nature genetics

Article

https://doi.org/10.1038/s41588-024-02030-z

Comprehensive genomic characterization of early-stage bladder cancer

Received: 31 October 2023

Accepted: 31 October 2024

Published online: 3 January 2025

Check for updates

Frederik Prip[1,2], Philippe Lamy[1], Sia Viborg Lindskrog[1,2], Trine Strandgaard[1,2], Iver Nordentoft[1], Karin Birkenkamp-Demtröder[1,2], Nicolai Juul Birkbak[1,2], Nanna Kristjánsdóttir[1,2], Asbjørn Kjær[1,2], Tine G. Andreasen[1,2], Johanne Ahrenfeldt[1,2], Jakob Skou Pedersen[1,2], Asta Mannstaedt Rasmussen[1,2], Gregers G. Hermann[3], Karin Mogensen[3], Astrid C. Petersen[4], Arndt Hartmann[5], Marc-Oliver Grimm[6], Marcus Horstmann[7], Roman Nawroth[8], Ulrika Segersten[9], Danijel Sikic[10], Kim E. M. van Kessel[11,12], Ellen C. Zwarthoff[13], Tobias Maurer[14], Tatjana Simic[15], Per-Uno Malmström[9], Núria Malats[16], Jørgen Bjerggaard Jensen[2,17], UROMOL Consortium*, Francisco X. Real[18,19] & Lars Dyrskjøt[1,2] ✉

Understanding the molecular landscape of nonmuscle-invasive bladder cancer (NMIBC) is essential to improve risk assessment and treatment regimens. We performed a comprehensive genomic analysis of patients with NMIBC using whole-exome sequencing (n = 438), shallow whole-genome sequencing (n = 362) and total RNA sequencing (n = 414). A large genomic variation within NMIBC was observed and correlated with different molecular subtypes. Frequent loss of heterozygosity in *FGFR3* and 17p (affecting *TP53*) was found in tumors with mutations in *FGFR3* and *TP53*, respectively. Whole-genome doubling (WGD) was observed in 15% of the tumors and was associated with worse outcomes. Tumors with WGD were genomically unstable, with alterations in cell-cycle-related genes and an altered immune composition. Finally, integrative clustering of multi-omics data highlighted the important role of genomic instability and immune cell exhaustion in disease aggressiveness. These findings advance our understanding of genomic differences associated with disease aggressiveness in NMIBC and may ultimately improve patient stratification.

Bladder cancer is the sixth most common malignancy among males globally[1] and exhibits highly heterogeneous molecular profiles and outcomes[2]. Most patients with nonmuscle-invasive bladder cancer (NMIBC; Ta, T1 and carcinoma in situ (CIS)) have a favorable prognosis, but up to 40% progress to muscle-invasive bladder cancer (MIBC; T2+) within 5 years, depending on clinical risk group[3]. Identifying patients with NMIBC who are likely to progress to MIBC is therefore critical to provide optimal treatment. Stage, grade and concomitant CIS are important risk factors[3]. However, patients with similar clinical and pathological risk profiles may show large differences in outcomes[4]. Thus, a better understanding of the underlying molecular landscape of NMIBC may improve the identification of patients with tumors that will eventually progress and may optimize treatment strategies.

The transcriptomic landscape of bladder cancer has been studied extensively, especially in MIBC[5]. In NMIBC, large studies have been carried out by the European Early-Stage Urothelial Cancer Molecular Biology (UROMOL) consortium, where four subtypes of NMIBC have been identified (class 1, 2a, 2b and 3)[6,7]. These were highly prognostic, with class 2a showing the highest risk of progression. Class 2b tumors had high immune cell infiltration and high expression of T cell exhaustion

A full list of affiliations appears at the end of the paper. ✉e-mail: lars@clin.au.dk

markers[6]. T cell exhaustion and immune suppression were recently associated with poor survival following treatment with Bacillus Calmette–Guérin (BCG)[8,9]. Class 1 and 3 tumors were characterized by a low risk of progression, and class 3 tumors had an immune-depleted phenotype with simultaneous enrichment for *FGFR3* mutations[6].

The genomic profile of NMIBC has previously been studied in smaller cohorts, revealing frequent mutations in *FGFR3*, *PIK3CA* and *STAG2*, as well as in chromatin modifier genes[10–15]. The copy-number alteration (CNA) landscape and the degree of genomic instability of NMIBC are highly variable between tumors, ranging from few alterations, mainly deletions in chromosome (chr) 9, to highly altered genomes[6,13,14]. However, due to limitations of the previously applied technologies, in-depth analyses have not been reported.

Here we present a comprehensive genomic characterization of tumors from 438 patients with NMIBC. Our results provide an increased understanding of disease aggressiveness in early-stage bladder cancer and may ultimately pave the way for new treatments and surveillance programs.

Results

Patient cohort

To understand disease aggressiveness in early-stage bladder cancer, we performed a genomic characterization of tumors from 438 patients with NMIBC from the European UROMOL consortium ($n = 296$)[6] and Aarhus University Hospital, Denmark ($n = 142$)[8]. The tumors represented the whole disease spectrum of NMIBC and included both incident ($n = 280$) and prevalent ($n = 158$) cases. Total RNA-sequencing (RNA-seq) data from 414 tumors was available[6] (see Supplementary Table 1 for details).

The mutational landscape of NMIBC

Whole-exome sequencing (WES) was performed on paired tumor and reference leukocyte DNA from 438 patients with NMIBC. Tumor and matched germline DNA were sequenced to a mean coverage of 132× (range = 35–338×) and 128× (range = 31–302×), respectively. We identified a median of 179 (range = 5–8,131) single-nucleotide variants (SNVs) and 11 (range = 0–769) insertions and deletions (InDels) within the exome target regions.

The tumor mutational burden (TMB; nonsynonymous mutations per megabase) was higher in T1 tumors (median = 5.7) compared to Ta tumors (median = 3.9, $P = 8.1 \times 10^{-6}$) and varied across the UROMOL2021 transcriptomic classes, with class 2a tumors having the highest TMB ($P = 4.5 \times 10^{-17}$; Extended Data Fig. 1a). An elevated TMB was strongly associated with mutations in *ERCC2* ($P = 5.2 \times 10^{-17}$; Extended Data Fig. 1b).

In total, 60 genes were significantly mutated in the cohort (mutsigCV; Supplementary Table 2), and 33 of these were mutated in at least 5% of the tumors (Fig. 1a). Of these 33 genes, the five most frequently mutated were *FGFR3* (58%), *KDM6A* (42%), *KMT2D* (36%), *PIK3CA* (31%) and *STAG2* (24%). Several of the identified genes were related to epigenetic modification (*ARID1A*, *KDM6A*, *KMT2C*, *CREBBP*, *EP300* and *UTY*). A total of 88.3% of the SNVs called in the 33 genes were present in the corresponding transcriptomes (Extended Data Fig. 1c). Interestingly, we observed a significant overexpression of the alternate allele for several of the genes, especially for *FGFR3* ($P = 1.5 \times 10^{-36}$) and *HRAS* ($P = 0.0011$), both part of the RAS-mitogen-activated protein kinase pathway (Extended Data Fig. 1d and Supplementary Table 3). Mutation frequencies for Ta and T1 tumors are listed in Supplementary Table 4. Mutations were similarly distributed across males and females except for a higher proportion of *KDM6A* (located on chrX) mutations in females (Extended Data Fig. 1e; $P = 0.017$), as described previously[13]. Of note, *STAG2* and *RBM10*, which are also located on chrX, were mutated with a similar frequency in males and females.

A high fraction of mutations in the 33 significantly mutated genes (Fig. 1a) were clonal, indicated by their high inferred cancer cell fraction (CCF; Extended Data Fig. 2a). The clonal behavior and biology of NMIBC were further documented from paired WES analysis of recurring NMIBCs from 60 patients. In this context, mutations in these genes were frequently rediscovered in recurrent tumors, particularly mutations in *TP53* (83%), *KDM6A* (79%) and *FGFR3* (77%; Fig. 1b).

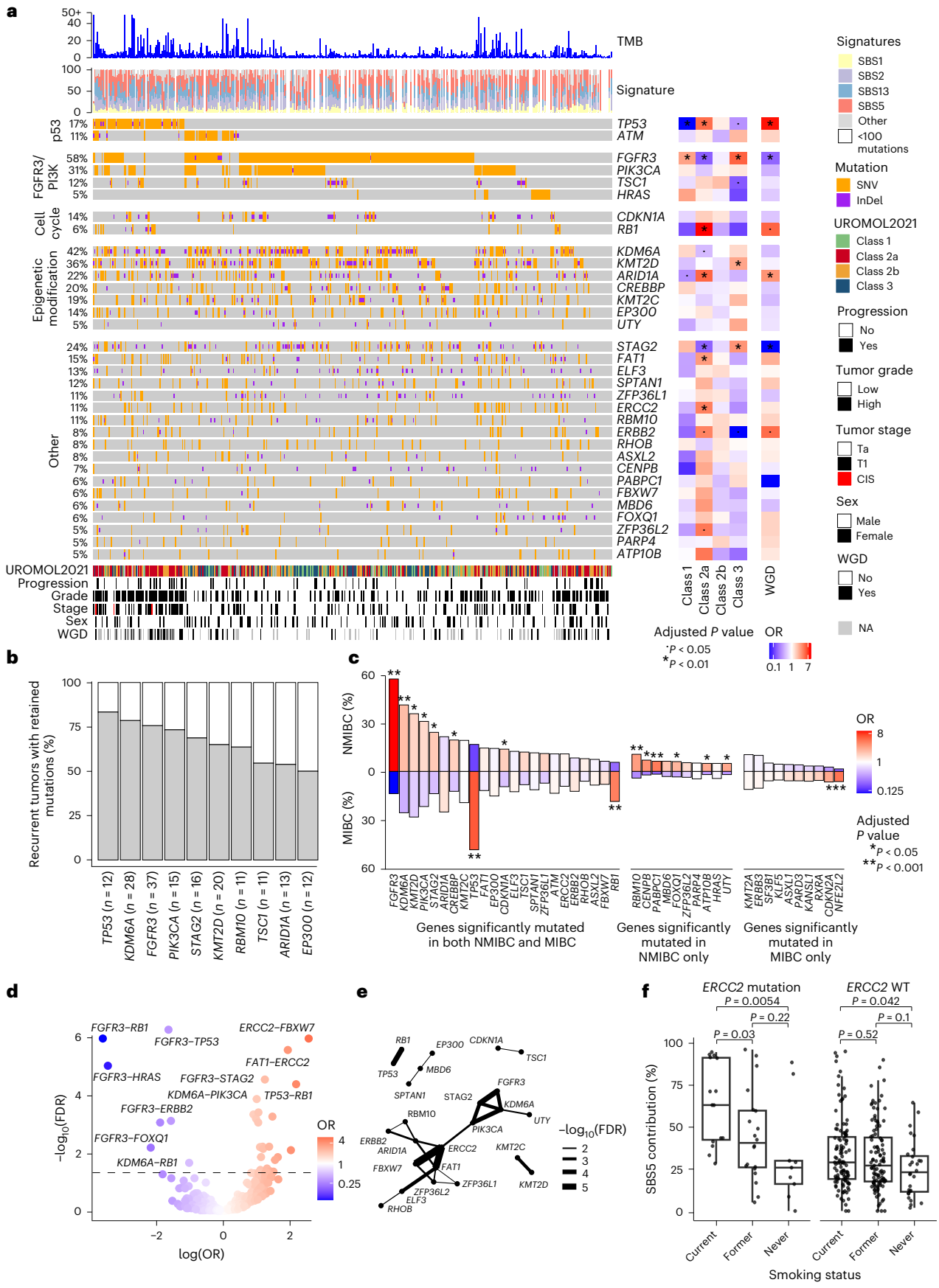
We observed that 70% (23/33) of the significantly mutated genes in NMIBC were also significantly mutated in MIBC (The Cancer Genome Atlas (TCGA)[16]; Fig. 1c). Some of the main differences in mutation frequencies were found for *FGFR3* (NMIBC = 58%, MIBC = 14%; $P = 9.8 \times 10^{-48}$), *KDM6A* (NMIBC = 42%, MIBC = 25%; $P = 6.5 \times 10^{-8}$), *TP53* (NMIBC = 17%, MIBC = 49%; $P = 2.2 \times 10^{-29}$) and *RB1* (NMIBC = 6%, MIBC = 19%; $P = 7.5 \times 10^{-10}$).

Several of the 33 significantly mutated genes in NMIBC had significantly different mutation frequencies across the transcriptomic UROMOL2021 classes (Fig. 1a and Extended Data Fig. 2b–e). *FGFR3* mutations were more frequent in class 1 and 3 tumors. Class 3 tumors additionally had frequent mutations in *KMT2D* and *STAG2*. Class 2a tumors showed a mutational profile highly similar to MIBC, including a higher proportion of mutations in *TP53* and *RB1* and fewer mutations in *FGFR3* (Extended Data Fig. 2d). *ARID1A*, *FAT1*, *ERCC2*, *ERBB2* and *ZFP36L2* were also more frequently mutated in class 2a tumors compared to the other classes. In contrast, no genes were enriched for mutations in class 2b tumors (Extended Data Fig. 2e), and pairwise comparisons between class 2b and the other classes did not identify a distinct mutational profile of class 2b tumors either (Extended Data Fig. 2f–i).

