon test $q < 0.05$).

Figure 1

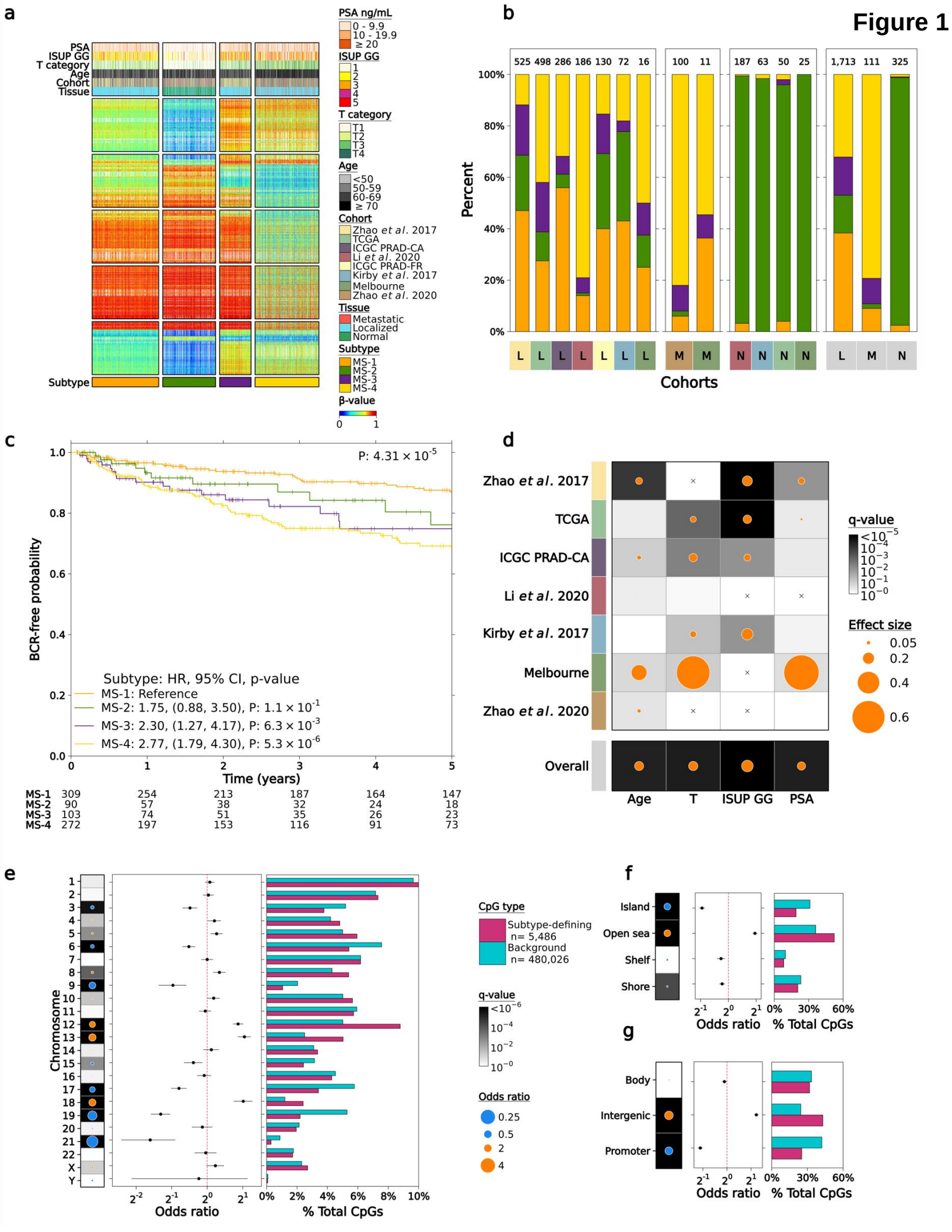

a
PSA
ISUP GG
T category
Age
Cohort
Tissue
Subtype
PSA ng/mL
0 - 9.9
10 - 19.9
≥ 20
ISUP GG
1
2
3
4
5
T category
T1
T2
T3
T4
Age
<50
50-59
60-69
≥ 70
Cohort
Zhao et al. 2017
TCGA
ICGC PRAD-CA
Li et al. 2020
ICGC PRAD-FR
Kirby et al. 2017
Melbourne
Zhao et al. 2020
Tissue
Metastatic
Localized
Normal
Subtype
MS-1
MS-2
MS-3
MS-4
β-value
0
1
b
525 498 286 186 130 72 16
100 11
187 63 50 25
1,713 111 325
100%
80%
60%
40%
20%
0%
Percent
L L L L L L L M M N N N N L M N
Cohorts
c
P: 4.31 × 10^-5
BCR-free probability
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5
Time (years)
Subtype: HR, 95% CI, p-value
MS-1: Reference
MS-2: 1.75, (0.88, 3.50), P: 1.1 × 10^-1
MS-3: 2.30, (1.27, 4.17), P: 6.3 × 10^-3
MS-4: 2.77, (1.79, 4.30), P: 5.3 × 10^-6
MS-1 309 254 213 187 164 147
MS-2 90 57 38 32 24 18
MS-3 103 74 51 35 26 23
MS-4 272 197 153 116 91 73
d
Zhao et al. 2017
TCGA
ICGC PRAD-CA
Li et al. 2020
Kirby et al. 2017
Melbourne
Zhao et al. 2020
Overall
Age
T
ISUP GG
PSA
q-value
<10^-5
10^-4
10^-3
10^-2
10^-1
10^-0
Effect size
0.05
0.2
0.4
0.6
e
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
2^-2 2^-1 2^0 2^1
Odds ratio
0% 2% 4% 6% 8% 10%
% Total CpGs
CpG type
Subtype-defining
n= 5,486
Background
n= 480,026
q-value
<10^-6
10^-4
10^-2
10^-0
Odds ratio
0.25
0.5
2
4
f
Island
Open sea
Shelf
Shore
2^-1 2^0 2^1
Odds ratio
0% 30% 60%
% Total CpGs
g
Body
Intergenic
Promoter
2^-1 2^0 2^1
Odds ratio
0% 30% 60%
% Total CpGs

Figure 2

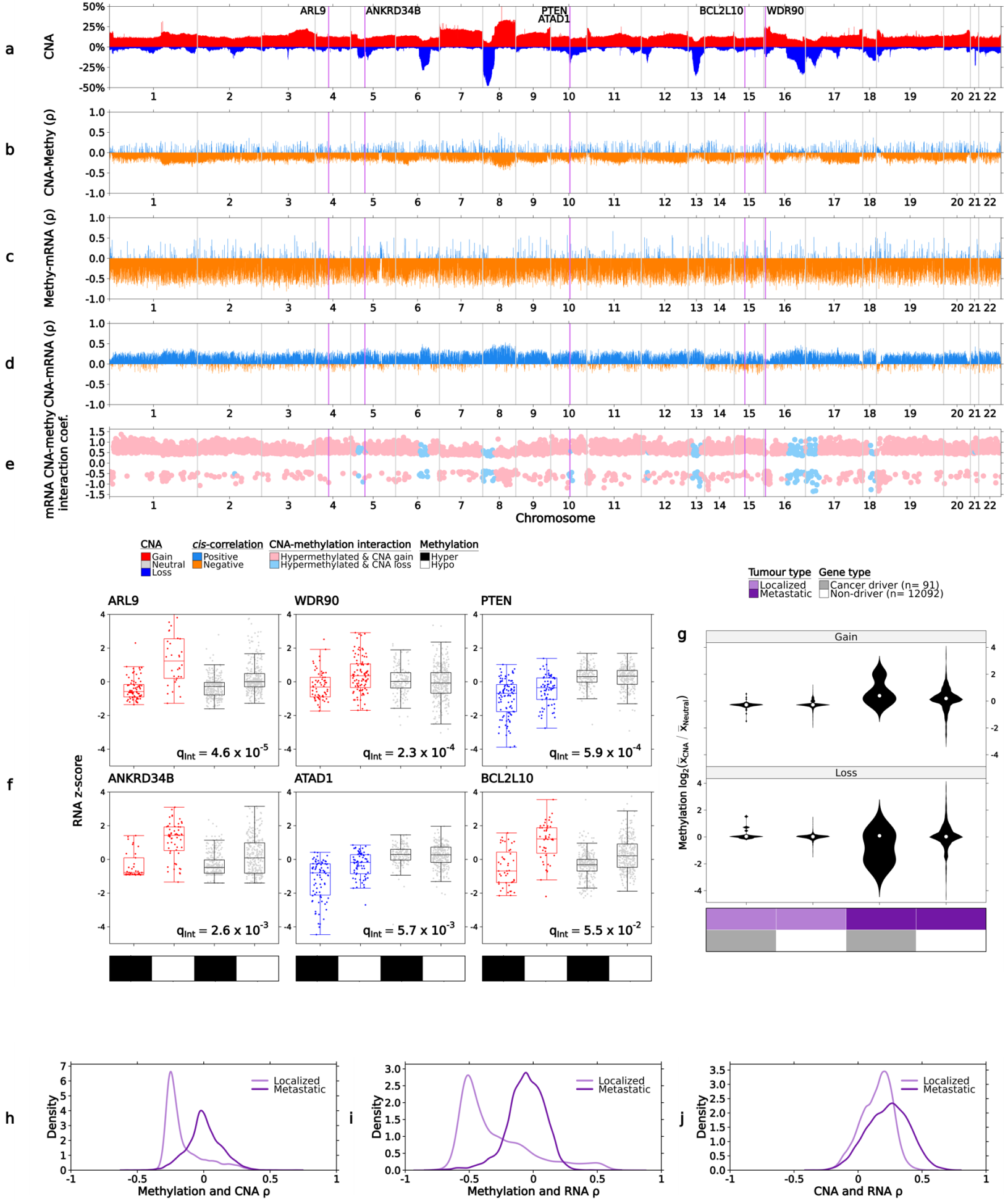

a
CNA
ARL9
ANKRD34B
PTEN
ATAD1
BCL2L10
WDR90
b
CNA-Methy (ρ)
c
Methy-mRNA (ρ)
d
CNA-mRNA (ρ)
e
CNA-methy interaction coef.
mRNA
Chromosome
CNA
Gain
Neutral
Loss
cis-correlation
Positive
Negative
CNA-methylation interaction
Hypermethylated & CNA gain
Hypermethylated & CNA loss
Methylation
Hyper
Hypo
f
RNA z-score
ARL9
WDR90
PTEN
ANKRD34B
ATAD1
BCL2L10
$q_{Int} = 4.6 \times 10^{-5}$
$q_{Int} = 2.3 \times 10^{-4}$
$q_{Int} = 5.9 \times 10^{-4}$
$q_{Int} = 2.6 \times 10^{-3}$
$q_{Int} = 5.7 \times 10^{-3}$
$q_{Int} = 5.5 \times 10^{-2}$
Tumour type
Localized
Metastatic
Gene type
Cancer driver (n= 91)
Non-driver (n= 12092)
g
Gain
Loss
Methylation $\log_2(\bar{x}_{CNA} / \bar{x}_{Neutral})$
h
Density
Localized
Metastatic
Methylation and CNA ρ
i
Methylation and RNA ρ
j
CNA and RNA ρ