TP53 was the only gene where mutations were associated with an increased risk of progression (hazard ratio (HR) = 5.2, 95% confidence interval (CI) = 2–14, $P = 0.03$; Extended Data Fig. 3a). However, the association was not significant after adjusting for tumor stage and

Fig. 1 | Genomic landscape of NMIBC. a, Oncoplot of the 33 significantly mutated genes in the cohort (MutSigCV; mutated in >5% of tumors). The annotations on the right show enrichment for mutations in the UROMOL2021 transcriptomic classes and in tumors with WGD. Statistically significant associations between mutations and the UROMOL2021 classes and WGD were determined using Fisher's exact test. *P* values were adjusted using the FDR approach. Dot indicates $P < 0.05$ and the asterisk indicates $P < 0.01$. **b**, Comparison of gene mutations in paired, recurring tumors from 60 patients. Only genes that were mutated in at least ten of the earliest tumors were included. On the *x* axis, *n* indicates the number of the earliest tumors with a mutation in the respective gene. Gray bars show the percentage of recurrent tumors where a mutation in the given gene was rediscovered. **c**, Comparison of mutation frequencies in NMIBC and MIBC. Selected genes include genes significantly mutated in the UROMOL cohort (NMIBC; 33 genes displayed in **a**) and genes significantly mutated in the TCGA cohort (MIBC). The set of genes significantly mutated in both the NMIBC and MIBC cohorts is listed on the left-hand side, and the set of genes significantly mutated in NMIBC only is listed in the middle and the set of genes significantly mutated in MIBC only is listed on the right-hand side. Fisher's exact tests were performed to assess differences in mutation frequencies between NMIBC and MIBC, and asterisks indicate genes with significantly different mutation frequencies. Colors represent the OR between the mutation frequency of a gene in NMIBC and MIBC. *P* values were adjusted using the FDR approach. **d**, Degree of co-occurrence (red dots) and mutual exclusivity (blue dots) between the 33 significantly mutated genes (**a**). *P* values and ORs are listed in Supplementary Table 5. **e**, Network of significantly co-occurring mutations (OR > 1 and FDR < 0.05 in **d**). The width of the lines indicates the significance level of the respective associations (thicker lines indicate lower FDR). **f**, Percentage of mutations attributable to the SBS5 signature stratified by smoking status and *ERCC2* mutation status (*ERCC2* mutation—15 current smokers, 20 former smokers and 9 never smokers; *ERCC2* WT—111 current smokers, 107 former smokers and 28 never smokers). Boxplots represent the median and lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range. Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests.

https://doi.org/10.1038/s41588-024-02030-z



grade (HR = 2.6, 95% CI = 0.91–7.6, $P = 0.075$). No specific mutation was associated with the risk of recurrence (Extended Data Fig. 3b).

Several of the significantly mutated genes ($n = 33$) showed strong mutual exclusivity and co-occurrence patterns (Fig. 1d,e and Supplementary Table 5). Most of the mutually exclusive gene pairs could be explained by the fact that these mutations were present in different transcriptomic classes (Fig. 1a). However, *FGFR3* and *HRAS*, which are frequently mutated in class 1, showed a strong mutually exclusive distribution (odds ratio (OR) = 0.03; 95% CI = 0.00075–0.2)), as previously observed[17]. The correlation analysis also showed that the most significant co-occurrence of mutations was observed for *ERCC2*–*FBXW7* (OR = 13.0; 95% CI = 5.3–35.6), mainly observed in class 2a; *ERCC2*–*FAT1* (OR = 7.0; 95% CI = 3.4–14.1), mainly observed in class 2a; *FGFR3*–*STAG2* (OR = 3.5; 95% CI = 2.1–6.1), mainly observed in class 1/3; and *TP53*–*RB1* (OR = 8.9; 95% CI = 3.6–23.4), mainly observed in class 2a (Fig. 1e). Only co-occurrences of mutations in *FGFR3*–*STAG2* and *TP53*–*RB1* have previously been reported[12,16,18]. *ERCC2*–*FBXW7* mutations were also significantly co-occurring in MIBC (TCGA cohort; $P = 0.006$; OR = 3.69; 95% CI = 1.32–9.41).

The mutational profile of each tumor was decomposed into a subset of nine known Catalogue of Somatic Mutations in Cancer (COSMIC) single-base substitution (SBS) signatures[19] (SBS1, SBS2, SBS4, SBS5, SBS10b, SBS13, SBS15, SBS29 and SBS31; Fig. 1a). SBS5 along with the two apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC)-related signatures, SBS2 and SBS13, showed the highest contribution to the mutational landscape (Fig. 1a and Extended Data Fig. 3c). APOBEC-related mutagenesis has been associated with both worse and improved outcomes in NMIBC[6,7,15]. Here we observed no association between APOBEC-related mutagenesis and UROMOL2021 classes or outcome when applying WES data for signature calling (Supplementary Note—Mutational signatures).

When comparing the mutational contribution of SBS5 in relation to the smoking status of the patients in tumors with and without *ERCC2* mutations, we found that tumors from patients with a history of smoking had a significantly higher SBS5 contribution (Fig. 1f). Smoking and *ERCC2* mutations have previously been associated with SBS5 mutations in MIBC[20], and our findings support that SBS5-related mutagenesis is accelerated by *ERCC2* mutations when exposed to tobacco carcinogens.

The CNA landscape of NMIBC

Total CNAs were estimated from shallow whole-genome sequencing (sWGS) of DNA from 362 tumors (mean coverage = 2.14× (range = 1.24–9.01×)) using ichorCNA[21]. Additionally, using Battenberg[22] and corresponding WES data, unbalanced copy-number segments were identified using the counts of each parental allele at common heterozygous single-nucleotide polymorphisms (Fig. 2a; Methods). For simplicity, we will use the term whole-genome doubling (WGD) to refer to tumors with high copy numbers. Tumors were classified as having WGD if >50% of the genome had a copy number >2 ($n = 54$; 15%; Fig. 2b)[23]. Copy-number segments with four copies were largely balanced in tumors classified as having WGD as opposed to tumors classified as diploid (Extended Data Fig. 4a,b), thereby reinforcing the validity of the respective classifications. Using a similar analytical approach on data from MIBC (TCGA cohort[16]), we expectedly found a higher proportion of tumors with WGD (58%; Fig. 2b). The median ploidy of tumors with WGD was 3.5 (range = 2.7–4.2) for NMIBC (Extended Data Fig. 4c) and 3.3 (range = 2.3–4.8) for MIBC, reflecting a high degree of deletions in tumors with WGD (compared to four copies; Extended Data Fig. 4d,e). Furthermore, tumors with WGD were enriched for copy-number gains (Extended Data Fig. 4f,g).

WGD was observed more frequently in T1 tumors (T1 versus Ta, $P = 6.2 \times 10^{-6}$), high-grade tumors (high grade versus low grade, $P = 1 \times 10^{-8}$) and in the UROMOL2021 class 2a and class 2b tumors (class 2a/class 2b versus class 1/class 3, $P = 2.7 \times 10^{-6}$; Extended Data Fig. 4h). Concordantly, patients having tumors with WGD had significantly shorter progression-free survival (PFS) compared to patients with diploid tumors ($P = 2.7 \times 10^{-6}$; Fig. 2c), independently of tumor stage and grade (HR = 6.4, 95% CI = 1.8–23, $P = 0.0039$). Notably, a significant association was still observed when restricting the analysis to patients with tumors classified as UROMOL2021 class 2a ($P = 0.0025$; Fig. 2c).

Tumors with WGD showed a significantly higher TMB compared to diploid tumors ($P = 3.6 \times 10^{-8}$; Extended Data Fig. 4i). The mutational spectrum in tumors with WGD was not enriched for specific mutational signatures when compared to diploid tumors (Extended Data Fig. 4j). Mutations showing the strongest association with WGD were *TP53* mutations ($P = 1.6 \times 10^{-6}$), as 40% of tumors with WGD harbored a *TP53* mutation compared to 9% of diploid tumors (Fig. 2d). *ARID1A* ($P = 0.007$), *ERBB2* ($P = 0.011$) and *RB1* ($P = 0.034$) mutations were likewise associated with WGD. Furthermore, *ARID1A* and *ERBB2* mutations were associated with WGD independently of *TP53* mutational status (*ARID1A*, $P = 0.006$; *ERBB2*, $P = 0.007$). On the contrary, *FGFR3* ($P = 1.1 \times 10^{-5}$) and *STAG2* ($P = 1.3 \times 10^{-4}$) mutations were significantly enriched in diploid tumors (Fig. 2d).

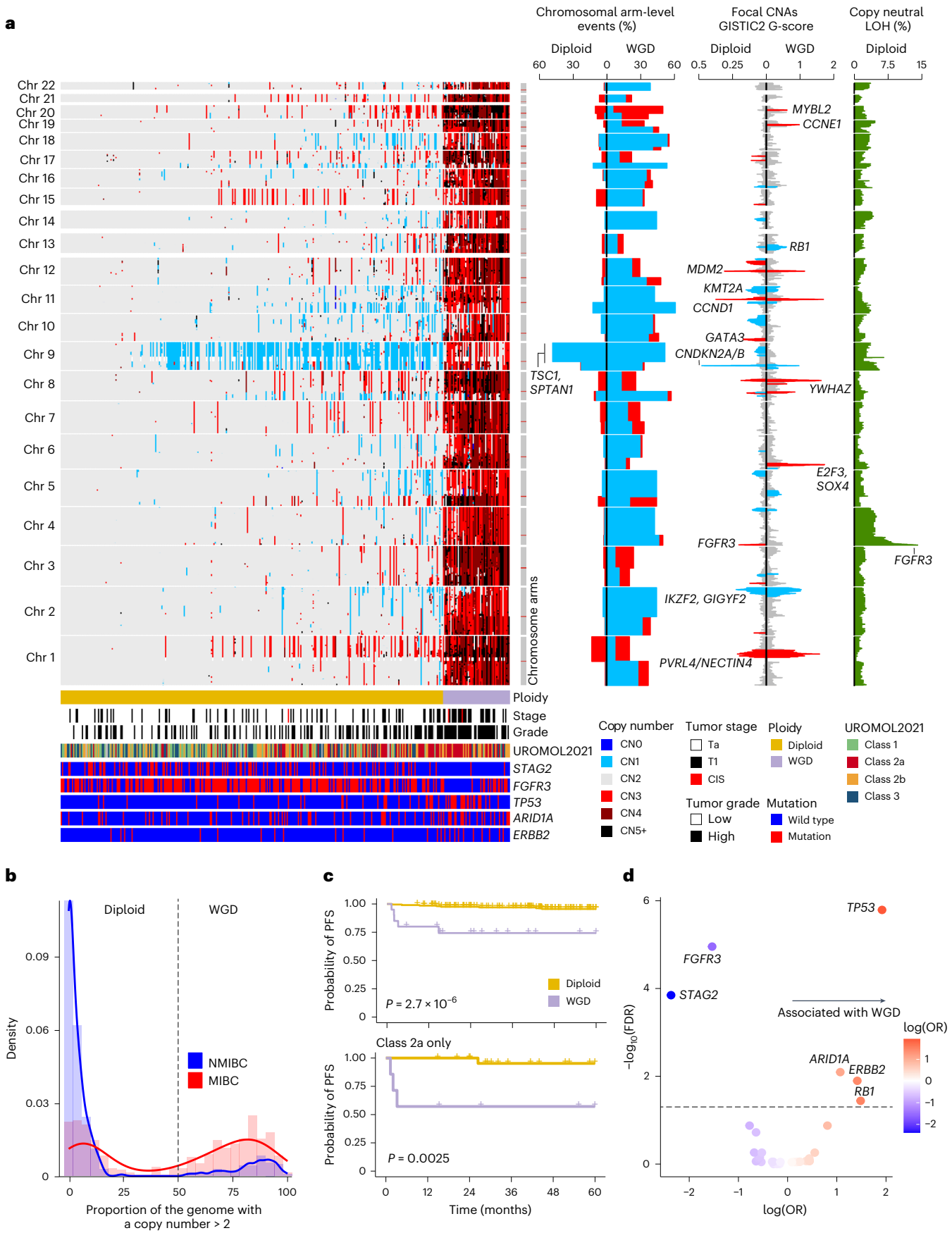
CNAs in diploid tumors and tumors with WGD

Analysis of chromosomal arm-level events (>70% of the arm altered) showed that loss of 9q occurred at a similar frequency in diploid tumors (47%) and tumors with WGD (52%; Fig. 3a). Mutations in *TSC1* and *SPTAN1* (located on 9q) were strongly associated with loss of 9q in diploid tumors (*TSC1*, $P = 2.1 \times 10^{-5}$; *SPTAN1*, $P = 7.1 \times 10^{-6}$; Extended Data Fig. 5a). The most frequently lost chromosomal arms in tumors with WGD were 11p, 8p, 18q and 17p (affecting *TP53*; all >50%). Interestingly, the copy-number state in these regions was frequently unbalanced with loss of heterozygosity (LOH; Fig. 3b). 17p and 9q had the highest proportion of LOH, which was predominantly copy number 2/0 (representing loss of one allele and duplication of the remaining), indicating that the chr arm loss likely occurred before WGD. Mutations in *TP53* (17p) were associated with LOH of the gene in both diploid tumors ($P = 0.00076$) and tumors with WGD ($P = 0.00073$; Fig. 3c). MIBCs with WGD similarly showed frequent LOH at 17p (65%). The most commonly gained chromosomal arm in tumors with WGD was 20q (44%), which was also observed in diploid tumors, although to a lesser extent (10%; Fig. 3a).

Genomic regions enriched for focal CNAs were identified using GISTIC2 (ref. 24; Fig. 3d,e and Supplementary Tables 6 and 7). Copy-number frequencies of selected genes are listed in Supplementary Table 8. Several genomic regions were significantly altered in both diploid tumors and tumors with WGD, including amplifications of 11q13.3 (*CCND1*), 12q15 (*MDM2*), 8q22.3 (*GRHL2* and *YWAZ*), different regions on 1q, for example,

Fig. 2 | WGD in NMIBC. a, Copy-number profile of 362 tumors arranged by the proportion of the genome with a copy number >2. Vertical annotations on the left, definition of chr arms with red lines indicating centromeres (left); chromosomal arm-level events (>70% of the chromosomal arm affected) for diploid tumors (compared to two copies) and tumors with WGD (compared to four copies; left-middle); genome-wide G scores computed by GISTIC2, red and blue bars highlight chromosomal regions that are significantly enriched for focal gains or losses, respectively (right-middle); percentage of diploid tumors with copy-neutral LOH (right). **b**, Proportion of the genome with a copy number above two in NMIBC and MIBC (TCGA cohort). The dashed line indicates the cut-off (50%) used to define tumors with WGD. **c**, Kaplan–Meier plot of the probability of PFS as a function of ploidy status for 226 patients that did not receive BCG treatment during their disease course and had available genomic data (top; 206 diploid tumors and 20 tumors with WGD) and a subanalysis including only the 32 tumors classified as UROMOL2021 class 2a (bottom; 25 diploid tumors and 7 tumors with WGD). Statistically significant differences between groups were assessed by two-sided log-rank tests. **d**, Association between mutations in specific genes and WGD. The dashed line represents an FDR-adjusted *P* value of 0.05. CN, copy number.

https://doi.org/10.1038/s41588-024-02030-z



https://doi.org/10.1038/s41588-024-02030-z

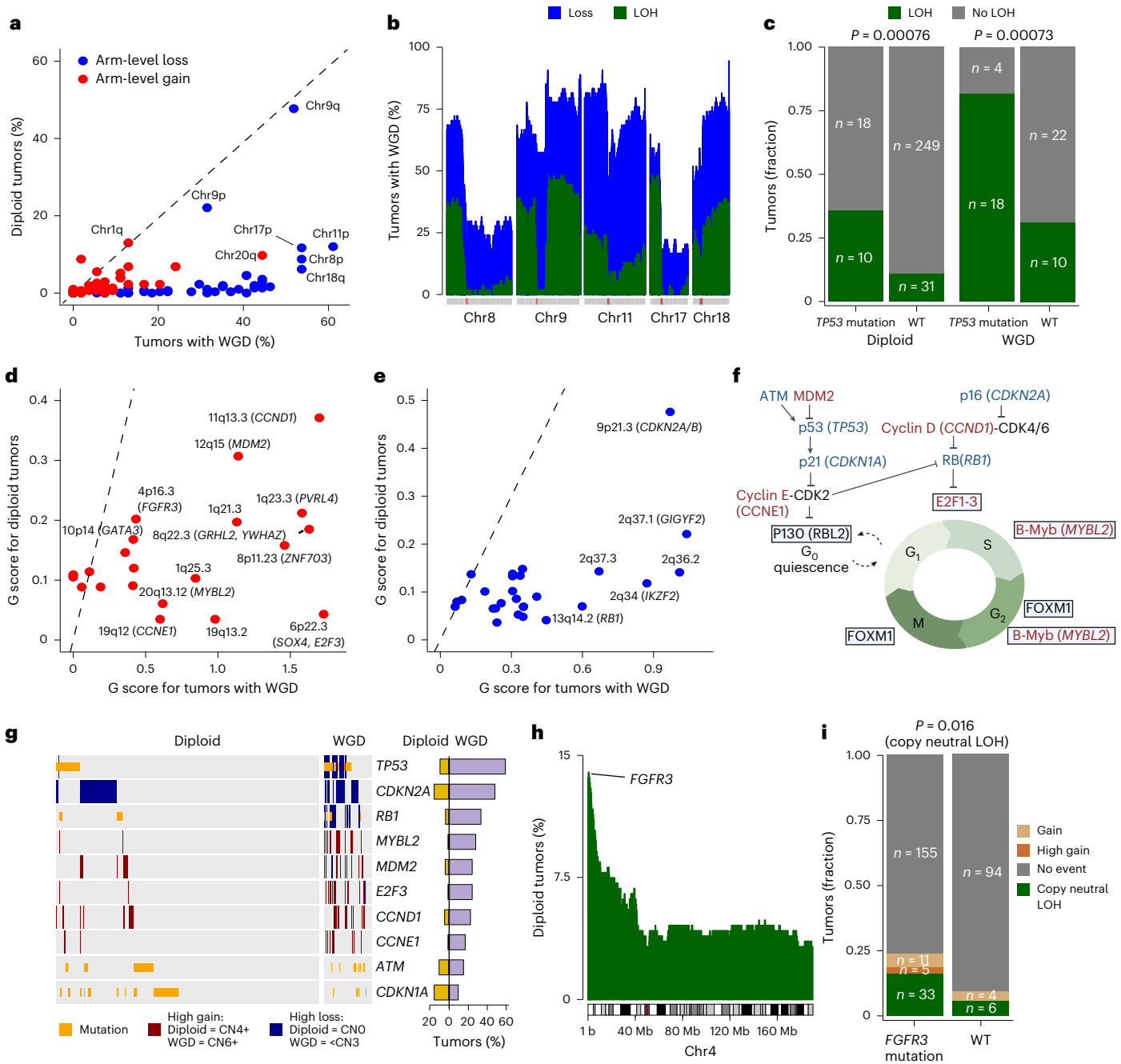


Fig. 3 | CNAs in diploid tumors and tumors with WGD. a, Percentage of diploid tumors and tumors with WGD with chr arm-level gains and losses when compared to two copies for diploid tumors and four copies for tumors with WGD. The dashed line shows a slope of 1. **b**, Percentage of tumors with WGD that have losses (copy number < 4) and/or LOH for selected chrs. **c**, Fraction of tumors with LOH in *TP53* stratified by *TP53* mutation status and ploidy. Statistically significant associations between groups were determined using the chi-square test. **d**, GISTIC2-computed G scores reflecting how significantly a region is affected by focal gains in tumors with WGD and diploid tumors. The dashed line shows a slope of 1. **e**, GISTIC2-computed G scores reflecting how significantly a region is affected by focal losses in tumors with WGD and diploid tumors. The dashed line shows a slope of 1. **f**, Central proteins in cell cycle regulation. Proteins encoded by genes found to be frequently affected by gain-of-function genomic alterations (mutations and/or gains) in the current study are colored red and proteins encoded by genes found to be frequently affected by loss-of-function genomic alterations (mutations and/or losses) in the current study are colored blue. **g**, Observed genomic alterations in genes involved in cell cycle regulation stratified by ploidy status. The type of genomic alteration (that is, mutation, gain or loss; left), and the percentage of diploid tumors (yellow) and tumors with WGD (purple; right) with a genomic alteration in the selected genes. **h**, Percentage of copy-neutral LOH on chr4 in diploid tumors. **i**, Fraction of copy-neutral LOH and gains affecting *FGFR3* in diploid tumors stratified by *FGFR3* mutation status. Statistically significant associations between groups were determined using the chi-square test.

1q23.3 (*PVRL4*/*NECTIN4*) and deletions of 9p21.3 (*CDKN2A*/*CDKN2B*), several regions in 2q, including 2q34 (*IKZF2*) and 2q37.1 (*GIGYF2*), and regions in 11q, such as 11q23.3 (*KMT2A*). Among diploid tumors, 4p16.3 (*FGFR3*) and 10p14 (*GATA3*) were some of the most significantly amplified regions, whereas among tumors with WGD, significant amplifications of 6p22.3 (*SOX4* and *E2F3*), 20q13.12 (*MYBL2*) and 19q12 (*CCNE1*) and loss of 13q14.2 (*RB1*) were observed (Fig. 3d,e). Loss of *RB1* was associated with mutations in *RB1* in both diploid tumors and tumors with WGD (Extended Data Fig. 5b). Of note, many of the genomic regions that were commonly altered in tumors with WGD but not in diploid tumors,

harbored genes involved in cell cycle regulation (*RB1*, *E2F3*, *CCNE1* and *MYBL2*; Fig. 3f,g). When also considering mutations in *TP53* and *RB1*, 93% of the tumors with WGD had at least one genomic alteration in a gene involved in cell cycle regulation (more than one for 85%; Fig. 3g). Among tumors with WGD, none of the genes significantly more mutated in tumors with WGD (Fig. 2d), and no genes involved in cell cycle regulation were significantly more mutated in *TP53* wild-type (WT) tumors compared to tumors with *TP53* mutations (Extended Data Fig. 5c). Consequently, we did not find any specific alteration that could explain WGD in *TP53* WT tumors.

Several of the highlighted focal CNAs were also identified when assessing CNAs in MIBC (Extended Data Fig. 6a–c). Interestingly, many of the regions mainly altered in NMIBCs with WGD were also frequent in diploid MIBCs. This may be explained by the high proportion of diploid tumors harboring *TP53* mutations in MIBC (37% compared to 9% in NMIBC). Indeed, *TP53*-mutated tumors with diploid genomes in MIBC were enriched for several of these regions, including 6p22 (*SOX4* and *E2F3*; 30%) and 13q14.2 (*RB1*, 43%; Extended Data Fig. 6d), suggesting that the specific CNAs enriched in tumors with WGD are not a direct result of WGD.

CNA assessment has earlier primarily focused on gains and losses, thereby missing copy-neutral LOH events. Here we assessed copy-neutral LOH (copy number 2/0) in diploid samples and identified one frequently affected region on 4p16, where *FGFR3* is located (14% of samples; Figs. 2a and 3h). Copy-neutral LOH in *FGFR3* was more frequent in *FGFR3*-mutated tumors compared to WT tumors ($P = 0.016$; Fig. 3i), indicating that losing the *FGFR3* WT allele may be an early driver of tumor development.

Pathway enrichment analysis in tumors with WGD

We used the overlapping RNA-seq data ($n = 312$) to assess the enrichment of specific pathways in tumors with WGD. We identified significant upregulation of cell-cycle-associated gene sets, especially cell cycle checkpoints ($P = 2 \times 10^{-39}$; Fig. 4a and Extended Data Fig. 7a), which is in line with the frequent genomic alterations observed in central cell cycle regulators (Fig. 3f,g). Regulon analysis of essential cell cycle regulators further supported the overall high cell cycle activity in tumors with WGD and showed that diploid tumors with a disrupted p53 pathway (*TP53* mutated and/or *MDM2* amplified) likewise had a high cell cycle activity (Fig. 4b and Extended Data Fig. 7b). Activity of the G_0 (quiescence)-related *RB1* ($P = 3.2 \times 10^{-9}$) and *RBL2* ($P = 8.3 \times 10^{-17}$) regulons was higher in diploid, p53 pathway WT tumors, whereas both diploid tumors and tumors with WGD with a disrupted p53 pathway had higher regulon activity of G_1 to M factors, including *E2F1–E2F3*, *MYBL2* and *FOXM1* (all $P < 0.001$). DNA repair pathways, including the homology-directed repair pathway ($P = 1.5 \times 10^{-21}$), were additionally among the most upregulated pathways in tumors with WGD (Extended Data Fig. 7a) along with a higher expression of *BRCA1* ($P = 1.5 \times 10^{-11}$), *H2AX* ($P = 6.3 \times 10^{-6}$) and *RAD51* ($P = 8.7 \times 10^{-10}$; Fig. 4a and Extended Data Fig. 7c). Interestingly, it seems that a disruption of the p53 pathway and WGD status independently had the same effect on the cell cycle regulon activity and that these effects were even higher when both were present (Fig. 4b).

The most downregulated pathways in tumors with WGD compared to diploid tumors were related to translation (Extended Data Fig. 7a), possibly reflecting a response to the cell stress (including DNA damage) induced by the impaired cell cycle regulation[25]. One exception was *EIF4G1*, which showed a higher expression in tumors with WGD ($P = 2.2 \times 10^{-7}$; Fig. 4a). eIF4G1 is a part of the eIF4F complex that is involved in cap-dependent translation initiation[26] and has been shown to be essential for cancer cells with a high degree of DNA damage[27].

Immune landscape of tumors with WGD

Immune cells seek to recognize and eliminate altered cells to hinder tumor development. The high number of genomic alterations in tumors with WGD should, in principle, make them more recognizable; however, the development of these highly altered tumors and their propensity to progress to MIBC (Fig. 2c) suggest that the tumors have developed mechanisms of immune evasion.

Deconvolution of RNA-seq data ($n = 312$) showed no significant differences in the inferred CD4 T cell ($P = 0.2$) or cytotoxic T lymphocyte (CTL; $P = 0.16$) infiltration between tumors with WGD and diploid tumors (Fig. 4c). However, tumors with WGD showed a higher score for T cell exhaustion (RNA-based estimation; $P = 0.011$; Fig. 4d). Immunohistochemistry (IHC) assessment of the T cell exhaustion marker PD-1 in a subset of samples ($n = 149$) also showed a higher proportion of PD-1-positive cells in tumors with WGD ($P = 0.00051$; Fig. 4e). Deconvolution analysis further revealed that tumors with WGD were enriched for myeloid cell types involved in the innate immune response, particularly neutrophils ($P = 5.1 \times 10^{-7}$), but also mast cells ($P = 0.0019$) and dendritic cells ($P = 0.03$; Fig. 4c). Tumors with WGD were frequently classified as the inflamed UROMOL2021 class 2b (Extended Data Fig. 4h); however, the association between neutrophil infiltration and WGD was also significant within class 2b ($P = 0.0012$; Extended Data Fig. 7d).

Studying the antigen processing and presentation machinery[28], tumors with WGD generally had a higher expression of genes in the major histocompatibility complex (MHC)-I peptide loading complex but a lower expression of the peptide-trimming genes *ERAP1* ($P = 2.1 \times 10^{-7}$) and *ERAP2* ($P = 0.0027$; Fig. 4f). A low expression of *ERAP1*/*ERAP2* in tumors with WGD was also observed in MIBC (TCGA[16]; *ERAP1*, $P = 0.0054$; *ERAP2*, $P = 0.0098$) and a low expression has also been reported in lung cancers with a high degree of genomic instability[29]. Furthermore, we found a lower expression of *TAPBPL* ($P = 2.9 \times 10^{-5}$) and *UGGT1* ($P = 0.0011$), which control the quality of peptides bound to MHC-I (Fig. 4f). A decreased activity of these functions may result in an altered repertoire of neoantigens presented on MHC-I. Only 3% of tumors had LOH of a human leukocyte antigen (HLA) locus, which may furthermore alter antigen presentation[30]. However, HLA LOH was enriched in tumors with WGD ($P = 0.002$; Fig. 4g).

Integrative clustering of multi-omics data

We delineated the biological framework of NMIBC using integrative clustering[31] to investigate whether the integration of multi-omics data provides additional biological and prognostic value beyond single-layer stratifications of tumors. For 230 tumors with overlapping data layers, we performed joint clustering of somatic mutations, CNAs and gene expression data and identified four 'iClusters' denoted iClus1–4 (Fig. 5a). The iClusters overlapped with the UROMOL2021 classes ($P = 1.9 \times 10^{-42}$) and WGD status ($P = 9.5 \times 10^{-17}$). iClus4 included a mixture of UROMOL2021 class 2a and 2b tumors, including 70% (26/37) of all class 2a tumors and 38% (29/77) of all class 2b tumors, and 93% (26/28) of tumors with WGD (Fig. 5a and Extended Data Fig. 8a). Accordingly, the iClusters were associated with PFS ($P = 0.0021$; Fig. 5b) with iClus4 tumors having the highest risk of progression, independent of tumor stage (HR = 4.6, 95% CI = 1.4–14.6, $P = 0.010$) and UROMOL2021 classification (HR = 5.5, 95% CI = 1.6–18.6, $P = 0.006$; Supplementary Table 9). The iClus4 tumors had a higher class 2a weighted in silico pathology (WISP) weight (estimate of the proportion of different subtypes within bulk tumor samples[6]; class 2a, $P = 5 \times 10^{-4}$; class 2b, $P = 9.4 \times 10^{-6}$; Extended Data Fig. 8b), a higher TMB ($P = 9.5 \times 10^{-13}$; Extended Data Fig. 8c,d) and were more genomically unstable, also when assessing diploid tumors only ($P = 1 \times 10^{-9}$; Fig. 5c).