Figure 3

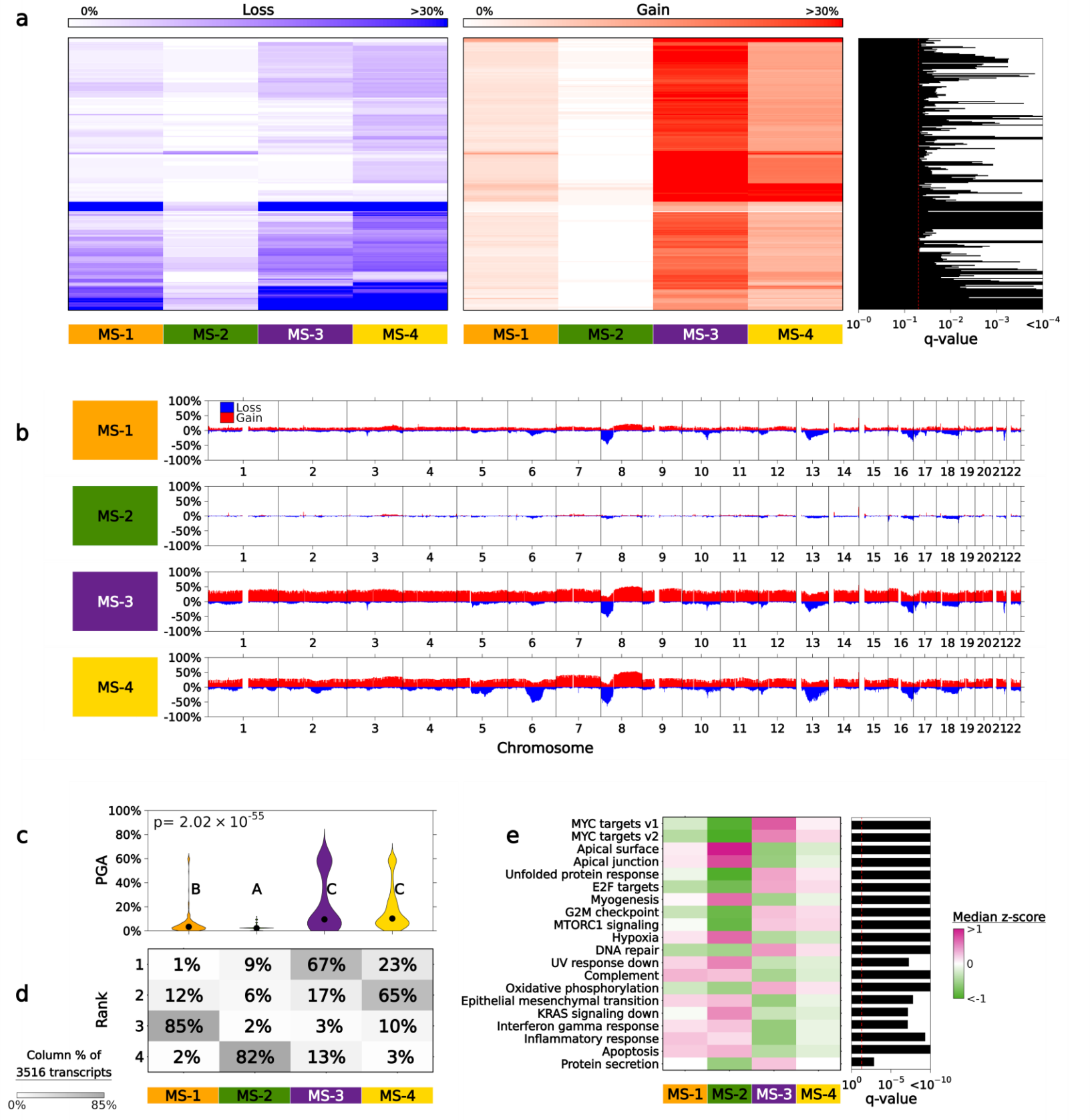
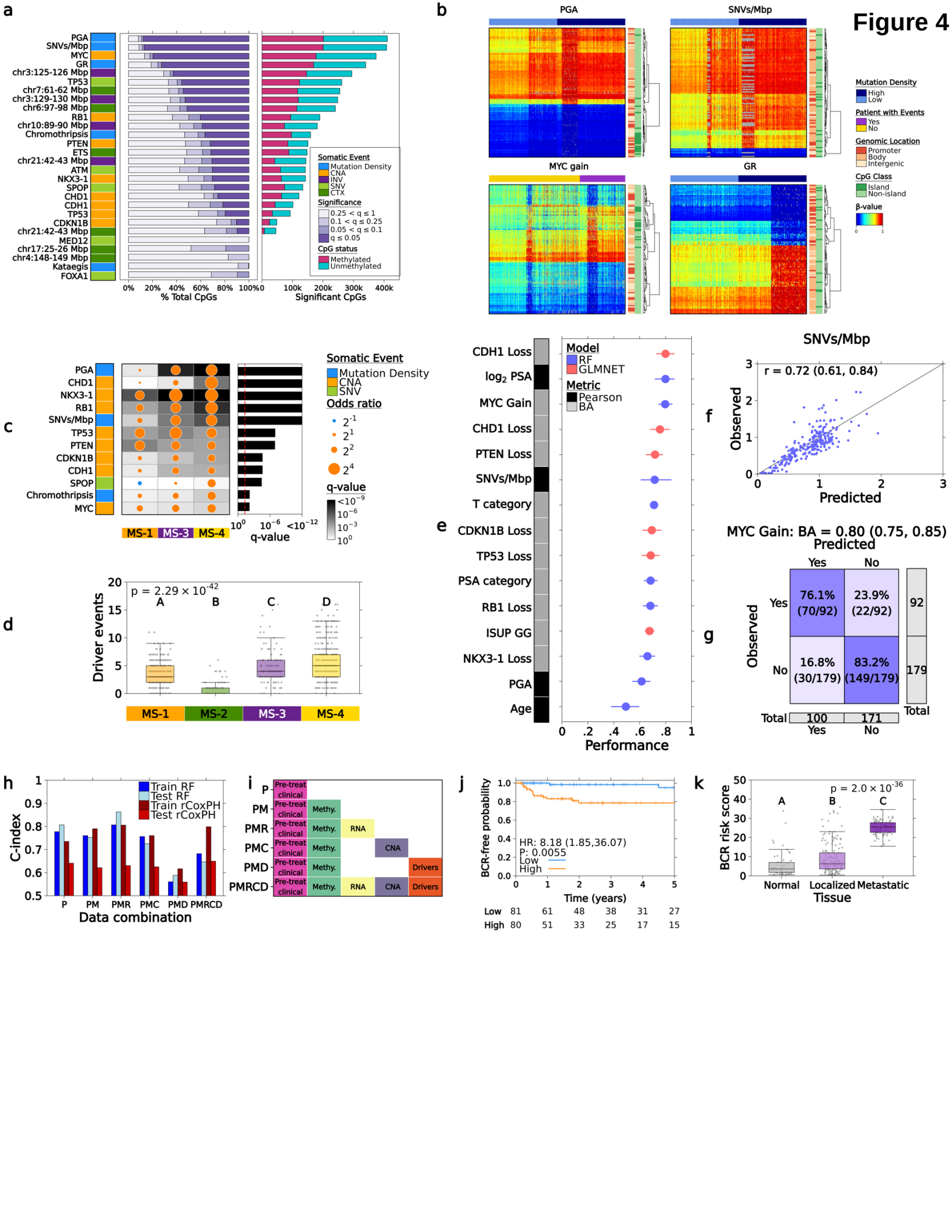
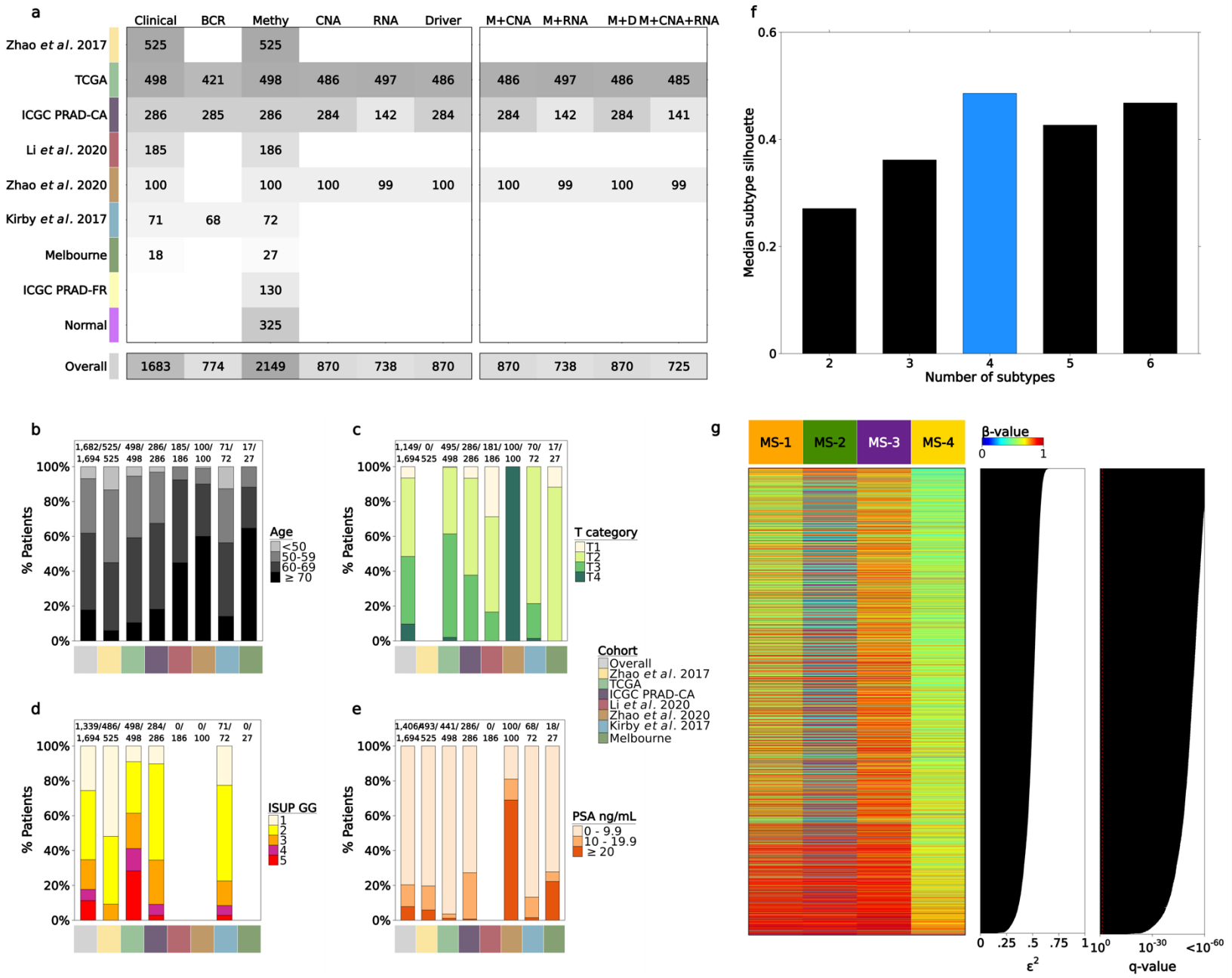

a
0%
Loss
>30%
0%
Gain
>30%
MS-1
MS-2
MS-3
MS-4
MS-1
MS-2
MS-3
MS-4
10^{-0}
10^{-1}
10^{-2}
10^{-3}
$<10^{-4}$
q-value
b
Loss
Gain
MS-1
MS-2
MS-3
MS-4
100%
50%
0%
-50%
-100%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
c
$p= 2.02 \times 10^{-55}$
PGA
100%
80%
60%
40%
20%
0%
B
A
C
C
d
Rank
1 1% 9% 67% 23%
2 12% 6% 17% 65%
3 85% 2% 3% 10%
4 2% 82% 13% 3%
MS-1
MS-2
MS-3
MS-4
Column % of
3516 transcripts
0%
85%
e
MYC targets v1
MYC targets v2
Apical surface
Apical junction
Unfolded protein response
E2F targets
Myogenesis
G2M checkpoint
MTORC1 signaling
Hypoxia
DNA repair
UV response down
Complement
Oxidative phosphorylation
Epithelial mesenchymal transition
KRAS signaling down
Interferon gamma response
Inflammatory response
Apoptosis
Protein secretion
MS-1
MS-2
MS-3
MS-4
10^{0}
10^{-5}
$<10^{-10}$
q-value
Median z-score
>1
0
<-1

Figure 4


a
PGA
SNVs/Mbp
MYC
GR
chr3:125-126 Mbp
TP53
chr7:61-62 Mbp
chr3:129-130 Mbp
chr6:97-98 Mbp
RB1
chr10:89-90 Mbp
Chromothripsis
PTEN
ETS
chr21:42-43 Mbp
ATM
NKX3-1
SPOP
CHD1
CDH1
TP53
CDKN1B
chr21:42-43 Mbp
MED12
chr17:25-26 Mbp
chr4:148-149 Mbp
Kataegis
FOXA1
0% 20% 40% 60% 80% 100%
% Total CpGs
0 100k 200k 300k 400k
Significant CpGs
Somatic Event
Mutation Density
CNA
INV
SNV
CTX
Significance
0.25 < q ≤ 1
0.1 < q ≤ 0.25
0.05 < q ≤ 0.1
q ≤ 0.05
CpG status
Methylated
Unmethylated
b
PGA
SNVs/Mbp
MYC gain
GR
Mutation Density
High
Low
Patient with Events
Yes
No
Genomic Location
Promoter
Body
Intergenic
CpG Class
Island
Non-island
β-value
0
1
c
PGA
CHD1
NKX3-1
RB1
SNVs/Mbp
TP53
PTEN
CDKN1B
CDH1
SPOP
Chromothripsis
MYC
MS-1
MS-3
MS-4
Somatic Event
Mutation Density
CNA
SNV
Odds ratio
2^{-1}
2^{1}
2^{2}
2^{4}
q-value
$<10^{-9}$
10^{-6}
10^{-3}
10^{0}
10^{0} 10^{-6} $<10^{-12}$
q-value
d
$p = 2.29 \times 10^{-42}$
A B C D
Driver events
0 5 10 15 20
MS-1 MS-2 MS-3 MS-4
e
CDH1 Loss
log_2 PSA
MYC Gain
CHD1 Loss
PTEN Loss
SNVs/Mbp
T category
CDKN1B Loss
TP53 Loss
PSA category
RB1 Loss
ISUP GG
NKX3-1 Loss
PGA
Age
Model
RF
GLMNET
Metric
Pearson
BA
0 .2 .4 .6 .8 1
Performance
f
SNVs/Mbp
r = 0.72 (0.61, 0.84)
Observed
Predicted
0 1 2 3
g
MYC Gain: BA = 0.80 (0.75, 0.85)
Predicted
Yes No
Observed
Yes
No
76.1% (70/92)
23.9% (22/92)
16.8% (30/179)
83.2% (149/179)
92
179
Total
100 171
Yes No
Total
h
Train RF
Test RF
Train rCoxPH
Test rCoxPH
C-index
0.5 0.6 0.7 0.8 0.9 1
P PM PMR PMC PMD PMRCD
Data combination
i
P
PM
PMR
PMC
PMD
PMRCD
Pre-treat clinical
Methy.
RNA
CNA
Drivers
j
BCR-free probability
0.0 0.2 0.4 0.6 0.8 1.0
HR: 8.18 (1.85,36.07)
P: 0.0055
Low
High
0 1 2 3 4 5
Time (years)
Low 81 61 48 38 31 27
High 80 51 33 25 17 15
k
$p = 2.0 \times 10^{-36}$
A B C
BCR risk score
0 10 20 30 40 50
Normal
Localized
Metastatic
Tissue