A higher RNA-based immune infiltration score was observed in iClus2 and iClus4 ($P = 6 \times 10^{-18}$; Extended Data Fig. 8e). In concordance, a higher relative fraction of T cells was observed in iClus2 and iClus4 when using the tool TcellExTRECT[32] on tumor WES data ($P = 3 \times 10^{-6}$; Extended Data Fig. 8f). We did not observe a difference in the RNA-based scores for CD4 T cells ($P = 0.38$) and CTLs ($P = 0.43$) between iClus2 and iClus4; however, similar to the analyses of WGD tumors, tumors within iClus4 showed higher scores for mast cells ($P = 0.011$) and neutrophils ($P = 0.0014$) compared with tumors in iClus2 (Fig. 5d).

https://doi.org/10.1038/s41588-024-02030-z

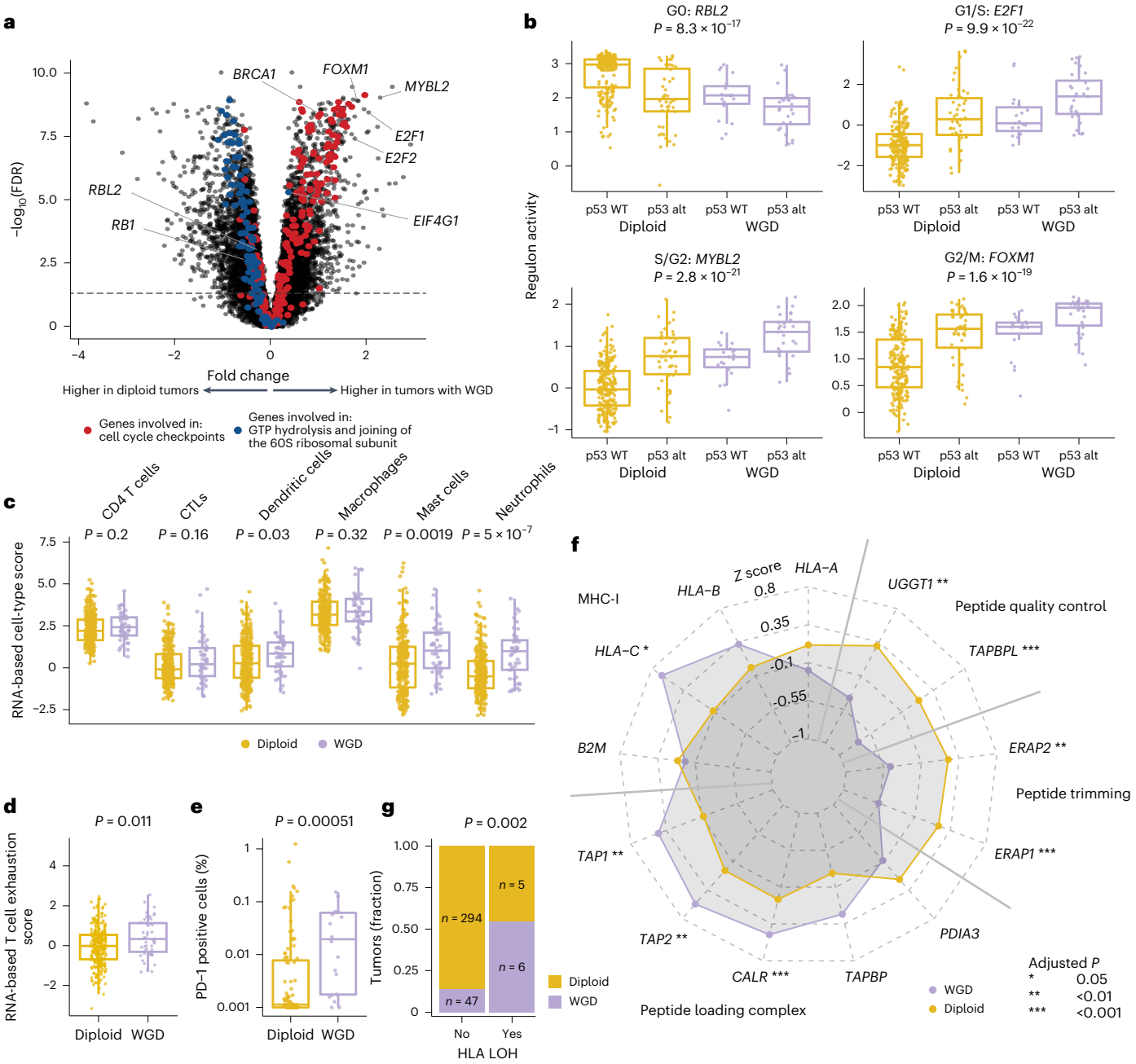


Fig. 4 | Gene expression and immunological features associated with WGD. **a**, Volcano plot of differentially expressed genes between diploid tumors and tumors with WGD. **b**, Regulon activity of genes involved in different steps of the cell cycle, from G0 (senescence) to M (mitosis), stratified by ploidy and involvement of the p53 pathway (*TP53* mutations and/or *MDM2* gain; 219 diploid, WT tumors; 46 diploid, p53 pathway-altered (p53 alt) tumors; 19 WGD, WT tumors; 28 WGD, p53 alt tumors). Statistically significant differences between groups were determined using Kruskal–Wallis tests. **c**, Estimated RNA-based cell-type scores stratified by ploidy status (265 diploid tumors and 47 tumors with WGD). Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests. **d**, Estimated RNA-based T cell exhaustion scores stratified by ploidy status (265 diploid tumors and 47 tumors with WGD). Statistically significant differences between groups were determined using a two-sided Wilcoxon rank sum test. **e**, Percentage of PD-1 positive cells in the cancer cell area (excluding stromal areas) assessed by IHC stratified by ploidy status (103 diploid tumors and 20 tumors with WGD). Statistically significant differences between groups were determined using a two-sided Wilcoxon rank sum test. **f**, Spider plot showing the median *z* score of gene expression for genes involved in antigen processing and presentation stratified by ploidy status. Asterisks indicate genes with a significantly different gene expression between diploid tumors and tumors with WGD. Statistically significant differences between groups were determined using a two-sided Wilcoxon rank sum test. *P* values were adjusted using the FDR approach. **g**, Fraction of tumors with HLA LOH in diploid tumors and tumors with WGD. Statistically significant associations between variables were determined using Fisher's exact test. Boxplots represent the median and lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

Given this enrichment for myeloid cells in iClus4 and the poor outcome observed for patients within iClus4, we hypothesized that iClus4 was characterized by a dysfunctional immune landscape with more innate activity compared to adaptive antitumor activity. To assess this, we investigated an adaptive-to-innate immune ratio[33] and found a lower score within iClus4 compared with iClus2, indicating that iClus4 tumors have a higher innate component ($P = 0.02$; Extended Data Fig. 8g). Furthermore, we observed higher T cell exhaustion[8] (RNA-based;

https://doi.org/10.1038/s41588-024-02030-z

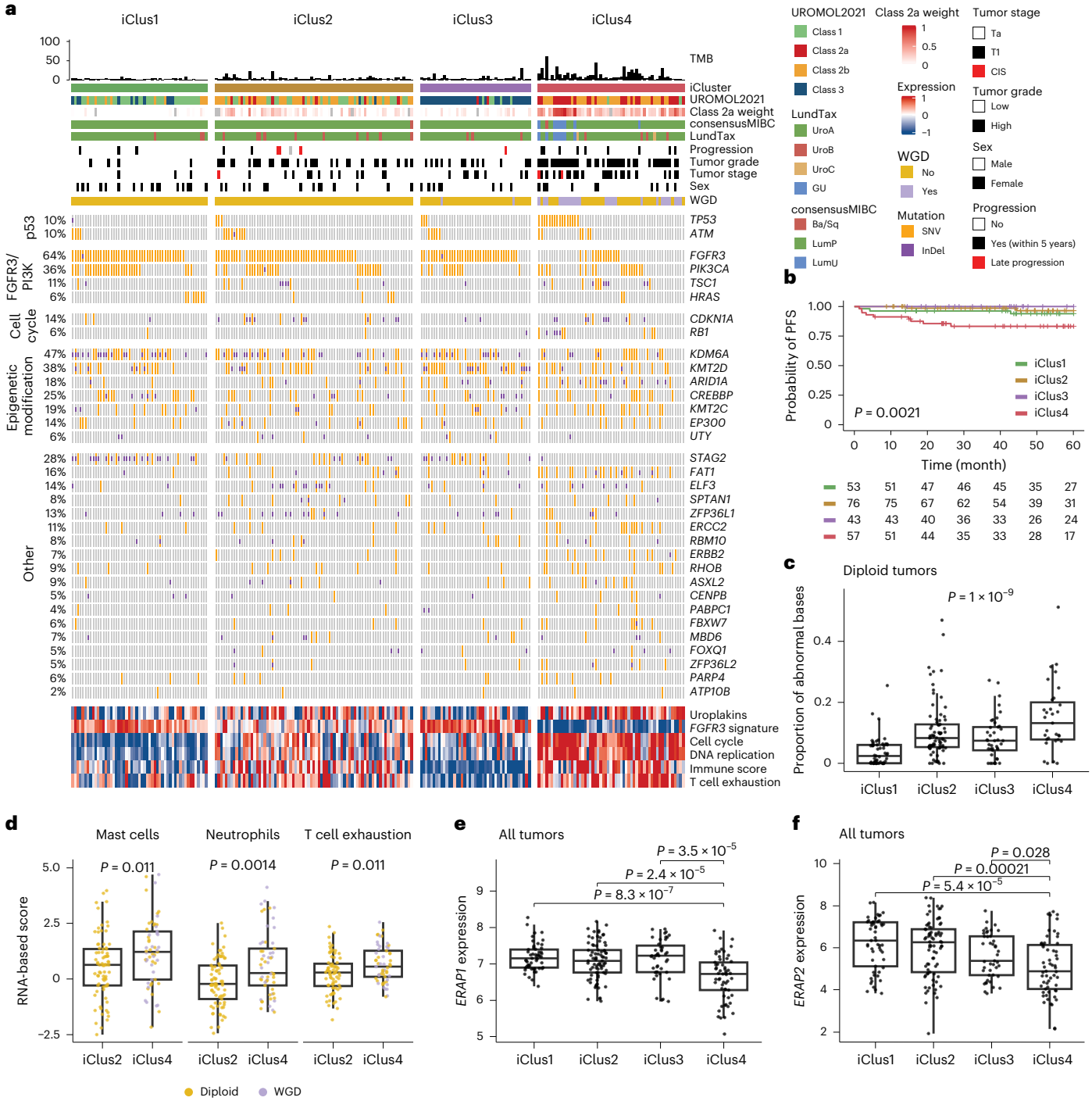


Fig. 5 | Integrative clustering analysis. a, Oncoplot of the 33 genes displayed in Fig. 1a for 230 tumors with somatic mutation, CNA and gene expression data. Tumors are stratified by the four iClusters. The UROMOL2021 class 2a weight was estimated using the WISP tool. The lower panel shows the scaled mean expression values of selected gene signatures. **b**, Kaplan–Meier plot of the probability of PFS as a function of the iClusters for 229 patients (iClus1, $n = 53$; iClus2, $n = 76$; iClus3, $n = 43$; iClus4, $n = 57$). Statistically significant differences between groups were determined using a two-sided log-rank test. **c**, Proportion of bases in an abnormal state for diploid tumors stratified by iClusters (iClus1, $n = 53$; iClus2, $n = 77$; iClus3, $n = 41$; iClus4, $n = 31$). Abnormal was defined as different from a copy number of two or an imbalanced copy number of two. Statistically significant differences between groups were determined using the Kruskal–Wallis test. **d**, Estimated RNA-based cell-type scores stratified by iClusters (iClus2, $n = 77$; iClus4, $n = 57$). Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests. **e**,**f**. *ERAP1* (**e**) and *ERAP2* (**f**) gene expression stratified by iClusters (iClus1, $n = 53$; iClus2, $n = 77$; iClus3, $n = 43$; iClus4, $n = 57$). Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests. Boxplots represent the median and lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

$P = 0.011$; Fig. 5d) and lower expression of *ERAP1* ($P < 0.001$) and *ERAP2* ($P < 0.05$) in iClus4 (Fig. 5e,f), indicating that T cell exhaustion and impaired peptide presentation may be key drivers of tumor aggressiveness and poor outcome. Notably, neither the iClusters nor progression was associated with lower levels of T cells or low T cell receptor (TCR) diversity in peripheral blood (determined using germline WES data), indicating a generally similar immune health state in these patients (Extended Data Fig. 8h–k).

See Supplementary Note for additional results (Extended Data Figs. 9 and 10 and Supplementary Tables 10–12).

Discussion

In summary, this study strengthens our understanding of the genomic framework underlying disease aggressiveness and different transcriptomic profiles within NMIBC. We demonstrated that NMIBC is a genomically heterogeneous disease with a subset of tumors showing high genomic instability. We uncovered that 15% of tumors had most likely undergone WGD, which was associated with an increased risk of progression. This is in line with previous studies demonstrating a link between genomic instability and poor outcomes in NMIBC[6,13,34] and the high occurrence of WGD in MIBC (58%).

WGD has previously been associated with *TP53* mutations across cancer types[16,23]. A recent study on a pancreatic cancer mouse model showed that *Trp53* (encoding p53) mutations lead to WGD in an ordered process including the loss of the WT *Trp53* allele[35]. Here we showed that *TP53* mutations were strongly associated with WGD and that *TP53*-mutated tumors with WGD predominately had *TP53* LOH. Looking beyond *TP53* mutations (as 60% of the tumors with WGD were *TP53* WT), more than 90% of the tumors with WGD had at least one genomic alteration in a central cell cycle regulator. Indeed, replicative stress caused by, for example, genomic alterations in cell cycle regulators has been shown to induce WGD in p53-proficient cell lines[36]. However, whether increased cell cycle activity in these tumors is a cause or a consequence of WGD is not clear.