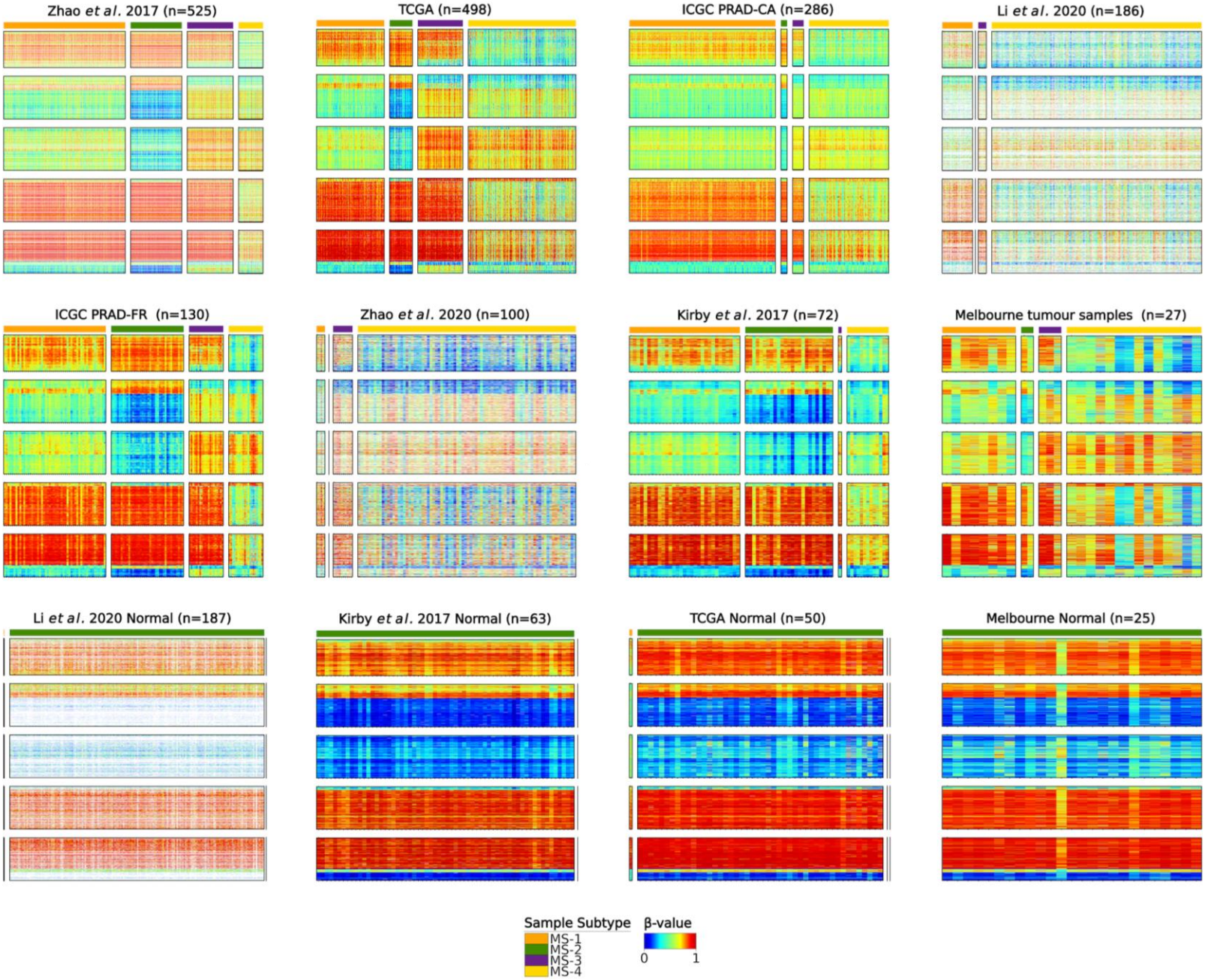
Supplementary Figure 1: Sample size, pre-treatment clinical features and subtype-defining CpGs

a. Sample size by cohort (row) and data type (column). Columns: Clinical stands for pre-treatment clinical features; BCR is time-until biochemical reoccurrence; Driver is the panel of 28 driver mutations; M = methylation, D = driver.

b-e. Stacked barplots showing the percent of patients in each category (y-axis) stratified by cohort (x-axis). T category: tumour category; ISUP GG: International Society of Urologic Pathologists grade group; PSA: prostate-specific antigen.

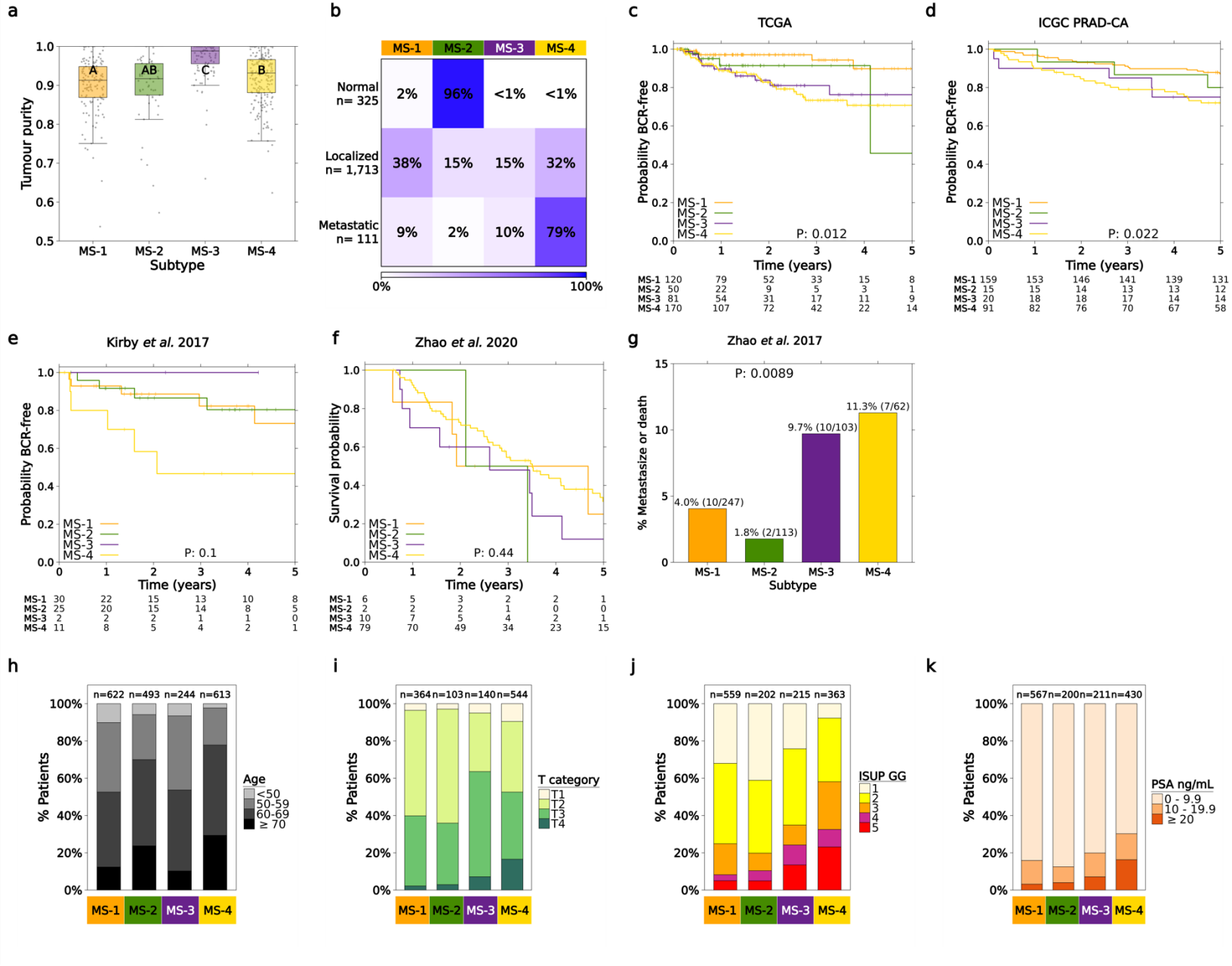
f. The silhouette method was used to choose the number of subtypes that maximized the median silhouette score, resulting in 4 subtypes (blue).

g. Each of 5,486 subtype-defining CpGs (rows) is summarized by the median methylation β-value within each subtype (column). Kruskal-Wallis test was used to compare each CpG between the subtypes, with q-values displayed in the right barplot and epsilon-squared ε^2 effect size in the middle barplot.



Supplementary Figure 2: Subtype methylation profiles by cohort, tissue type

Subtype tumour-methylation profiles stratified by cohort (top two rows), with profiles of normal samples in the bottom row. Patient-wise consensus clustering was performed using IMPACC to derive four subtypes within the TCGA cohort (n=498). The heatmaps include 5,486 CpGs which were selected by IMPACC as important for deriving the subtypes. A nearest shrunken centroid classifier was used to classify the remaining 1,326 tumours and 325 normal samples to their respective subtypes identified in the TCGA cohort. CpGs were sorted using hierarchical clustering with complete linkage and Euclidean distance in the TCGA cohort, then split into five equally sized groups for easier visualization. Missing values are colored in white.



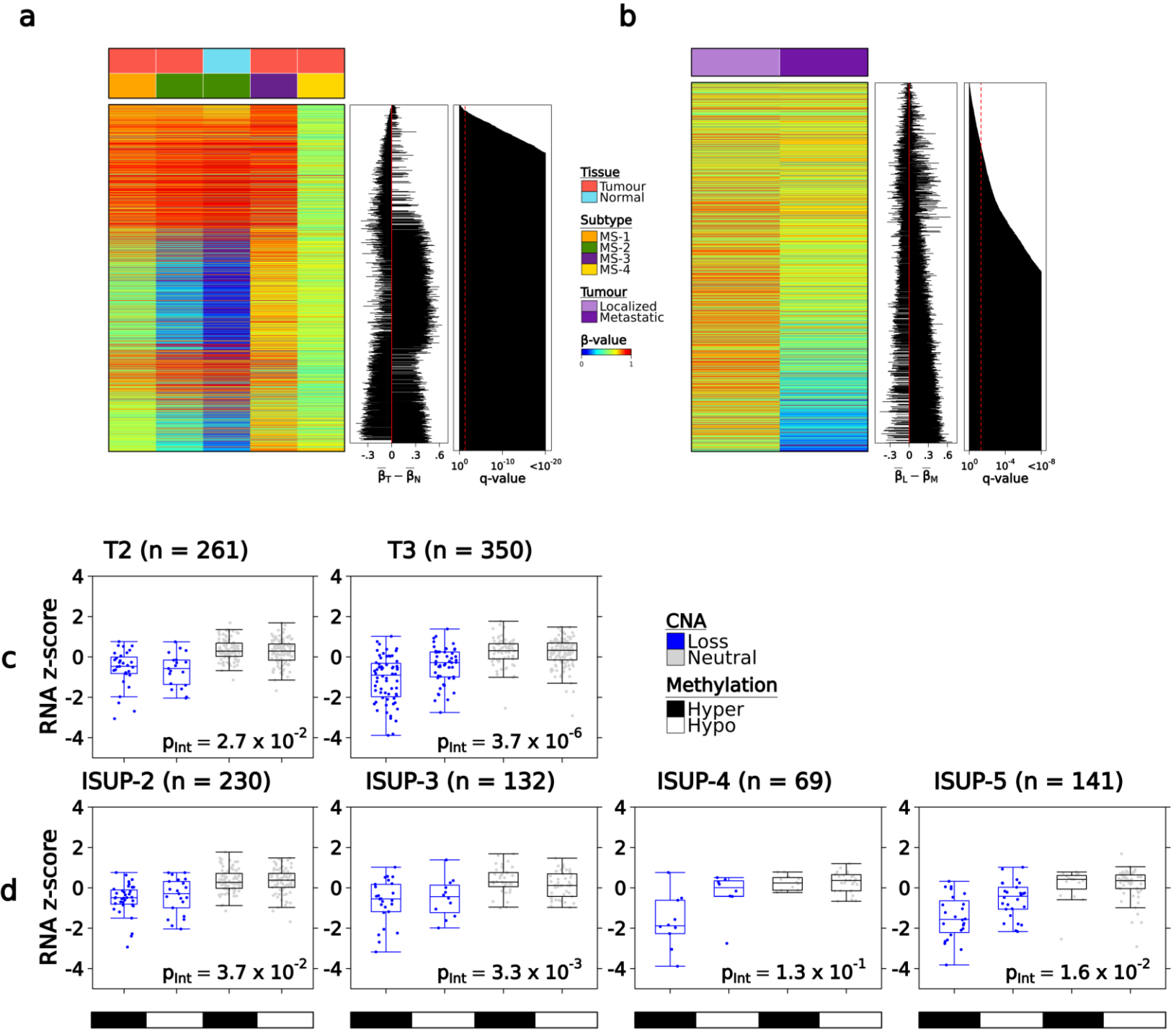
Supplementary Figure 3: Methylation subtypes *vs.* tumour purity and clinical features

a. Tumour purity was compared between subtypes within the TCGA derivation cohort. Groups sharing the same letter do not significantly differ, U-test q > 0.05.

b. Within each tissue type (row), the percent of samples assigned to each subtype (column) is shown.

c-g. The subtypes were tested for association with time until biochemical recurrence (BCR) using a logrank test in cohorts TCGA (**c**), ICGC PRAD-CA (**d**) and a permutation logrank test in Kirby *et al.* 2017 (**e**), due to small/uneven group sample sizes (10,000 permutations). The permutation logrank test was used for time until death in Zhao *et al.* 2020 (**f**), while Fisher's Exact Test was used in Zhao *et al.* 2017 for a binary outcome of metastasize or death during followup (**g**).

h-k. Stacked barplots showing % of patients in each category (y-axis) stratified by methylation subtype (x-axis). T category: tumour category; ISUP GG: International Society of Urologic Pathologists grade group; PSA: prostate-specific antigen.

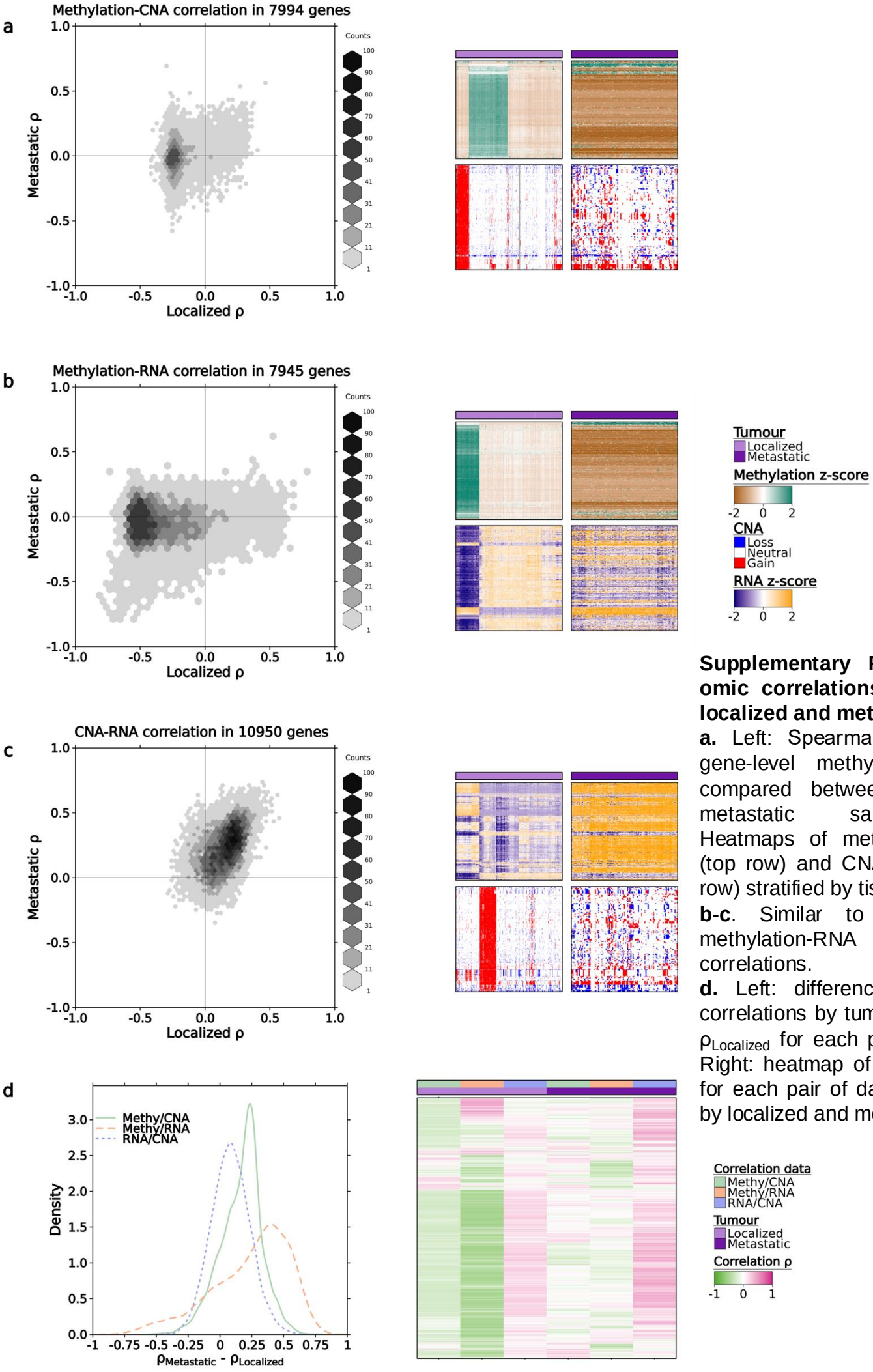


Supplementary Figure 4: Subtype defining CpGs *vs.* tissue type; PTEN interaction sensitivity analysis

a. Tumour *vs.* normal samples: Heatmap of 5,486 subtype-defining CpGs, sorted by $|\bar{\beta}_T - \bar{\beta}_N|$, where $\bar{\beta}_T$ is the median methylation beta-value in tumour samples, and $\bar{\beta}_N$ is the median beta-value in normal samples. The Mann–Whitney U test was used to test for differences between the tumour and normal distributions of each CpG, with q-values displayed in the right barplot.

b. Localized *vs.* metastatic samples: Similar to **a.**, except CpGs sorted by $|\bar{\beta}_L - \bar{\beta}_M|$, where $\bar{\beta}_L$ is the median methylation beta-value in localized prostate cancers, and $\bar{\beta}_M$ is the median beta-value in metastatic prostate cancers.

c,d. A sensitivity analysis was performed for the PTEN methylation-CNA interaction on mRNA abundance (from Figure 2e,f) to determine whether the same association held across tumour categories (**c**) and ISUP grade groups (**d**). Linear regression was used to test the interaction between binary methylation (hyper vs. hypomethylation; median dichotomized) and binary CNA status (loss vs. neutral), with the interaction p-value given by p_{Int}. The dependent variable is RNA abundance (z-score transformed $\log_2$ TPM). Boxplots are only shown for tumour categories and ISUP grades where each of the four groups had sample size greater than five.

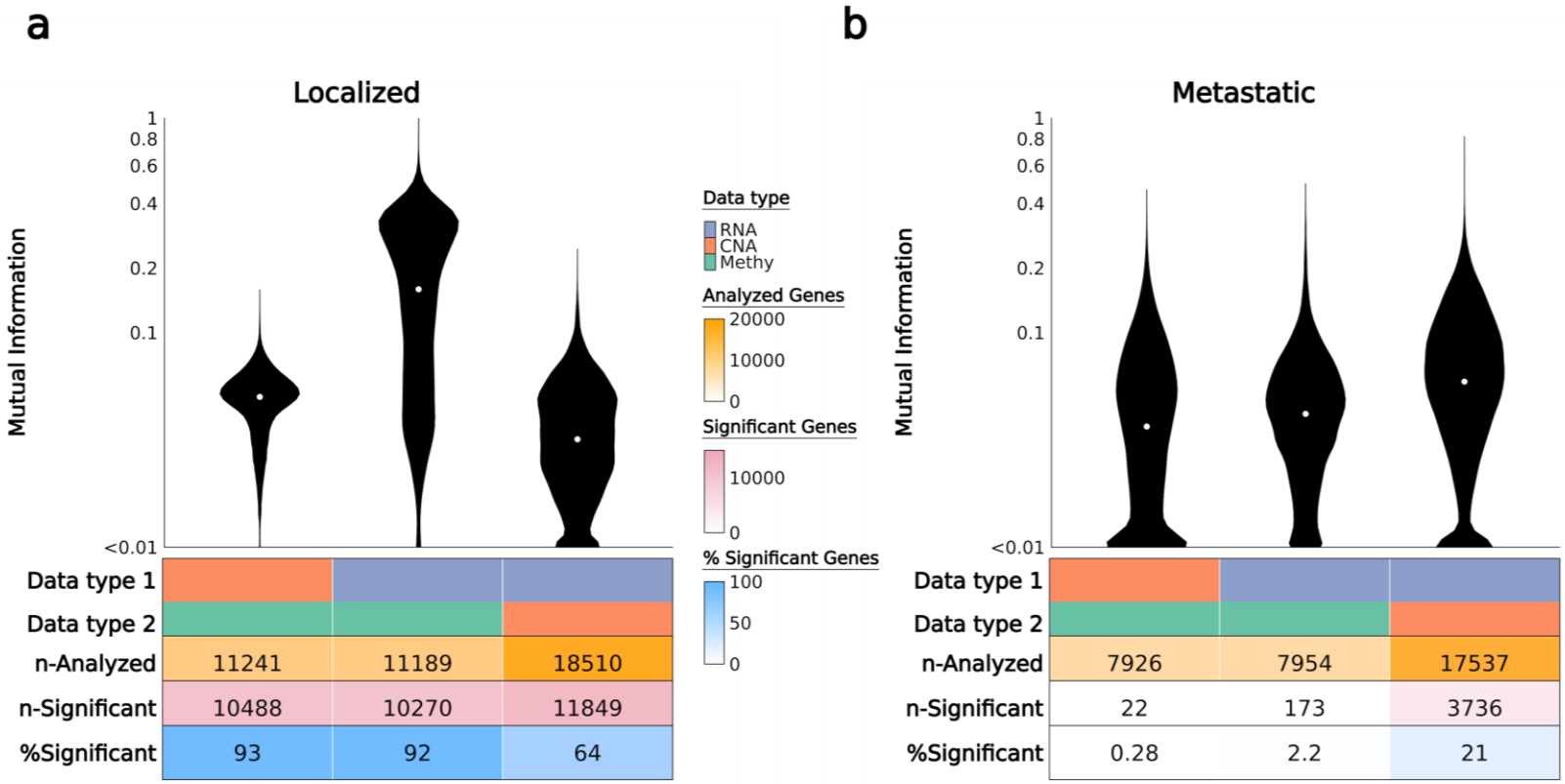


Supplementary Figure 5: Multi-omic correlations differ between localized and metastatic samples

a. Left: Spearman's correlation of gene-level methylation and CNA compared between localized and metastatic samples. Right: Heatmaps of methylation z-scores (top row) and CNA profiles (bottom row) stratified by tissue type.

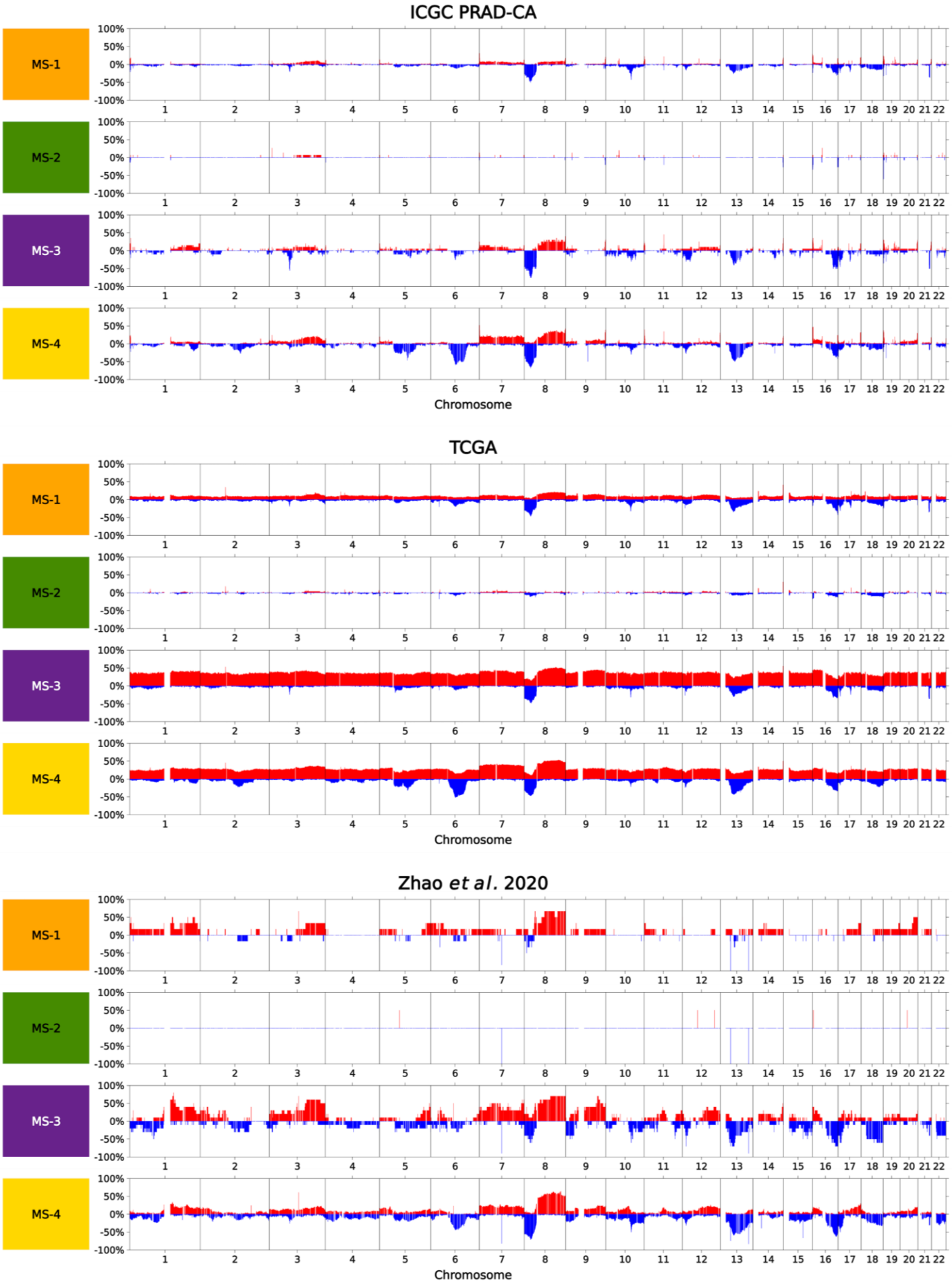
b-c. Similar to (a) except for methylation-RNA and CNA-RNA correlations.

d. Left: differences in gene-level correlations by tumor type $\rho_{Metastatic}$ - $\rho_{Localized}$ for each pair of data types. Right: heatmap of gene correlations for each pair of data types stratified by localized and metastatic samples.