Integrative clustering of multi-omics data identified four iClusters that captured the biological and molecular framework of NMIBC beyond single-layer stratifications of tumors. Overall, the iClusters underlined that both genomic instability, transcriptomic subtype scores and the functional status of infiltrating immune cells are important features to identify clinically high-risk tumors. iClus4 constituted a group of highly aggressive NMIBCs and included nearly all tumors with WGD as well as diploid tumors with high genomic instability. The tumors were of UROMOL2021 classes 2a and 2b, and had high expression of genes related to cell cycle activity. The aggressive nature of iClus4 tumors, and tumors with WGD in general, may indicate that the tumors have managed to evade immune recognition and elimination despite their high neoantigen load (high mutational burden). Our results suggest that one mechanism of evasion could be a change in the repertoire of the immunopeptidome presented on MHC-I caused by a low expression of especially *ERAP1*. A decreased *ERAP1* expression has been reported in several cancer types[37] and has been associated with loss of p53 and genomic instability[29,38]. Additionally, *ERAP1* is believed to be central for inducing a strong CTL response[39].

iClus4 tumors and tumors with WGD showed an altered immune cell composition with, for example, high levels of neutrophils. Neutrophils have a dual role in the tumor microenvironment with both protumorigenic and antitumorigenic activities[40]. Although the estimated levels of CTLs were similar between molecular subgroups, the level of T cell exhaustion was higher in iClus4 and tumors with WGD, suggesting impaired immune functionality as another potential mechanism of immune evasion. The high neoantigen load, likely caused by genomic instability in these tumors, may drive T cell exhaustion as has previously been described[8,41]. Moreover, T cell exhaustion has been associated with disease aggressiveness in BCG-treated patients with NMIBC[8], and here we document that it may have a role in tumor aggressiveness in NMIBC overall, suggesting a need for identifying patients with aggressive tumors (iClus4 and/or WGD) for improved treatment stratification. Several clinical trials are currently ongoing investigating the impact of immune checkpoint inhibitors in NMIBC[42], and this study demonstrates the biological rationale for this oncological treatment.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-02030-z.

References

1. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Dyrskjøt, L. et al. Bladder cancer. *Nat. Rev. Dis. Primers* **9**, 58 (2023).
3. Babjuk, M. et al. European Association of Urology Guidelines on non-muscle-invasive bladder cancer (Ta, T1, and carcinoma in situ). *Eur. Urol.* **81**, 75–94 (2022).
4. Dyrskjøt, L. & Ingersoll, M. A. Biology of nonmuscle-invasive bladder cancer: pathology, genomic implications, and immunology. *Curr. Opin. Urol.* **28**, 598–603 (2018).
5. Kamoun, A. et al. A consensus molecular classification of muscle-invasive bladder cancer. *Eur. Urol.* **77**, 420–433 (2020).
6. Lindskrog, S. V. et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat. Commun.* **12**, 2301 (2021).
7. Hedegaard, J. et al. Comprehensive transcriptional analysis of early-stage urothelial carcinoma. *Cancer Cell* **30**, 27–42 (2016).
8. Strandgaard, T. et al. Elevated T-cell exhaustion and urinary tumor DNA levels are associated with Bacillus Calmette–Guérin failure in patients with non–muscle-invasive bladder cancer. *Eur. Urol.* **82**, 646–656 (2022).
9. De Jong, F. C. et al. Non-muscle-invasive bladder cancer molecular subtypes predict differential response to intravesical Bacillus Calmette–Guérin. *Sci. Transl. Med.* **15**, eabn4118 (2023).
10. Van Rhijn, B. W. G. et al. Novel fibroblast growth factor receptor 3 (FGFR3) mutations in bladder cancer previously identified in non-lethal skeletal disorders. *Eur. J. Hum. Genet.* **10**, 819–824 (2002).
11. López-Knowles, E. et al. PIK3CA mutations are an early genetic alteration associated with FGFR3 mutations in superficial papillary bladder tumors. *Cancer Res.* **66**, 7401–7404 (2006).
12. Balbás-Martínez, C. et al. Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. *Nat. Genet.* **45**, 1464–1469 (2013).
13. Hurst, C. D. et al. Genomic subtypes of non-invasive bladder cancer with distinct metabolic profile and female gender bias in KDM6A mutation frequency. *Cancer Cell* **32**, 701–715.e7 (2017).
14. Hurst, C. D. et al. Stage-stratified molecular profiling of non-muscle-invasive bladder cancer enhances biological, clinical, and therapeutic insight. *Cell Rep. Med.* **2**, 100472 (2021).
15. Goel, A. et al. Combined exome and transcriptome sequencing of non-muscle-invasive bladder cancer: associations between genomic changes, expression subtypes, and clinical outcomes. *Genome Med.* **14**, 59 (2022).
16. Robertson, A. G. et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171**, 540–556.e25 (2017); erratum **174**, 1033 (2018).
17. Jebar, A. H. et al. FGFR3 and Ras gene mutations are mutually exclusive genetic events in urothelial cell carcinoma. *Oncogene* **24**, 5218–5225 (2005).
18. Taylor, C. F., Platt, F. M., Hurst, C. D., Thygesen, H. H. & Knowles, M. A. Frequent inactivating mutations of STAG2 in bladder cancer are associated with low tumour grade and stage and inversely related to chromosomal copy number changes. *Hum. Mol. Genet.* **23**, 1964–1974 (2014).
19. Sondka, Z. et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
20. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).

21. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
22. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
23. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
24. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
25. Bartkova, J. et al. Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444**, 633–637 (2006).
26. Prévôt, D., Darlix, J.-L. & Ohlmann, T. Conducting the initiation of protein synthesis: the role of eIF4G. *Biol. Cell* **95**, 141–156 (2003).
27. Badura, M., Braunstein, S., Zavadil, J. & Schneider, R. J. DNA damage and eIF4G1 in breast cancer cells reprogram translation for survival and DNA repair mRNAs. *Proc. Natl Acad. Sci. USA* **109**, 18767–18772 (2012).
28. Jhunjhunwala, S., Hammer, C. & Delamarre, L. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nat. Rev. Cancer* **21**, 298–312 (2021).
29. Tripathi, R., Modur, V., Senovilla, L., Kroemer, G. & Komurov, K. Suppression of tumor antigen presentation during aneuploid tumor evolution contributes to immune evasion. *Oncoimmunology* **8**, 1657374 (2019).
30. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271.e11 (2017).
31. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
32. Bentham, R. et al. Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature* **597**, 555–560 (2021).
33. Ahrenfeldt, J. et al. The ratio of adaptive to innate immune cells differs between genders and associates with improved prognosis and response to immunotherapy. *PLoS ONE* **18**, e0281375 (2023).
34. Norming, U., Tribukait, B., Nyman, C. R., Nilsson, B. & Wang, N. Prognostic significance of mucosal aneuploidy in stage Ta/T1 grade 3 carcinoma of the bladder. *J. Urol.* **148**, 1420–1426 (1992).
35. Baslan, T. et al. Ordered and deterministic cancer genome evolution after p53 loss. *Nature* **608**, 795–802 (2022).
36. Zeng, J., Hills, S. A., Ozono, E. & Diffley, J. F. X. Cyclin E-induced replicative stress drives p53-dependent whole-genome duplication. *Cell* **186**, 528–542.e14 (2023).
37. Babaie, F. et al. The roles of ERAP1 and ERAP2 in autoimmunity and cancer immunity: new insights and perspective. *Mol. Immunol.* **121**, 7–19 (2020).
38. Wang, B., Niu, D., Lai, L. & Ren, E. C. p53 increases MHC class I expression by upregulating the endoplasmic reticulum aminopeptidase ERAP1. *Nat. Commun.* **4**, 2359 (2013).
39. Yan, J. et al. In vivo role of ER-associated peptidase activity in tailoring peptides for presentation by MHC class Ia and class Ib molecules. *J. Exp. Med.* **203**, 647–659 (2006).
40. Hedrick, C. C. & Malanchi, I. Neutrophils in cancer: heterogeneous and multifaceted. *Nat. Rev. Immunol.* **22**, 173–187 (2022).
41. Dolina, J. S., Van Braeckel-Budimir, N., Thomas, G. D. & Salek-Ardakani, S. $CD8^+$ T cell exhaustion in cancer. *Front. Immunol.* **12**, 715234 (2021).
42. Valenza, C. et al. Emerging treatment landscape of non-muscle invasive bladder cancer. *Expert Opin. Biol. Ther.* **22**, 717–734 (2022).

[1]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [3]Department of Urology, Herlev Hospital, Copenhagen University, Copenhagen, Denmark. [4]Department of Pathology, Aalborg University Hospital, Aalborg, Denmark. [5]Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Comprehensive Cancer Center EMN, Erlangen, Germany. [6]Department of Urology, Jena University Hospital, Jena, Germany. [7]Department of Urology, University Hospital Essen, Essen, Germany. [8]Department of Urology, Technical University of Munich, Klinikum rechts der Isar, Munich, Germany. [9]Department of Surgical Sciences, Uppsala University, Uppsala, Sweden. [10]Department of Urology and Pediatric Urology, University Hospital Erlangen, Friedrich-Alexander-University Erlangen–Nuremberg, Erlangen, Germany. [11]Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, the Netherlands. [12]Department of Urology, Amphia Ziekenhuis, Breda, the Netherlands. [13]Department of Pathology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, the Netherlands. [14]Department of Urology and Martini-Klinik, University of Hamburg–Eppendorf, Hamburg, Germany. [15]Institute of Medical and Clinical Biochemistry, Center for Redox Medicine, Faculty of Medicine, University of Belgrade, Belgrade, Serbia. [16]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Center (CNIO) and CIBERONC, Madrid, Spain. [17]Department of Urology, Aarhus University Hospital, Aarhus, Denmark. [18]Epithelial Carcinogenesis Group, Spanish National Cancer Research Center (CNIO) and CIBERONC, Madrid, Spain. [19]Medicine and Life Sciences Department, Universitat Pompeu Fabra, Barcelona, Spain. ✉e-mail: lars@clin.au.dk

UROMOL Consortium

Gregers G. Hermann[3], Karin Mogensen[3], Astrid C. Petersen[4], Arndt Hartmann[5], Marc-Oliver Grimm[6], Marcus Horstmann[7], Roman Nawroth[8], Ulrika Segersten[9], Danijel Sikic[10], Kim E. M. van Kessel[11,12], Ellen C. Zwarthoff[13], Tobias Maurer[14], Tatjana Simic[15], Per-Uno Malmström[9], Núria Malats[16], Francisco X. Real[18,19] & Lars Dyrskjøt[1,2]

Methods

Patients

Patients with NMIBC were included in the European UROMOL project ($n = 296$) and from Aarhus University Hospital, Denmark ($n = 142$). Detailed information on the UROMOL cohort can be found in refs. 6,7, and information on the Aarhus cohort is described in ref. 8. This study complies with all relevant ethical regulations. Informed written consent to take part in research projects was obtained from all patients, and all ethical regulations for working with human participants were followed according to national guidelines. The study was approved by the Central Denmark Region Committees on Biomedical Research Ethics (1994/2920; Skejby, Aalborg, Frederiksberg); the Danish National Committee on Health Research Ethics (1906019;1708266); the Ethics Committee of the University Hospital Erlangen (3755); the ethics committee of the Technical University of Munich (2792/10); Medical Ethics Committee of Erasmus MC (MEC 168.922/1998/55, Rotterdam); the Uppsala Region Committee on Biomedical Research Ethics (2008/252); the Ethical Committee of Faculty of Medicine, University of Belgrade (440/VI-7); the Ethics Committee (Comissão de Ética para a Investigação Clínica) of Institut Municipal d'Assistència Sanitària/Hospital del Mar (2008/3296/I); and the ethics committee of the University Hospital Jena (4774-4/16). Patients did not receive any compensation for participation. A summary of the clinical characteristics of the cohort is provided in Supplementary Table 1.

DNA extraction

Tumor tissue was collected fresh from resection in each clinical center, embedded in Tissue-Tek O.C.T. and snap-frozen in liquid nitrogen before storage at −80 °C. Before DNA extraction, the cancer cell content was assessed by evaluation of hematoxylin, and eosin-stained sections were cut immediately before and after those used for extraction. The median cancer cell content was 89%. DNA was extracted from sections of Tissue-Tek O.C.T Compound using a Puregene DNA purification kit (Gentra Systems). Leukocyte DNA was extracted from the buffy coat of all patients using the QIAsymphony DSP DNA midi kit (Qiagen). DNA concentration was measured using a Qubit Fluorometer (Thermo Fisher Scientific).

WES

Libraries of tumor and matching germline DNA were prepared using the Twist Library preparation EF kit (Twist Bioscience) with an input of 50 ng DNA. For buffy-coat DNA and DNA from fresh frozen tumors, a 10-min fragmentation time was used. The protocol was optimized using the xGEN UDI-UMI Adapters (Integrated DNA Technologies) with seven cycles of PCR postligation and seven cycles of PCR postcapture. Libraries for the Aarhus cohort were also prepared using the Twist Library Preparation EF Kit (Twist Bioscience), xGEN UDI-UMI Adapters, and the input of 50 ng DNA, however, using a 16-min fragmentation time and ten cycles of PCR postligation and seven cycles of PCR postcapture. Libraries for the few FFPE DNA samples included in the Aarhus cohort were prepared using the same conditions. Of note, a small number of libraries were made using the Illumina TruSeq DNA Kit and NimbleGen SeqCap EZ v3.0 capture[43].

Mutation calling

WES reads were trimmed for adaptor sequences and low-quality bases using cutadapt (v3.7)[44] before being mapped to hg38 using bwa (v0.7.17)[45]. Picard tool MarkDuplicates (v2.27.00) was used to mark duplicate reads. Mutect2 (v2.2)[46] was used to call somatic mutations and InDels. Strelka (v2.9.10)[47] was subsequently run and mutations filtered by Mutect2 but detected with a high confidence by Strelka were reintroduced. Only the genomic alterations within the WES target were kept and functionally annotated using SnpEff (v.4.3t)[48].

Identification of significantly mutated genes

MutSigCV (v1)[49] was used to identify genes with a significantly higher frequency of nonsynonymous mutations than what would be expected by chance based on a computed background model. Before the analysis, variant positions were lifted from hg38 to hg19 using the R packages GenomicRanges (v1.44.0) and rtracklayer (v1.52.00). Only genes with a q value below 0.05 and a mutation frequency above 5% were included (Supplementary Table 2).

Mutation signatures

A matrix summarizing the type of mutations and their trinucleotide context for each sample was generated, and the sigProfiler framework[50], including sigprofilerassignment (v0.0.11), sigprofilerextractor (v1.1.0), sigprofilermatrixgenerator (v1.2.9) and sigprofilerplotting (v1.2.2), was used to extract de novo mutational signatures. These de novo signatures were then decomposed into a set of the following nine known COSMIC SBS 96 signatures[19]: SBS1, SBS2, SBS4, SBS5, SBS10b, SBS13, SBS15, SBS29 and SBS31. Only tumors with at least 100 mutations were included in the decomposition of signatures.

CCF

For each mutation, the estimated CCF in the corresponding tumor was used to infer the fraction of cancer cells carrying the mutation. To determine CCF, variant allele frequency (VAF) was integrated with tumor purity estimated using PurBayes (v1.3)[51] and the local copy number at the mutated site, as described in ref. 52, according to the following formula:

$$\mathrm{CCF} = \mathrm{VAF} \times 1/\mathrm{purity} \times ((\mathrm{purity} \times \mathrm{CNt}) + \mathrm{CNn} \times (1 - \mathrm{purity}))\,,$$

where CNt is the local copy number at the mutation and CNn is the diploid copy-number state. CCFs >1 were classified as having a CCF of 1.

sWGS

sWGS was performed using aliquots of libraries set aside before exome capture ('WES'). All libraries were paired-end sequenced (2 × 150 bp) on the NovaSeq 6000 platform (Illumina) using S4 flow cells. Before sequencing, all runs were calibrated on a MiSeq Nano (2 × 150 bp) to obtain even coverage.

Copy-number estimation

The segmentation and initial ploidy call were estimated from the sWGS data by ichorCNA (v0.3.2)[21]. ichorCNA was run with a bin size of 50 kb, and bin count normalization was independently conducted for tumors from male and female patients. For normalization, we used germline samples from 15 females and 15 males as references (mean coverage = 2.45×, range = 0.38–4.30×).

The genome-wide and uniform coverage of sWGS offers a more precise identification of copy-number breakpoints (segmentation) compared to estimations based on WES data. However, sWGS lacks the capability to provide information on allele-specific copy-number distribution, preventing the identification of unbalanced copy-number regions. This is particularly important for accurate ploidy estimation, which can be guided by the fact that balanced regions should exhibit an even copy number, typically a copy number of two or four. To address this limitation, overlapping WES data (tumor and germline) were used to assess allele-specific copy-number distribution using Battenberg software (v2.2.10)[22]. Information on unbalanced regions from Battenberg was integrated with the copy-number estimations from ichorCNA. Subsequently, each sample underwent manual inspection by two individuals to ensure that only even copy-number states contained balanced segments. If this criterion was not met, manual adjustment of the ploidy was made.

The copy-number calls from Battenberg guided the ploidy estimation; however, there were cases where deviations occurred (Supplementary Fig. 1a). U0091, for example, was classified as near tetraploid by Battenberg but was considered near diploid in our analysis. This sample exhibited one balanced copy-number state surrounded by

two smaller unbalanced copy-number states (Supplementary Fig. 1b,c). We did not find compelling evidence that these three states should not be a copy number of one, two and three. Another example is U0457, which was classified as polyploid in our analysis but near diploid by Battenberg. This sample contained two copy-number states with balanced copy-number segments that were separated by an unbalanced copy-number state (Supplementary Fig. 1d,e). In line with the previously mentioned concept of balanced copy-number segments having an even copy number, the two balanced copy-number states must have at least a copy number of two and four, respectively. This results in a ploidy estimate of 3.6.

For 19 samples, it was not possible to reach a confident copy-number estimation, leading to their exclusion from further analysis.

Definition of WGD

In accordance with ref. 23, tumors were classified as having WGD if more than 50% of the genome had a copy number above two ($n = 54$; 15%; Fig. 2b). A large fraction of the genome with a copy number above two could also be explained by frequent independent amplifications. In that case, it would be expected that segments with a copy number of four would be unbalanced (n major allele = 3 and n minor allele = 1) as opposed to tumors with WGD, where segments with a copy number of four mainly should be balanced (n major and minor allele = 2). Indeed, segments with a copy number of four in tumors classified as having WGD were predominantly balanced (B-allele frequency located around 0.5), as opposed to segments with a copy number of four in tumors classified as being diploid (Extended Data Fig. 4a,b).

Identification of significant CNAs

GISTIC2.0 (ref. 24) was used to identify regions enriched for focal CNAs. GISTIC2.0 computes G scores across the genome that reflect both the frequency of events in a region and the amplitude of these events. GISTIC2.0 was first run for the whole cohort to identify significantly altered regions and then separately for diploid tumors and tumors with WGD to compare the degree of focal CNAs (G score) between the two groups along the genome.

Long-read sequencing

Sequencing libraries were prepared according to the Nanopore protocol Genomic DNA by Ligation (SQK-LSK114) with an input of 1.5 µg of DNA and incubation times for end-prep at both 20 °C and 60 °C increased to 10 min. We used an input of 78–158 ng DNA for the loading mix with the aim of getting an equal loading input of 15 fmol. To increase the output, we performed a flow cell wash and loaded the same libraries after both 1 and 2 days of sequencing. The sequencing produced between 62 and 139 Gb of raw data per sample with an N50 of 13–17 kb for the blood samples and 7–16 kb for the tissue samples. The data were base called with Guppy (v6) with the high-accuracy model, read splitting and 5mC detection.

Fastq files were mapped using minimap2 (v2.24)[53]. Very low mapping quality reads and unmapped reads were filtered out using SAMtools (v1.15.1)[54]. Somatic structural variants were called using SAVANA (v0.2.3)[55]. We used the strict VCF file for the rest of the analysis. Plots were produced using the CIRCOS package (v1.3.5)[56].

TCGA data analysis

To compare the genomic landscape in NMIBC with MIBC, we studied data from MIBCs from the TCGA cohort[16]. Mutation annotation format files containing annotated variant calls and clinical information were imported for using the R package maftools[57]. Allele-specific copy-number segments from 412 tumors were retrieved from the Genomic Data Commons Data Portal. These data were computed by ASCAT2, based on Affymetrix SNP6.0 array data[16].

RNA-seq

Total RNA-seq data from 414 tumors were reported in refs. 6,8. In short, RNA was extracted from tissue using RNeasy Mini and Micro Kits (Qiagen), and libraries were prepared using ScriptSeq (EpiCentre) or a KAPA RNA HyperPrep Kit (RiboErase HMR; Roche). Transcript quantification was done using Salmon (v1.4)[58] using gencode v33 and hg38. The R package txImport (v1.20.0)[59] was used to summarize expression at the gene level, and edgeR (v3.34.1)[60] was used to normalize the data and output a logCPM data matrix. Samples with a library size below 2 million reads ($n = 43$) were only used to study transcriptomic subtypes (UROMOL2021) and discarded for the remaining analyses. Only genes expressed in at least 25% of samples were kept.

Subtype classifications and RNA-based estimation of immune cell populations

Tumors that were not included in ref. 6 were classified according to the UROMOL2021 scheme using the R package classifyNMIBC (v1.1.0). Tumors were previously classified according to the LundTax and consensusMIBC classification schemes[5,6,61]. The presence of different immune cell populations in the tumor samples was estimated from the RNA-seq data as in ref. 62 using established gene expression signatures[63,64]. A score for each immune cell population was calculated as the mean expression of the respective cell-type marker genes. Individual scores for B cells, Th1 cells and natural killer cells were not evaluated as most or all marker genes were missing for these cell populations. T cell exhaustion was estimated from RNA as previously described[8].

Pathway enrichment analysis

For 312 tumors (corresponding to the number of tumors with available data on CNAs and >2 million reads from the RNA-seq data), pathway enrichment analysis was performed for genes with higher expression in tumors with WGD and in diploid tumors, separately (fold change >0.5 and false discovery rate (FDR)-adjusted $P < 0.05$). The analysis was performed based on the Reactome Pathway database[65] using the function enrichPathway from the R package ReactomePA (v1.47.0). The Reactome Pathways are hierarchical with parent pathways and different layers of underlying pathway layers. All enriched pathways were labeled according to their parent pathway to identify if the enriched pathways predominantly belonged to specific parent pathways.

Regulon analysis

Gene regulatory networks (regulons) were reverse-engineered for all transcription factors using the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe)-AP (v1)[66]. All transcription factors were analyzed to remove indirect targets in which two genes are coregulated by estimating data processing inequality[67]. ARACNe was performed with 100 bootstrap iterations using a mutual information threshold of $P < 1 \times 10^{-8}$. A consensus network was built by keeping edges that occurred a significant number of times across the iterations.

Based on the ARACNe regulons, regulon activity was estimated by the Virtual Inference of Protein activity by Enriched Regulon analysis (VIPER, v1.22.0)[68]. VIPER first calculates single-sample gene expression signatures by comparing the expression of a target gene in a sample to the average expression level across all samples. Then it calculates the enrichment of each regulon by using the analytic ranked-based enrichment analysis (aREA) algorithm. aREA rank-sorts the previously calculated gene expression signatures in a regulon and applies quantile normalization of the ranks.

Fusion gene discovery

The RNA-seq fastq files from the four tumor samples included in the Nanopore long-read sequencing were mapped using STAR (v2.7.6a)[69], and fusion genes were called using Arriba (v2.1.0)[70] and Star-fusion (v1.10.0)[71]. Fusion-inspector (v2.5.0)[72] was run on the output of star-fusion.

microRNA (miRNA) activity analysis

miRNA activity was predicted using the miReact (v1.0.0)[73,74] run on the RNA-seq data from 312 tumors (corresponding to the number of tumors with available data on CNAs and >2 million reads from the RNA-seq data), and miRNA annotations were extracted from miRBase (v20)[75]. The predicted miRNA activity is an expression proxy and a relative measure for the level of target repression. Differential miRNA activity between diploid tumors and tumors with WGD was assessed using two-sided Wilcoxon rank sum tests. The Bonferroni method was used to adjust for multiple testing as it applies a stringent control on type I errors (family wise error rate <0.1).

The miRNA seed site was defined as position 2–8 in the miRNA sequence. Subsequently, miRNA target genes were defined as containing at least two matches to the reverse complementary target site in their 3′ untranslated region and having a negative Pearson correlation <−0.5 between their expression and the miRNA activity.

Gene Ontology enrichment analysis was conducted for functional profiling of the miRNA target genes. The analysis was performed using g:Profiler (version e110_eg57_p18_4b54a898)[76], and *P* values were adjusted using the Bonferroni method.

Estimation of LOH of HLA

Tumor-specific LOH in the HLA genes was called using the tool loss of heterozygosity in human leukocyte antigen (LOHHLA; v1)[30] based on WES data. Additionally, LOHHLA requires information about germline HLA type, tumor fraction and ploidy. Germline HLA types were inferred by POLYSOLVER (v4)[77] using buffy-coat WES data. As described in previous sections, the tumor fraction was estimated using PurBayes[51] based on WES data, and ploidy was determined using ichorCNA (v0.3.2)[21] based on sWGS and WES data. As described in the original paper, HLA genes with a copy number <0.5 and an allelic imbalance $P < 0.01$ were classified as LOH. Samples with at least one LOH event were classified as LOH of HLA.

PD-1 protein expression assessed by IHC

Sections of 3 µm thickness from tissue microarrays consisting of 1 mm triplicate tissue cores from 156 of the tumors were used for immunohistochemical analysis. IHC was performed as previously described[78] with an antibody directed against PD-1 (Clone EP239; Cell Marque). To identify tissue location (carcinoma or stroma), a sequential section was stained against pan-cytokeratin (Clone AE1/3; Dako, Agilent).

Digital pathology was used to automate the quantification of PD-1-positive cells and was carried out using Visiopharm Software (v2018.9.5.5952; Visiopharm A/S). Each cytokeratin-stained image was aligned to its sequential PD-1-stained image using the Tissue Align module to differentiate between carcinoma and stromal areas of the core biopsies.

Integrative clustering

iClusterBayes[79] was applied to data on somatic mutations (derived from WES), copy-number estimations (derived from sWGS) and gene expression levels (derived from RNA-seq) for 230 tumors. The somatic mutation data were summarized as a binary matrix (1: mutation and 0: no mutation) at the gene level. Only mutations characterized as high or moderate by snpEff and only genes mutated in more than 5% of the samples were kept, resulting in 212 mutated genes. CNAs were summarized by collecting all breakpoints from the original segmentation, thus defining all unique genomic segments (with different sizes) present in the dataset. For each segment and each tumor, we attributed the raw segment value (from 0 copies to 6.0 copies). Only segments with no missing values belonging to autosomal chromosomes were kept, resulting in 5,445 segments. Finally, we log-transformed the matrix (negative value for losses and positive values for gains). Only genes with a mean expression above zero and a s.d. above 1.5 were kept, resulting in a gene expression matrix for 1,529 genes. iClusterBayes was run with two to ten clusters and with default settings, and the solution with four clusters was chosen.

WES-based estimation of T cell fraction and TCR diversity

Based on tumor and germline WES data, we estimated tumor and blood T cell fractions (TCR), respectively, using TcellExTRECT (v1.0.1) with default settings[32]. For diversity estimation, productive TCR β sequences were extracted using the MiXCR (v3.0.13) analyze shotgun functionality[80]. Patients with more than two reads and at least two distinct clones were included in the analysis. When assessing associations to outcome, disease progression status was defined as progression to MIBC within 5 years after transurethral resection of the bladder tumor.

Statistical analysis

Statistical analyses were performed in R (v4.1). Statistically significant differences between groups were assessed by the Kruskal–Wallis or two-sided Wilcoxon rank sum test for numeric variables and the Fisher's exact test or the chi-square test for categorical variables. Survival analyses were visualized by the Kaplan–Meier method, and differences between groups were assessed by two-sided log-rank tests using the R packages survival and survminer. Cox proportional hazards analyses were performed to adjust for other variables in the survival analyses using the R packages survival and survminer. For multiple testing, *P* values were adjusted using the Benjamini–Hochberg/FDR approach unless otherwise stated.

Catalog numbers for all reagents are provided in Supplementary Table 13.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequencing data are deposited at the European Genome–Phenome Archive (EGA), which is hosted by the European Bioinformatics Institute and the Centre for Genomic Regulation. The data are available under controlled access at EGA due to privacy laws and legal restrictions associated with sharing sensitive data under the General Data Protection Regulation. Access to data that are under controlled access requires that the data requestor (the legal entity) enter into collaboration and data processing agreements with the Central Denmark Region (the legal entity controlling and being responsible for the data). Request to access data furthermore requires that the purpose of data re-analysis is approved by the Danish National Committee on Health Research Ethics. Upon request, the authors, on behalf of the Central Denmark Region, will enter into a collaboration with the data requestor to apply for approval. Any requests will be reviewed within a time frame of 2–3 weeks by the data assessment committee. This applies to the following datasets:

sWGS data are available under accession number EGAS50000000513. WES data are available under accession number EGAS50000000511. RNA-seq data are available under accession number EGAS50000000512. Nanopore long-read sequencing data are available under accession number EGAS50000000510.