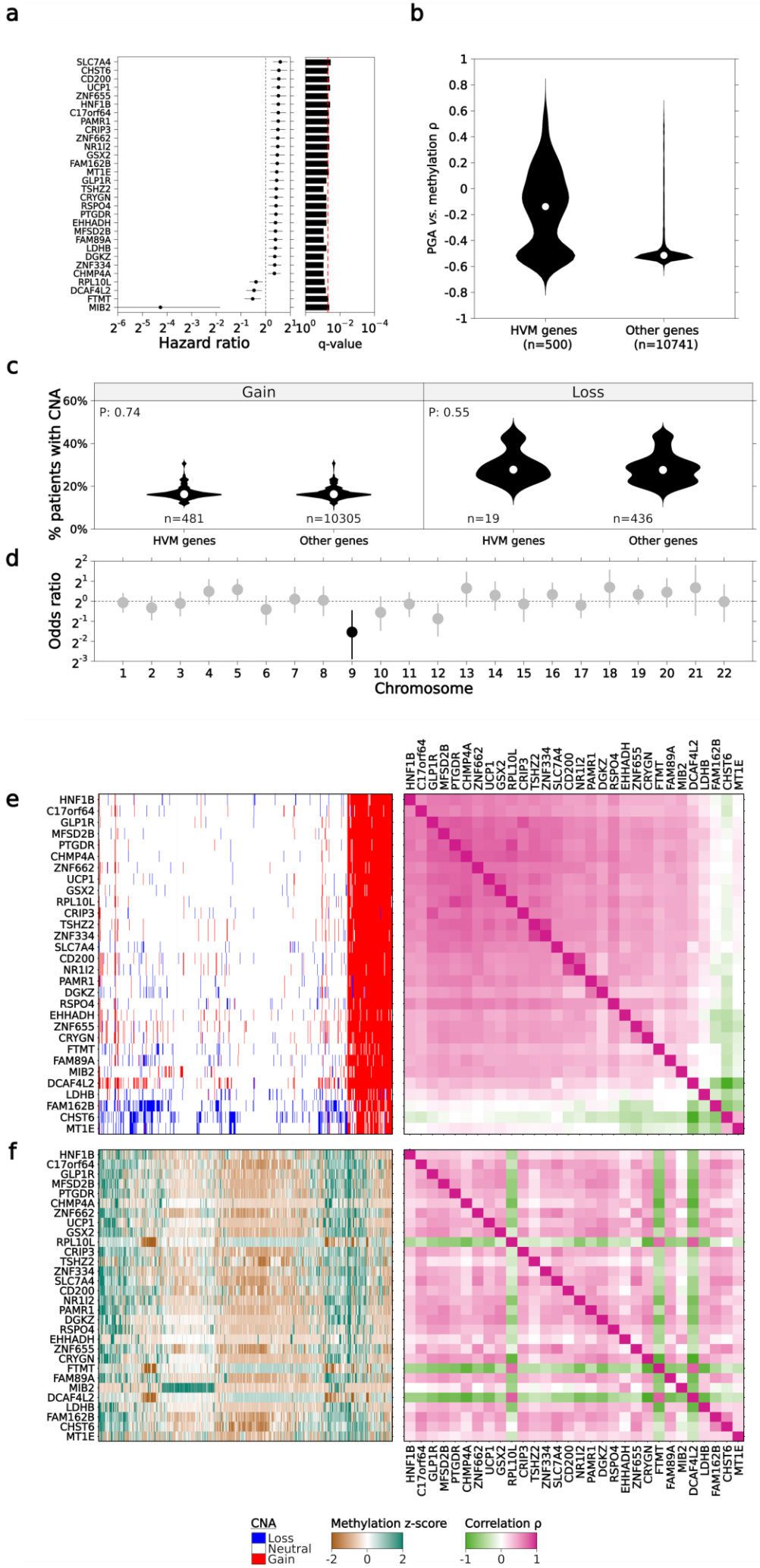


Supplementary Figure 6: Mutual information of multi-omic data differs in localized and metastatic samples

a-b. Mutual information was calculated for all features within each pair of data types (RNA, CNA, methylation) and displayed in a violin plot. A permutation test with 1000 permutations was used to calculate p-values and results with multiple-testing adjusted q<0.05 were regarded as statistically significant. Results were stratified by localized (**a**) and metastatic (**b**) prostate cancers.



Supplementary Figure 7: Genome-wide CNA profiles by subtype and cohort
The genome-wide copy number alteration (CNA) profile is shown for each subtype (row), stratified by the ICGC PRAD-CA, TCGA and Zhao *et al.* 2020 cohorts. Bars indicate the percent of patients with copy number gain or loss (y-axis) at each location of the genome (x-axis).



Supplementary Figure 8: Genes with high variation in methylation (HVM) *vs.* BCR and CNA

a. 30/500 HVM genes significantly associated with time-until-BCR (Cox proportional hazard Wald test q < 0.05). Gene methylation measured by median beta-value of CpG islands in the gene promoter region.

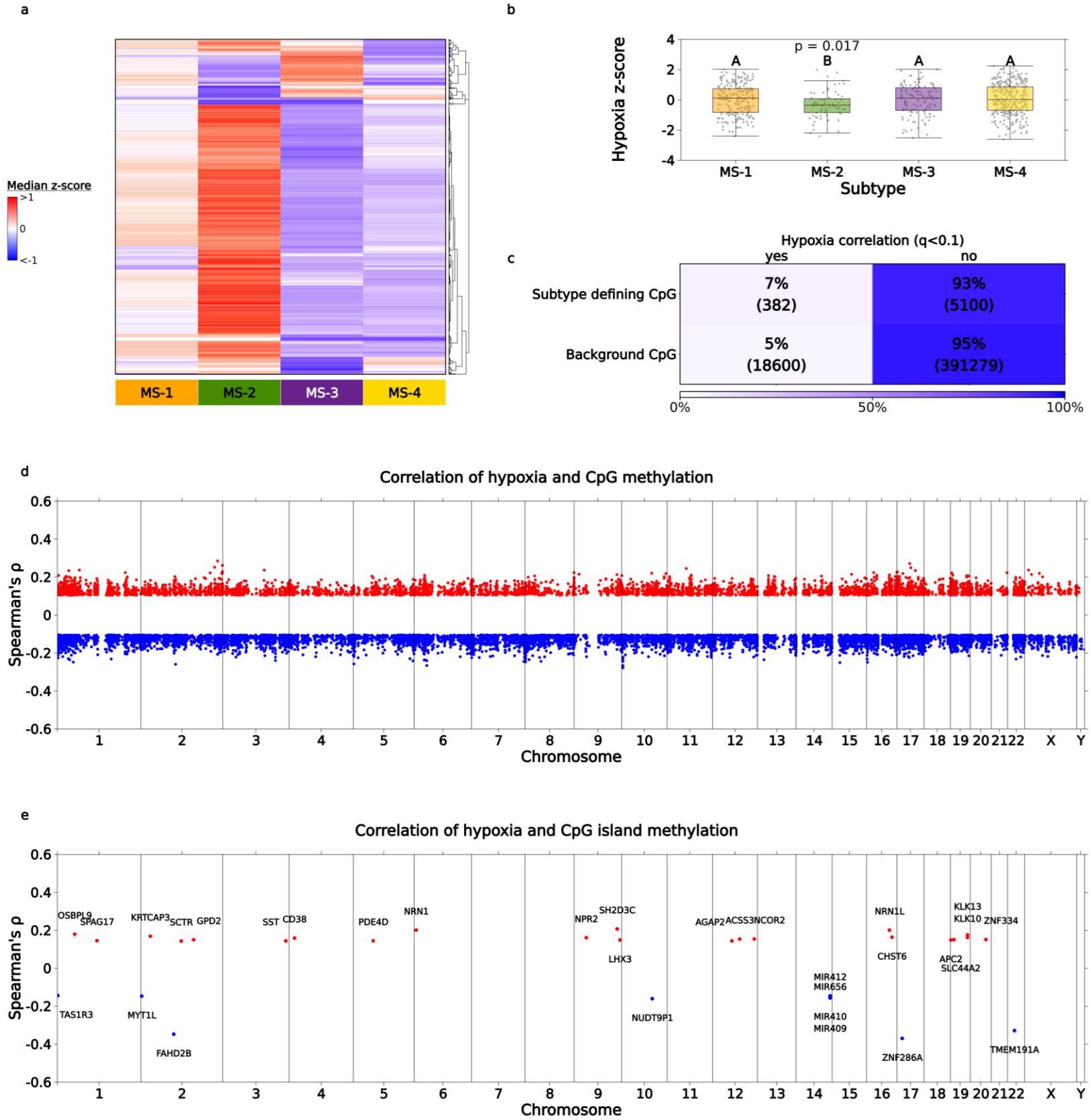
b. Spearman correlation of HVM and non-HVM gene methylation vs. PGA.

c. Percent of patients with CNA loss or gain was calculated separately for HVM and non-HVM genes and compared using Mann-Whitney U-test.

d. The number of HVM and non-HVM genes was compared within each chromosome using Fisher's exact test and odds ratio. Significant results shown in black, q < 0.05.

e. Left: CNA profiles for the 30 HVM genes in (a). Both rows and columns were sorted using hierarchical clustering with Euclidean distance and complete linkage. Right: Spearman's correlation heatmap of the CNA profiles.

f. Left: Methylation profiles for the 30 HVM genes in (a). Rows/columns were sorted to match (e). Right: Spearman's correlation heatmap of the methylation profiles.



Supplementary Figure 9: RNA *vs.* subtypes and hypoxia *vs.* methylation

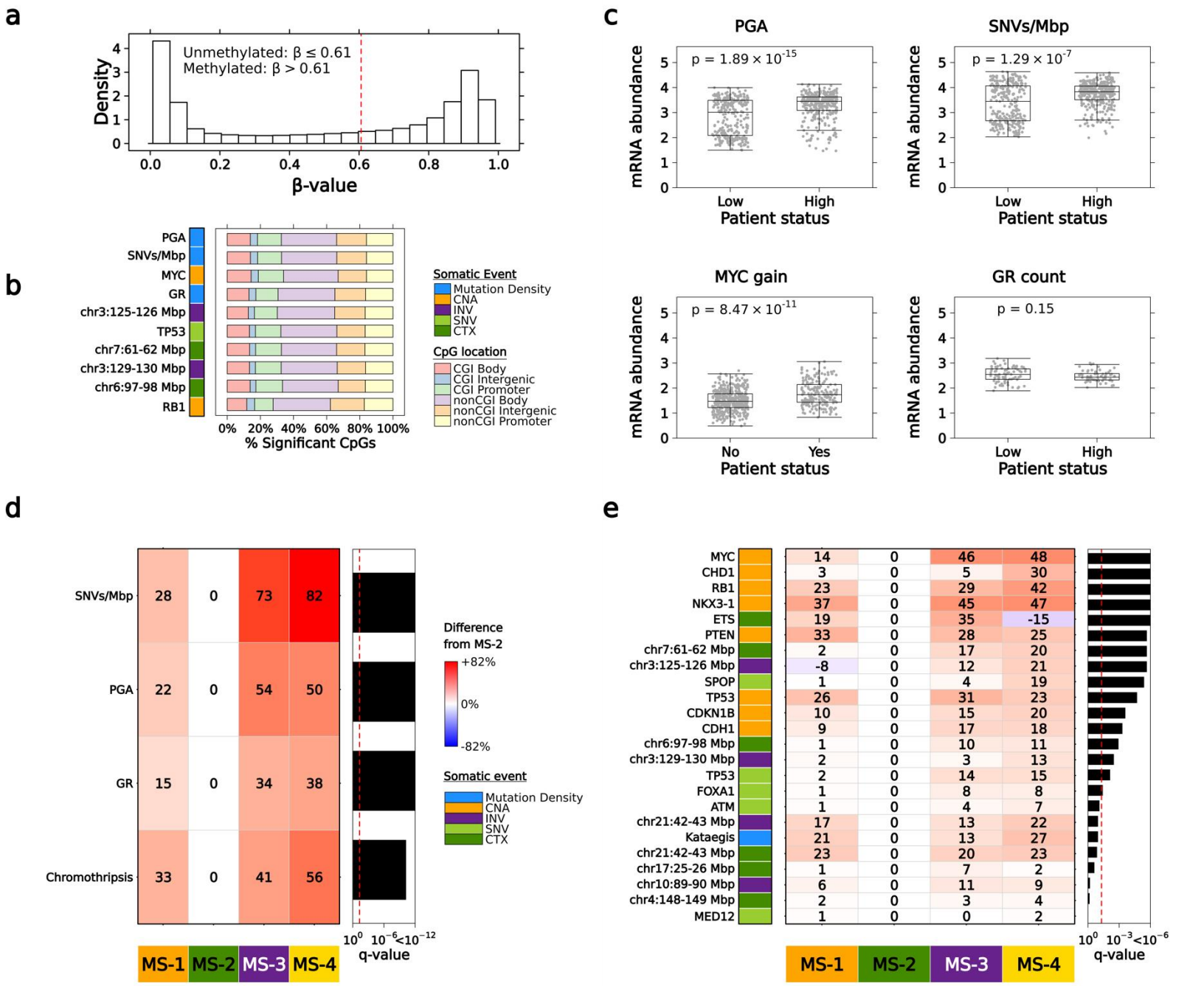
a. Heatmap of median adjusted RNA abundance z-score (adjusted for cohort and PGA) by subtype for 3,516 significant transcripts (Kruskal Wallis q < 0.05, standard deviation (SD) of subtype median z-scores > 0.4).