Clinical information, molecular tumor characteristics, annotated somatic variant calls, copy-number estimations and gene expression levels are provided in the source data.

The Reactome Pathway database[65] was accessed via the function enrichPathway from the R package ReactomePA (v1.47.0)[81]. Source data are provided with this paper.

Code availability

No custom code was applied as we used publicly available data processing and visualization tools as described in Methods.

References

43. Lamy, P. et al. Paired exome analysis reveals clonal evolution and potential therapeutic targets in urothelial carcinoma. *Cancer Res.* **76**, 5894–5906 (2016).
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Benjamin, D. et al. Calling somatic SNVs and indels with Mutect2. Preprint at *bioRxiv* https://doi.org/10.1101/861054 (2019).
47. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
48. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
49. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
50. Islam, S. M. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.* **2**, 100179 (2022).
51. Larson, N. B. & Fridley, B. L. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29**, 1888–1889 (2013).
52. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. Elrick, H. et al. SAVANA: reliable analysis of somatic structural variants and copy number aberrations in clinical samples using long-read sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2024.07.25.604944 (2024).
56. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
57. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
58. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
59. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
61. Marzouka, N.-A.-D. et al. A validation and extended description of the Lund taxonomy for urothelial carcinoma using the TCGA cohort. *Sci. Rep.* **8**, 3737 (2018).
62. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
63. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
64. Danaher, P. et al. Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* **5**, 18 (2017).
65. Gillespie, M. et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).
66. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).
67. Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).
68. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
69. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
70. Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
71. Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
72. Haas, B. J. et al. Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector. *Cell Rep. Methods* **3**, 100467 (2023).
73. Nielsen, M. M. & Pedersen, J. S. miRNA activity inferred from single cell mRNA expression. *Sci. Rep.* **11**, 9170 (2021).
74. Nielsen, M. M., Tataru, P., Madsen, T., Hobolth, A. & Pedersen, J. S. Regmex: a statistical tool for exploring motifs in ranked sequence lists from genomics experiments. *Algorithms Mol. Biol.* **13**, 17 (2018).
75. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
76. Kolberg, L. et al. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).
77. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
78. Taber, A. et al. Molecular correlates of cisplatin-based chemotherapy response in muscle invasive bladder cancer by integrated multi-omics analysis. *Nat. Commun.* **11**, 4858 (2020).
79. Mo, Q. et al. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71–86 (2018).
80. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
81. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).

Acknowledgements

The work of the laboratory of L.D. is supported in part by the following funding sources: Aarhus University, Aarhus University Hospital, the Danish Cancer Society and the Danish Cancer Biobank. Work in the laboratory of F.X.R. is supported, in part, by the Fundación Científica de la Asociación Española Contra el Cáncer (grant PRYGN223005REAL). J.S.P. and A.M.R. were supported by the Novo Nordisk Foundation (NNF18OC0053222). F.P. was supported by grants from the Neye Foundation, Højmoselegatet, Direktør Emil C. Hertz og hustru Inger Hertz' Foundation, Dagmar Marshalls Foundation, A.P. Møller Foundation, Fabrikant Einar Willumsens Memorial Trust and Aase og Ejnar Danielsens Foundation.

Author contributions

L.D. supervised the work. L.D., F.P., P.L., S.V.L., T. Strandgaard and I.N. conceived and designed the experiments. I.N. performed the

experiments. F.P., P.L., S.V.L., T. Strandgaard and A.K. performed statistical analysis. F.P., P.L., S.V.L., T. Strandgaard, N.J.B., A.K., T.G.A., J.A., J.S.P., A.M.R., F.X.R. and N.K. analyzed the data. K.B., G.G.H., K.M., A.C.P., A.H., M.G., M.H., R.N., U.S., D.S., K.E.M.K., E.C.Z., T.M., T. Simic, P.M., N.M., J.B.J. and F.X.R. contributed to reagents/materials/analysis tools. L.D., F.P., P.L. S.V.L. and T. Strandgaard wrote the paper.

Competing interests
The authors declare no competing interests.

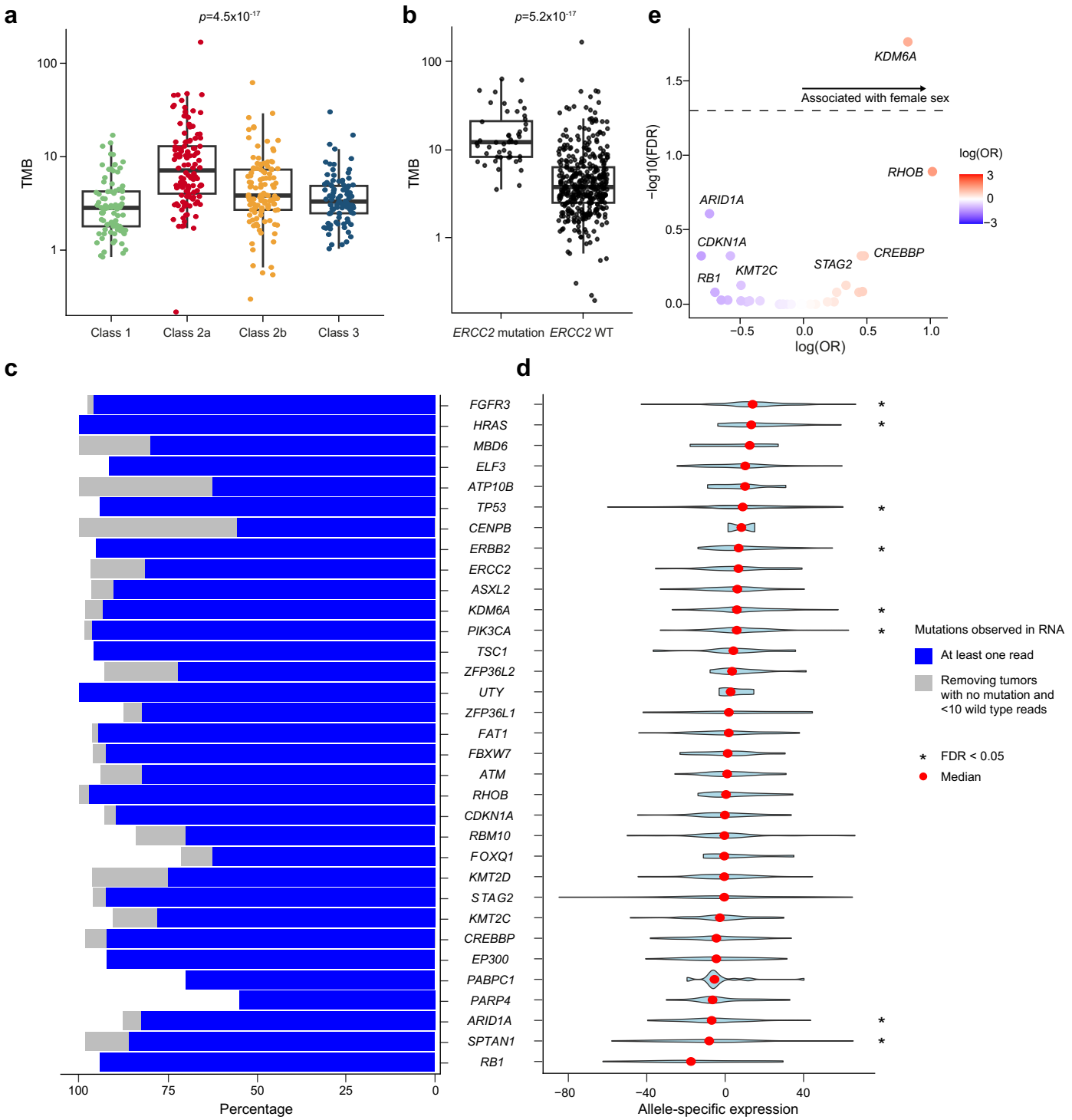
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-02030-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-02030-z.

Correspondence and requests for materials should be addressed to Lars Dyrskjøt.

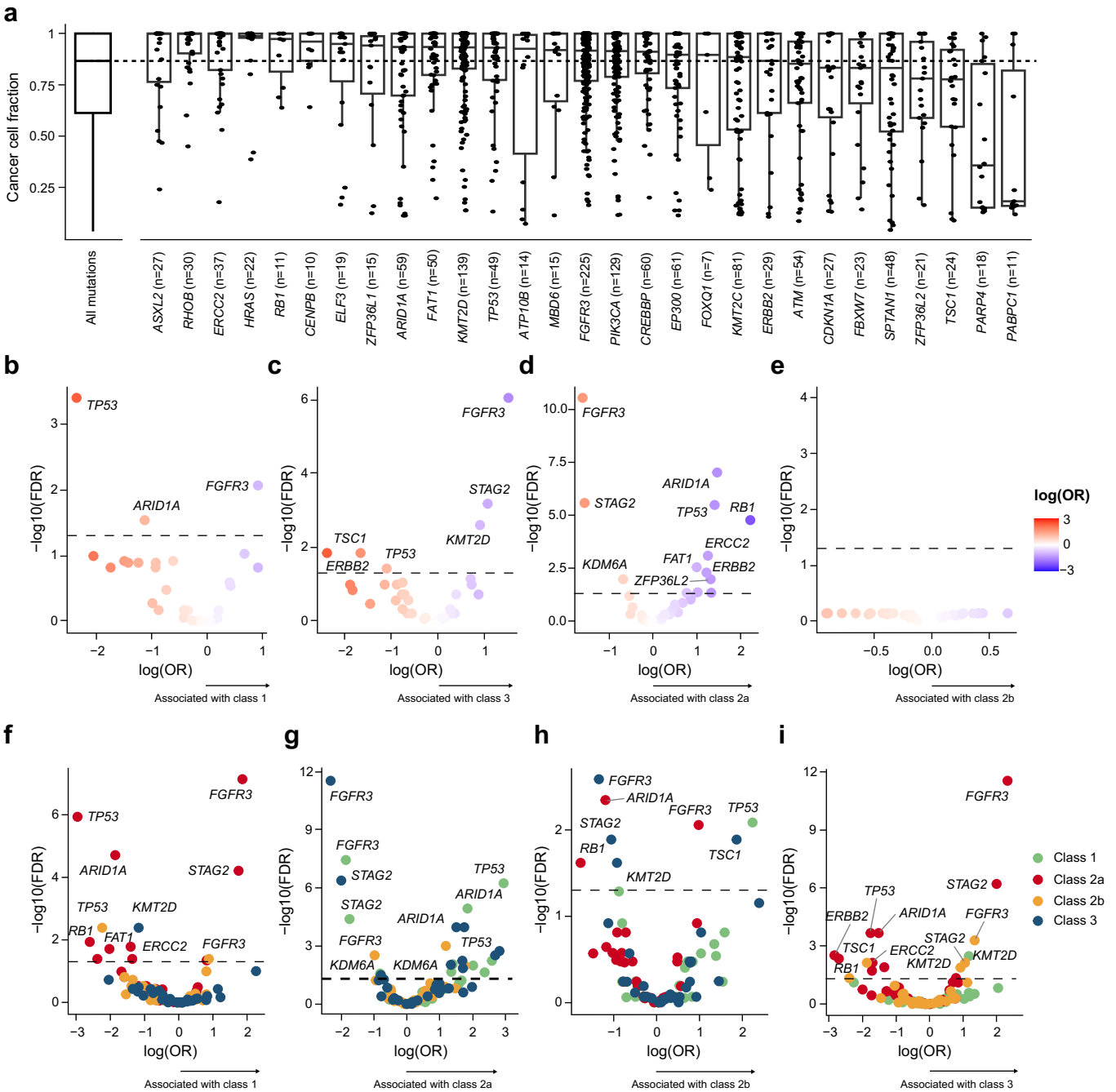
Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-02030-z



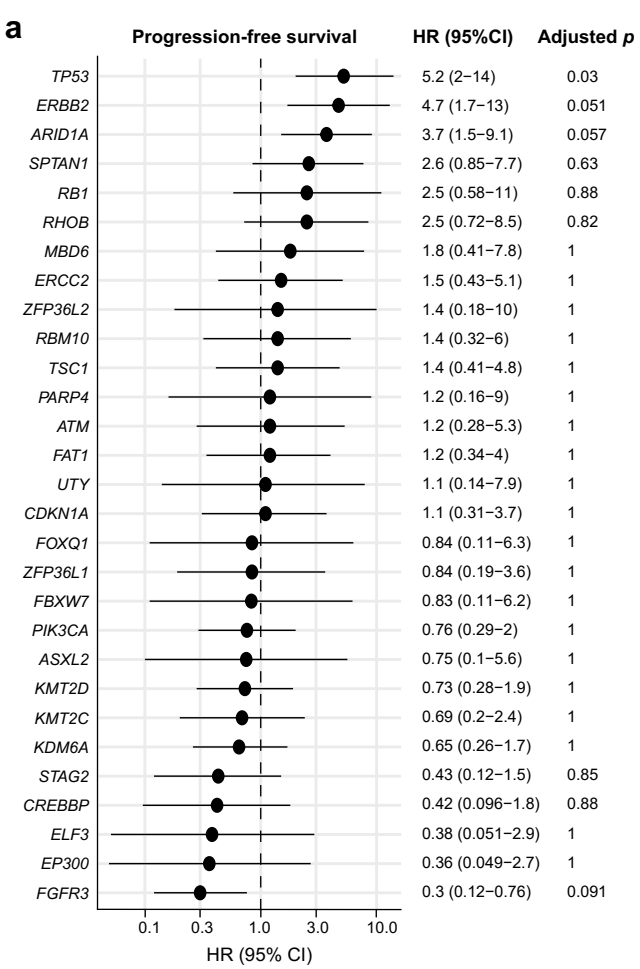
Extended Data Fig. 1 | TMB, validation of mutations in RNA-sequencing data and mutations enriched in females. a, Tumor mutational burden (TMB) stratified by the UROMOL2021 transcriptomic classes (class 1, n = 85; class 2a, n = 122; class 2b, n = 113; class 3, n = 94). Statistically significant differences between groups were determined using a Kruskal–Wallis test. **b**, TMB in *ERCC2* mutated (n = 48) and *ERCC2* wild-type (WT; n = 390) tumors. Statistically significant differences between groups were determined using a two-sided Wilcoxon rank sum test. **c**, Validation of single-nucleotide variants called in the significantly mutated genes from Fig. 1a using RNA-sequencing data. Blue bars show the percentage of mutations observed with at least one read at the RNA level in the subset of patients with both whole-exome and RNA-sequencing data available. Gray bars show the percentage of mutations observed in RNA when removing tumors without an RNA-based mutation with too low coverage (<10 WT reads from RNA-sequencing) to reliably confirm absence of a mutation. **d**, Allele-specific expression (alternate allele frequency in RNA minus alternate frequency in DNA). Asterisks indicate genes that deviate significantly from zero after correction for multiple testing (false discovery rate (FDR); Supplementary Table 3). **e**, Enrichment of mutations in females. The dashed lines represent a FDR-adjusted *p* value of 0.05. OR = odds ratio. Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