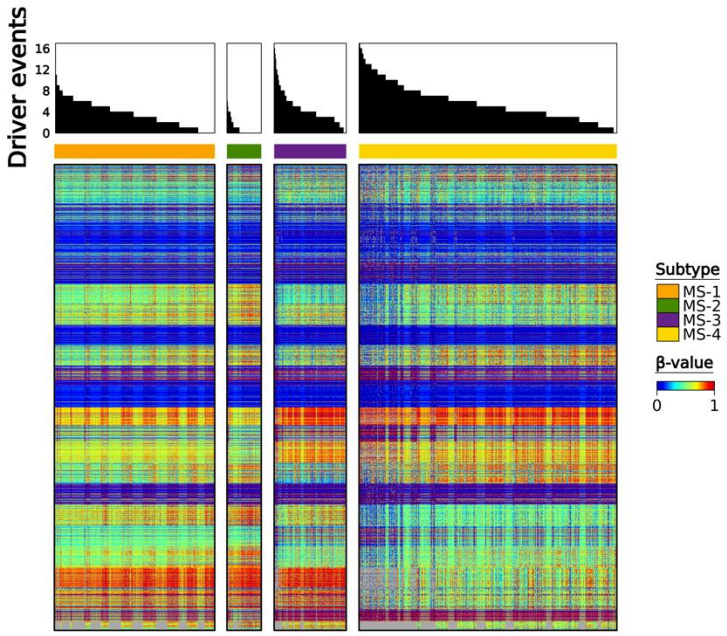
b. Hypoxia z-score was compared between subtypes using Kruskal-Wallis rank sum test. Pairwise comparisons between groups were conducted using the Mann-Whitney U-test, such that groups with different letters significantly differ (q < 0.05).

c. Spearman's correlation was calculated for each CpG with hypoxia score. The table compares the rate of significant CpG-hypoxia correlation (q < 0.05) between subtype-defining and non-subtype-defining background CpGs. Subtype defining CpGs have greater odds of hypoxia correlation: OR = 1.6 (95% CI: 1.4, 1.8).

d. Spearman's correlation between each CpG with hypoxia score. 18,982 significant CpGs shown (q<0.1) out of 415,361 tested; 69% negative correlations.

e. Spearman's corr. of hypoxia and methylation of CpG islands in gene promoters. 32 significant genes shown (q<0.1) out of 13,816 tested.





Supplementary Figure 10: Driver events *vs.* methylation

a. The median methylation β-value was calculated for each CpG and displayed as a histogram (498 TCGA patients; 418,822 CpGs). The red dotted line represents a median β-value = 0.61. This was used to assess a binary cut off for methylated and unmethylated status in Figure 4a.

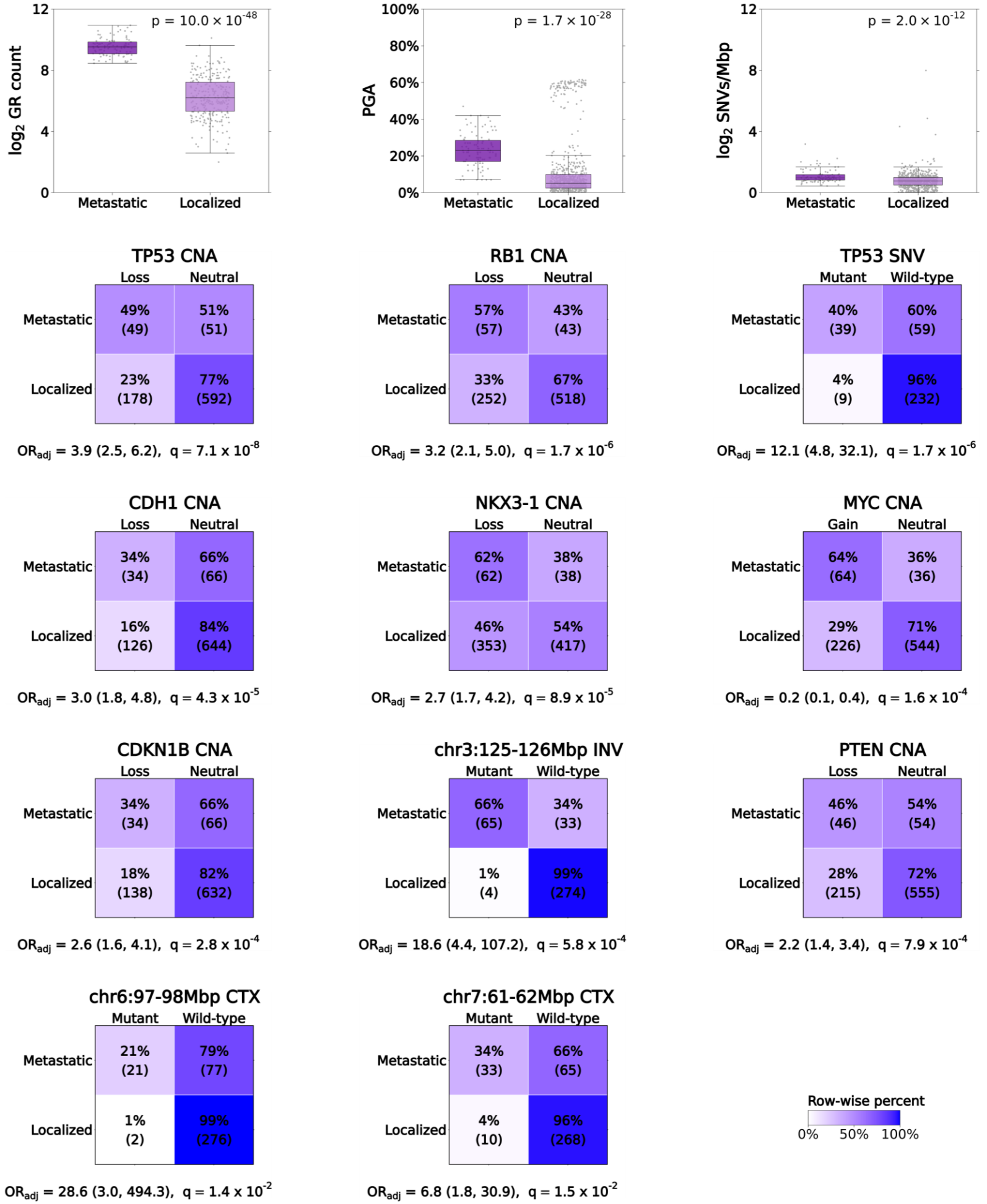
b. Top 10 driver events with the most number of significant CpG associations (q < 0.05). CpG locations were categorized into six groups with a combination of CpG island/non-CpG island and different genomic regions (gene promoter, body and intergenic). Percentages were calculated based on how many significant differentially methylated CpGs in each group.

c. mRNA abundance levels of the top four driver events with the most number of significant CpG associations. For a given driver event, patient-level mRNA abundance was calculated as the median abundance among all genes that map to the top 100 most significant CpG associations. SNV count and PGA were median-dichotomized into high *vs.* low groups. P-value was generated from the Mann-Whitney U-test.

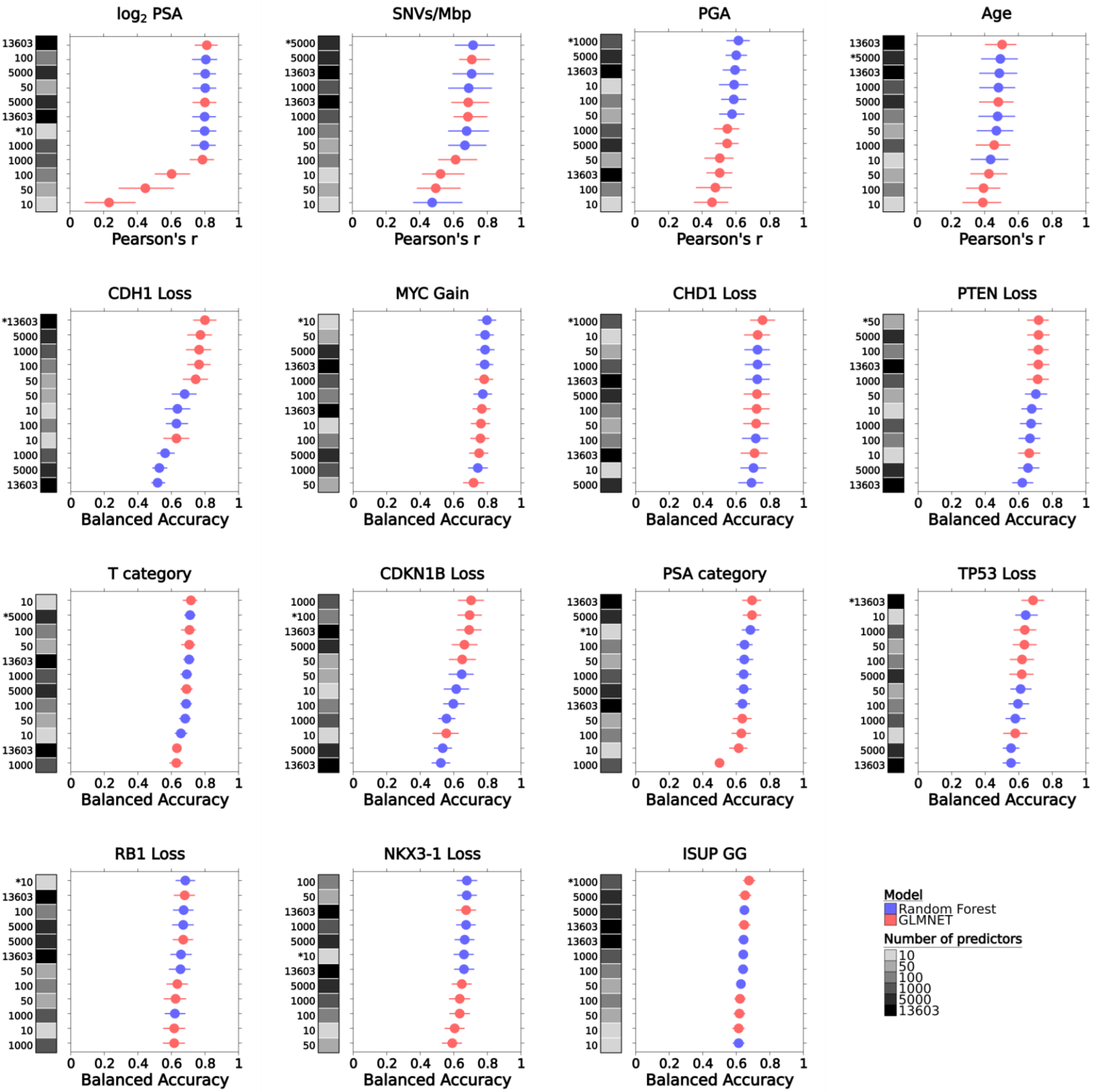
d. Numeric drivers *vs.* subtypes: the median percentile was calculated for each subtype then subtracted from MS-2; Kruskal-Wallis test q-value.

e. Binary drivers *vs.* subtypes: the percent of patients with the driver event was calculated for each subtype then subtracted from MS-2; chi-square test q-value (or Fisher's exact test if 2x2 table or cell with <5 expected counts).

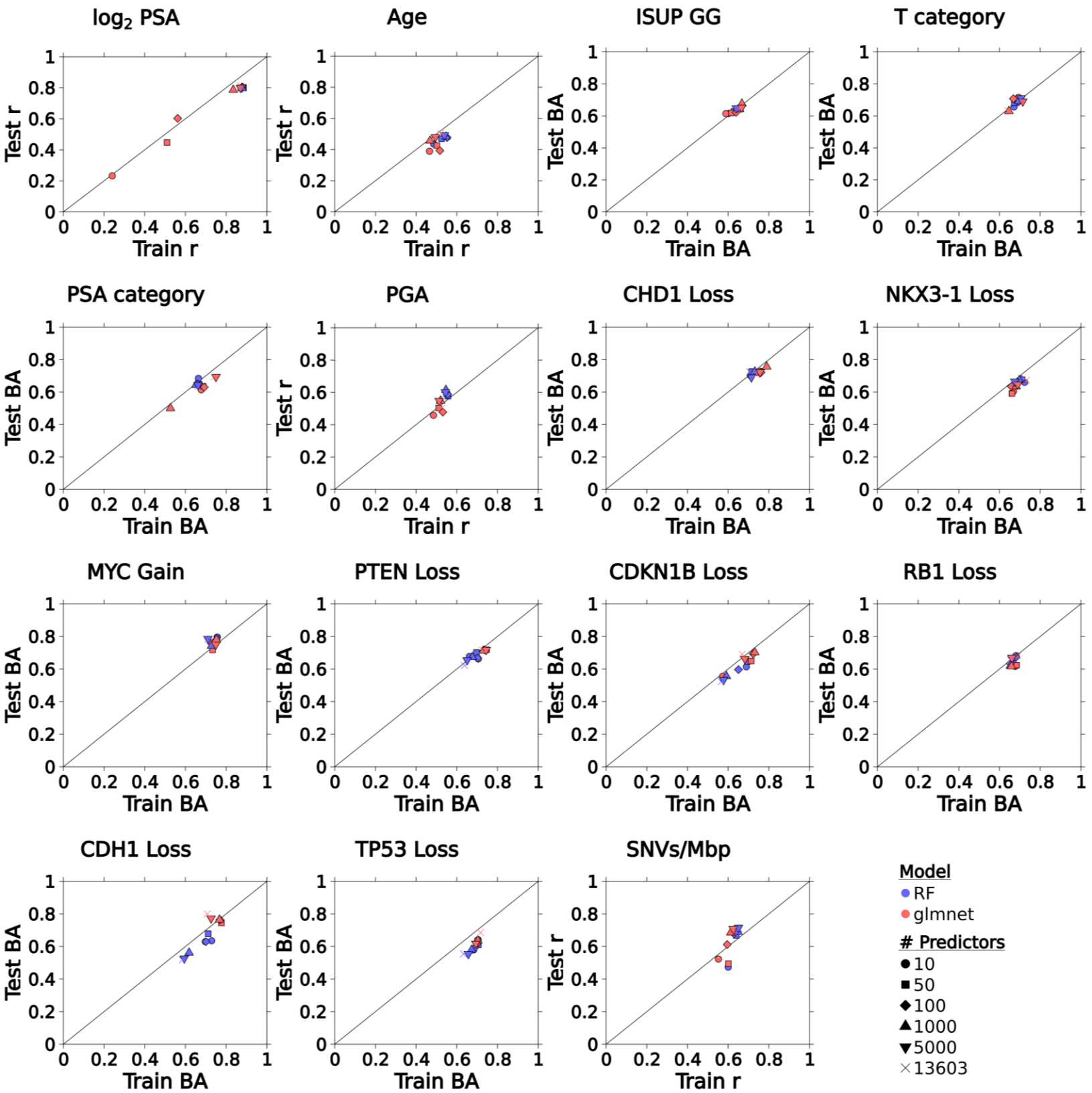
f. Heatmap of union of top 100 significant CpGs from each driver event (2,196 total CpGs). For each patient, the total number of driver events was calculated by summing binary indicator variables. Discrete drivers were binarized as presence/absence of a mutation, while continuous drivers were median dichotomized into high *vs.* low, with high adding to the total score.



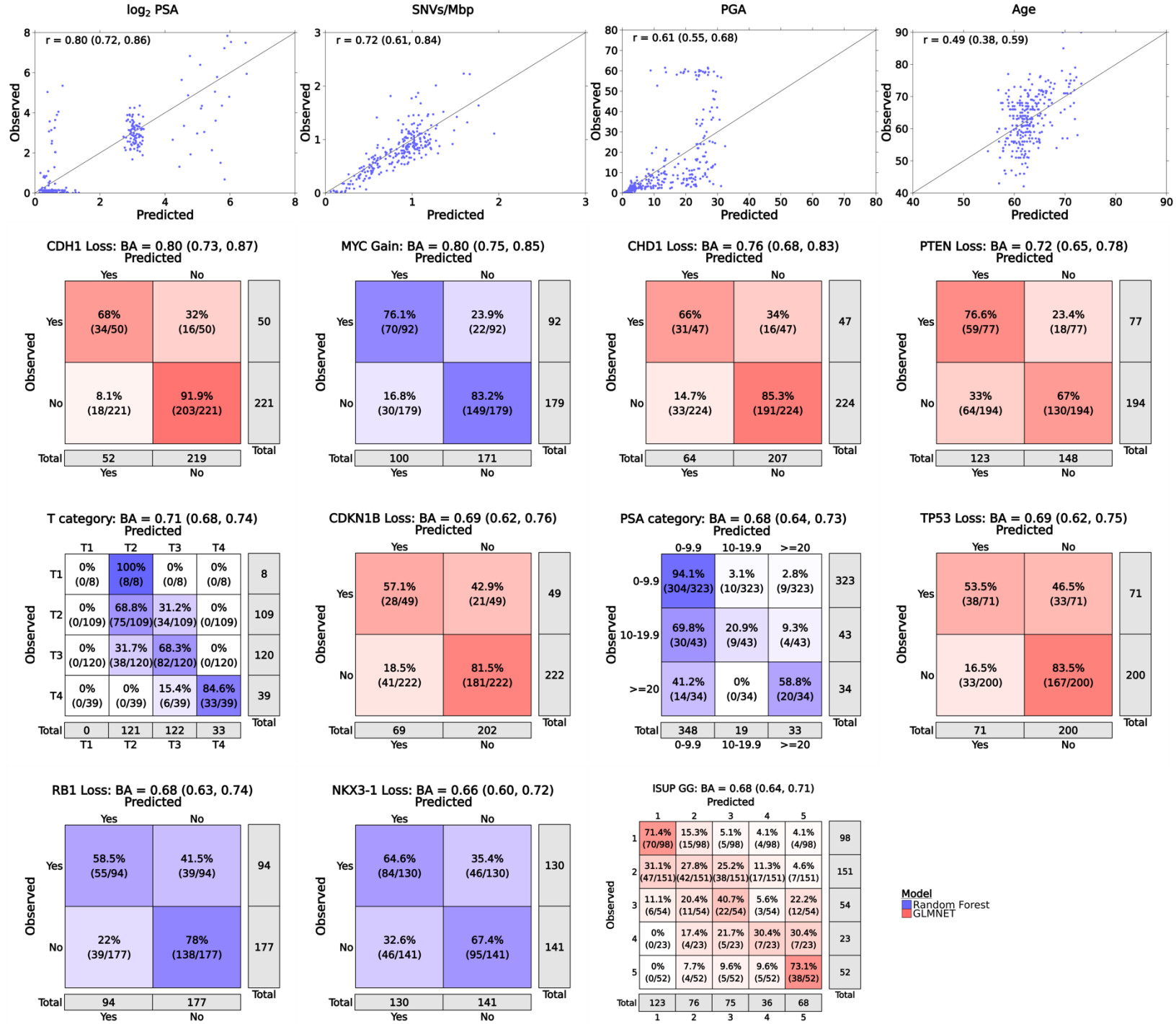
Supplementary Figure 11: Driver mutations in localized *vs.* metastatic prostate cancers
Mann-Whitney U-test p-value shown for continuous drivers (top row). For categorical drivers, logistic regression odds ratio (OR_{adj} with 95% confidence interval) adjusting for total mutation burden (CNA drivers adjusted for PGA, SNVs adjusted for log_2 SNVs/Mbp, and genomic rearrangements adjusted for log_2 GR count). INV = inversion, CTX = translocation.



Supplementary Figure 12: Performance on the test dataset after feature selection
For a given plot, the left black-grey covariate bar indicates how many features were kept after feature selection (see Methods). Pearson's correlation and balanced accuracy were used for numeric and categorical features, respectively. An asterisk in the black-grey covariate bar indicates the final selected model, which was the smallest model within 0.02 of the best model. Results for random forest are shown in blue and glmnet in red. 95% bootstrap confidence intervals are indicated by the lines besides each point.

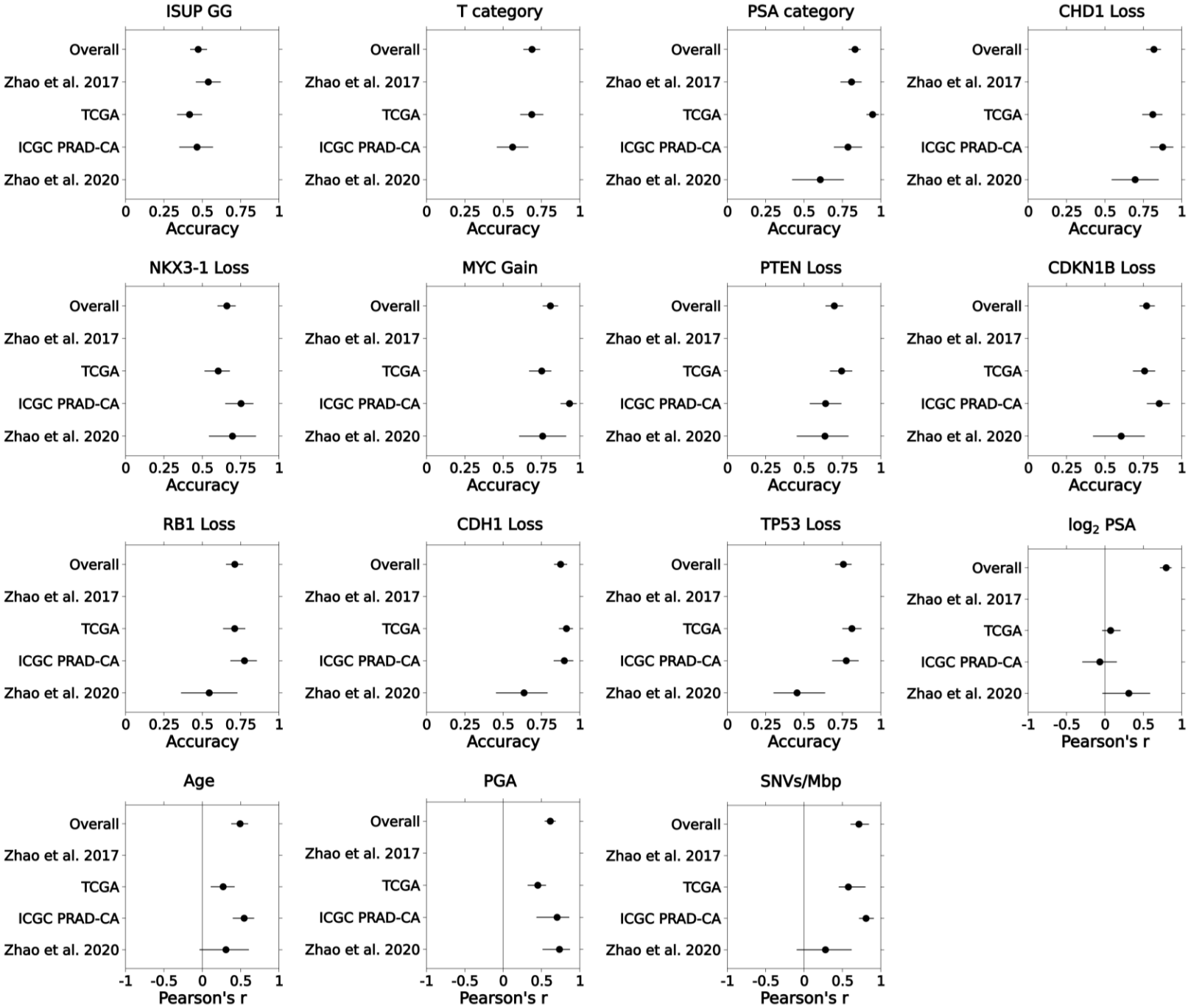


Supplementary Figure 13: Performance on the training and test datasets for all models
For a given scatterplot, the model performance is shown for the training dataset (x-axis) and test dataset (y-axis) for random forest (RF; blue) and glmnet (red), and different combinations of number of predictors used after feature selection (see legend for shapes of points). Pearson's correlation (r) and balanced accuracy (BA) were used for numeric and categorical features, respectively.

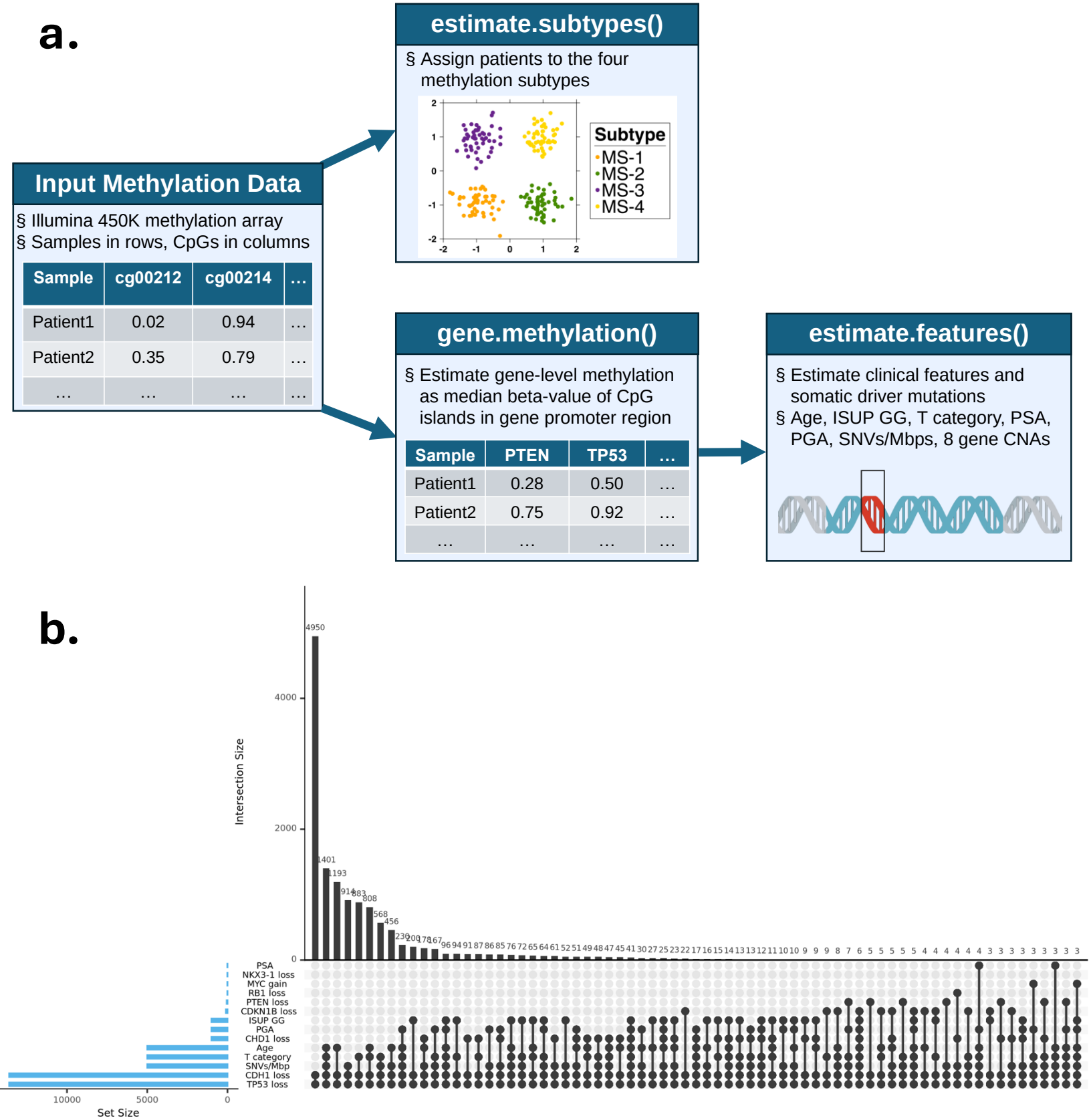


Supplementary Figure 14: Performance of the final models on the test dataset

Results are shown for the final selected models (see Methods for model section details), using scatterplots for numeric outcomes and confusion matrices for categorical. Results for random forest are shown in blue and glmnet in red. The shading of the confusion matrices is determined by the row-wise percentages, with higher percentages given darker color. Pearson's correlation is in the upper left corner of scatterplots and balanced accuracy (BA) in the title of the confusion matrix; 95% bootstrap confidence intervals are shown in parentheses.



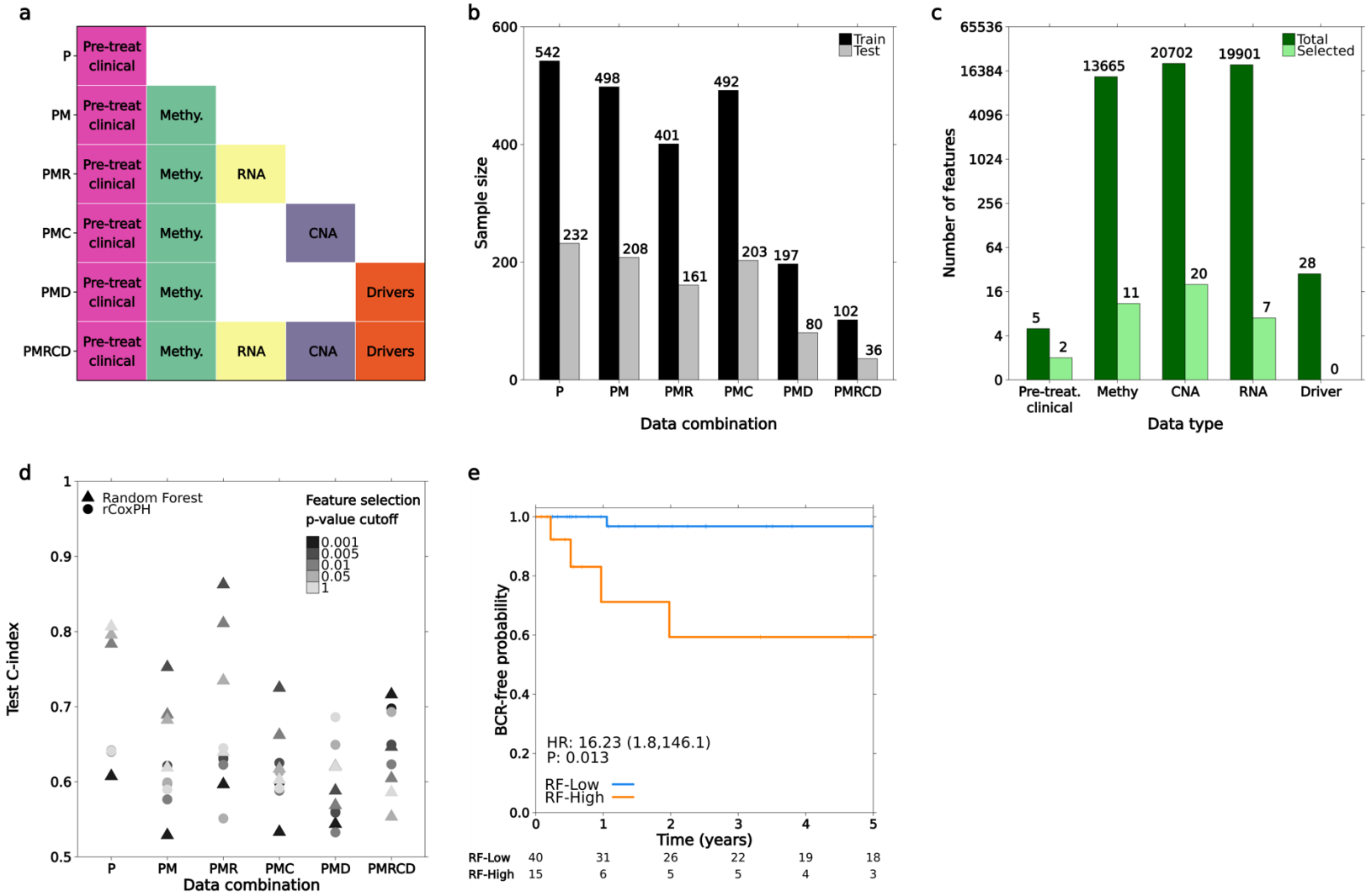
Supplementary Figure 15: Model performance on the test dataset stratified by cohort
For a given plot, results are shown for the final model when pooling all cohorts together ("Overall") or when fit separately by cohort as indicated in the y-axis labels. Pearson's r is used for numeric outcomes and accuracy for categorical outcomes (% of observations correctly classified). Although balanced accuracy (BA) was used to tune and select the final models for categorical outcomes, accuracy is shown here since we were unable to calculate BA for some of the cohorts due to small cell counts resulting in the BA being undefined. 95% bootstrap confidence intervals are indicated by the lines besides each point.



Supplementary Figure 16: PrCaMethy R package

a. User's input a data.frame of methylation data from the Illumina 450K methylation array. Samples are in rows and CpGs in columns. The matrix contains beta-values (continuous proportions ranging from 0 to 1). Using this input dataset, the flowchart demonstrates the workflow of the 3 main R functions.

b. Overlap of predictors used between the 14 models. Left barplot shows total genes per model. Top barplot shows overlap of genes for various combinations of models. For a given column of the barplot, the black dots with lines indicate which models are represented by that bar. The number above the bar indicates the number of genes shared in common between those models. Combinations that share at least 3 genes are shown.



Supplementary Figure 17: Multi-omic data for predicting time-until-BCR

a. Models for predicting time-until-BCR were built using different combinations of data types (Pre-treatment clinical, methylation, RNA, CNA and somatic driver mutations). The acronyms on the yaxis-labels indicate the combinations of data types and correspond to the xaxis-labels in parts (b) and (d).

b. A total of 774 localized prostate cancer patients with available time-until-BCR data were were randomly split into 70% Train (black) and 30% Test datasets (grey). The height of bars indicates the train and test set sample sizes for the different combinations of data types in (a).

c. The height of bars indicates the total number of features (dark green) and the number of features after performing feature selection (light green) with p-value cutoff 0.005. To perform feature selection, a separate Cox model was fit to each feature treating it as a predictor of time-until-BCR, while adjusting for the pre-treatment clinical features. Then only features with Wald-test p-value < 0.005 were retained.

d. The following p-value cutoffs were considered for the feature selection in part b: 0.001, 0.005, 0.01, 0.05 and 1 (1 meaning no feature selection, *i.e.* all features were considered in the model). Notice a random forest model with p-value cutoff 0.005 resulted in the highest C-index in the test dataset.

e. The best random forest model from Figure 4h. was used to predict BCR risk score in the test dataset and median-dichotomized into RF-High vs. RF-Low risk. The Kaplan-Meir plot shows the BCR-free probabilities for a subset of 55 patients classified as intermediate risk according to the National Comprehensive Cancer Network (NCCN) risk categories. The hazard ratio (HR), 95% confidence interval, and Wald-test p-value were estimated from a Cox proportional hazards model. The table below the plot gives the number of patients at risk from 0-5 years. Results for all 161 test patients are shown in Figure 4j.

Supplementary Table 1

	Subtype				
	Total (N=1824)	**MS-1 (N=667)**	**MS-2 (N=253)**	**MS-3 (N=266)**	**MS-4 (N=638)**
Cohort					
Zhao *et al.* 2017	525 (28.8%)	247 (37.0%)	113 (44.7%)	103 (38.7%)	62 (9.7%)
TCGA	498 (27.3%)	137 (20.5%)	56 (22.1%)	96 (36.1%)	209 (32.8%)
ICGC PRAD-CA	286 (15.7%)	160 (24.0%)	15 (5.9%)	20 (7.5%)	91 (14.3%)
Li *et al.* 2020	186 (10.2%)	26 (3.9%)	2 (0.8%)	11 (4.1%)	147 (23.0%)
ICGC PRAD-FR	130 (7.1%)	52 (7.8%)	38 (15.0%)	20 (7.5%)	20 (3.1%)
Zhao *et al.* 2020	100 (5.5%)	6 (0.9%)	2 (0.8%)	10 (3.8%)	82 (12.9%)
Kirby *et al.* 2017	72 (3.9%)	31 (4.6%)	25 (9.9%)	3 (1.1%)	13 (2.0%)
Melbourne	27 (1.5%)	8 (1.2%)	2 (0.8%)	3 (1.1%)	14 (2.2%)
Age					
<50	116 (6.9%)	62 (10.1%)	24 (11.2%)	16 (6.6%)	14 (2.3%)
50-59	526 (31.3%)	231 (37.6%)	77 (36.0%)	96 (39.5%)	122 (20.0%)
60-69	741 (44.1%)	247 (40.2%)	92 (43.0%)	106 (43.6%)	296 (48.4%)
>=70	299 (17.8%)	74 (12.1%)	21 (9.8%)	25 (10.3%)	179 (29.3%)
Missing	142	53	39	23	27
PSA ng/mL					
0 - 9.9	1120 (79.7%)	477 (84.1%)	174 (87.9%)	169 (80.1%)	300 (69.8%)
10 - 19.9	176 (12.5%)	72 (12.7%)	17 (8.6%)	27 (12.8%)	60 (14.0%)
>=20	110 (7.8%)	18 (3.2%)	7 (3.5%)	15 (7.1%)	70 (16.3%)
Missing	418	100	55	55	208
ISUP GG					
1	342 (25.5%)	179 (32.0%)	83 (41.1%)	52 (24.2%)	28 (7.7%)
2	532 (39.7%)	241 (43.1%)	79 (39.1%)	88 (40.9%)	124 (34.2%)
3	228 (17.0%)	93 (16.6%)	19 (9.4%)	23 (10.7%)	93 (25.6%)
4	86 (6.4%)	18 (3.2%)	11 (5.4%)	23 (10.7%)	34 (9.4%)
5	151 (11.3%)	28 (5.0%)	10 (5.0%)	29 (13.5%)	84 (23.1%)
Missing	485	108	51	51	275
T category					
T1	75 (6.5%)	13 (3.6%)	3 (3.0%)	7 (5.0%)	52 (9.6%)
T2	517 (45.0%)	206 (56.6%)	61 (60.4%)	44 (31.4%)	206 (37.9%)
T3	446 (38.8%)	137 (37.6%)	34 (33.7%)	79 (56.4%)	196 (36.0%)
T4	111 (9.7%)	8 (2.2%)	3 (3.0%)	10 (7.1%)	90 (16.5%)
Missing	675	303	152	126	94