https://doi.org/10.1038/s41588-024-02030-z



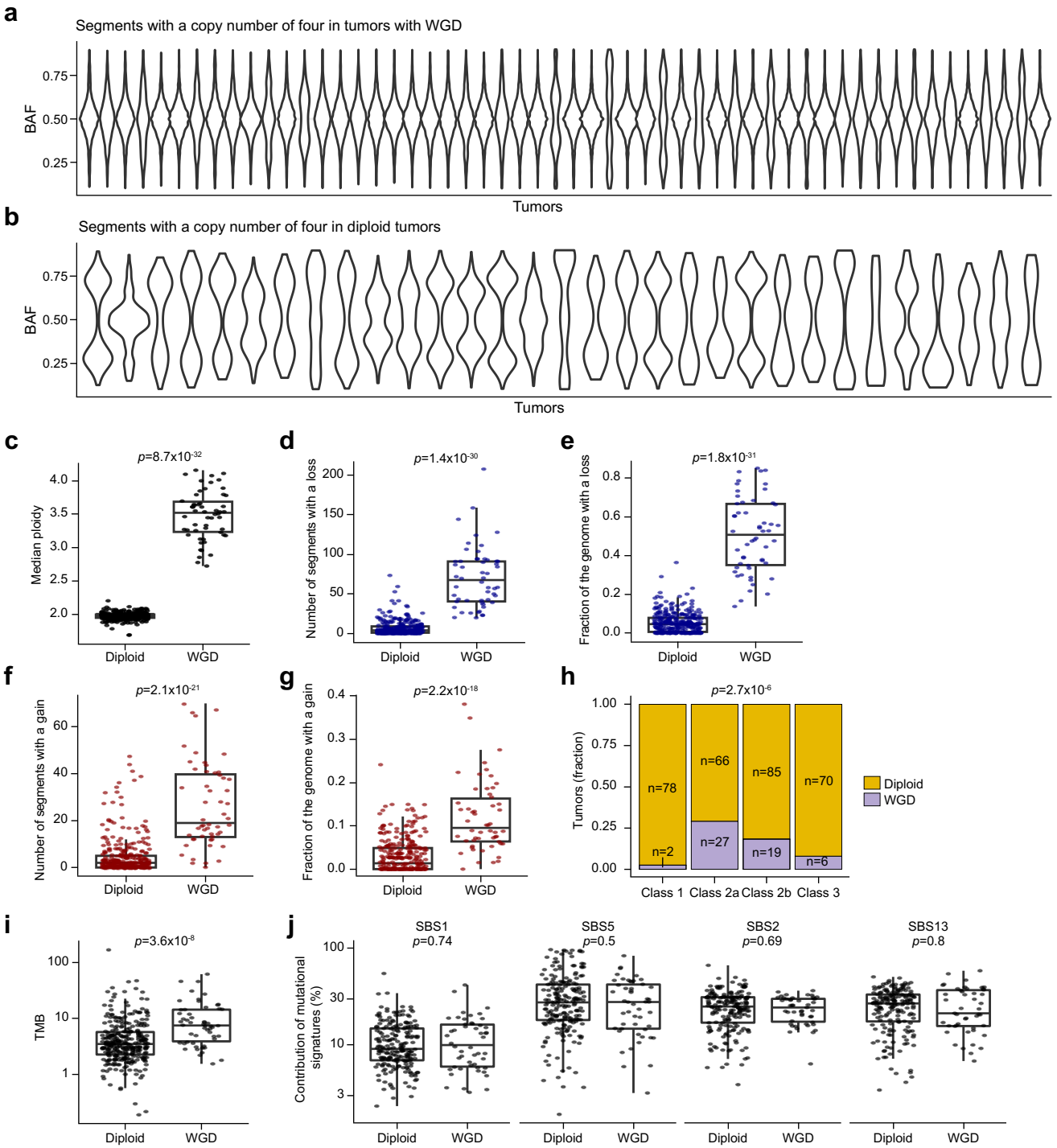
Extended Data Fig. 2 | Estimated cancer cell fractions and mutations associated with the UROMOL2021 classes. a, Estimated fraction of cancer cells harboring the mutation outlined for each of the 33 genes displayed in Fig. 1a (excluding genes located on sex chromosomes). Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range. On the x-axis, n indicates the number of tumors with a mutation in the given gene. The dashed horizontal line indicates the median cancer cell fraction when assessing all mutations. **b**–**e**, Enrichments of mutations in the UROMOL2021 classes. The dashed lines represent a false discovery rate (FDR)-adjusted *p* value of 0.05. OR = odds ratio. **f**–**i**. Pairwise comparisons of mutation frequencies in different genes between the UROMOL2021 classes. The dashed lines represent a FDR-adjusted *p* value of 0.05.

https://doi.org/10.1038/s41588-024-02030-z



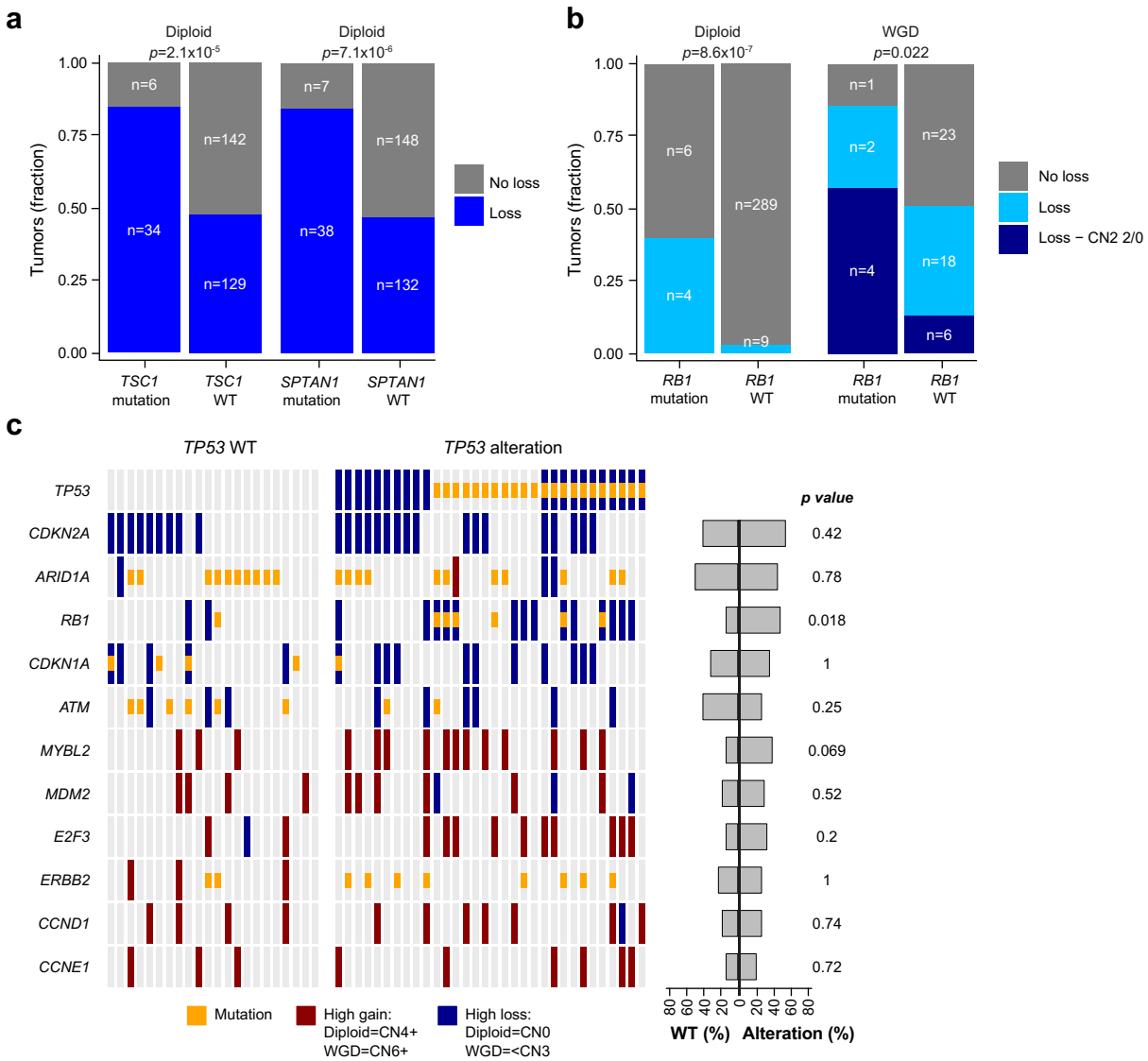
Extended Data Fig. 3 | Prognostic value of mutations and contribution of mutational signatures. a, Forest plot based on univariate Cox proportional hazards regression models of the association between gene mutations and progression-free survival for patients that did not receive BCG treatment (n = 245, 19 events). Black dots indicate hazard ratios (HRs) and horizontal lines show the corresponding 95% confidence intervals (95% CI). *P*-values were adjusted using the false discovery rate (FDR) approach. **b**, Forest plot based on univariate Cox proportional hazards regression models of the association between gene mutations and recurrence-free survival for patients that did not receive BCG treatment (n = 232, 135 events). Black dots indicate hazard ratios (HRs) and horizontal lines show the corresponding 95% confidence intervals (95% CI). *P*-values were adjusted using the FDR approach. **c**, Contribution of different single-base substitutions (SBS) 96 signatures to the mutational landscape of the 325 tumors having more than 100 single-nucleotide variants. Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

https://doi.org/10.1038/s41588-024-02030-z



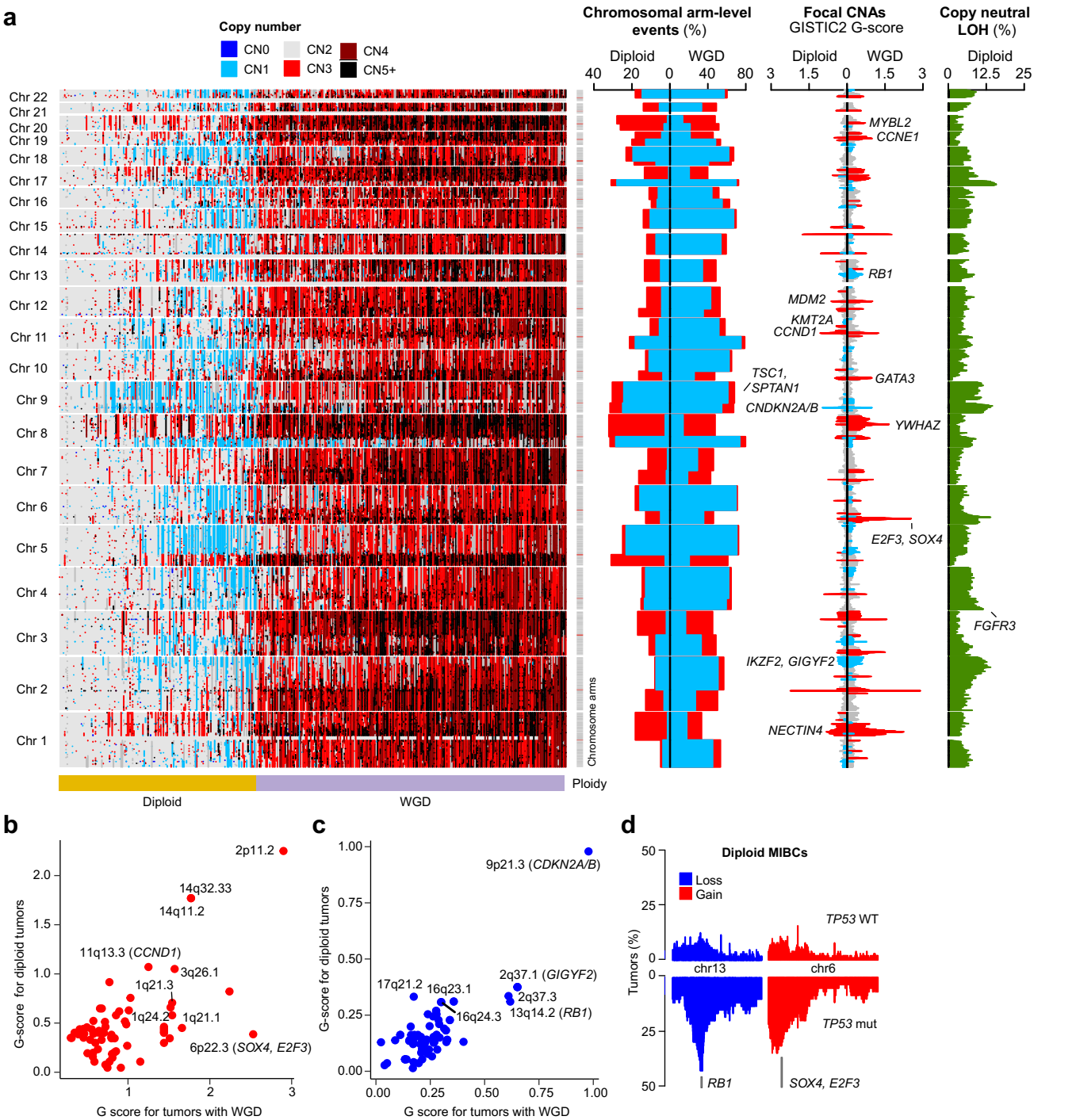
Extended Data Fig. 4 | Molecular characteristics of diploid tumors and tumors with WGD. a, B-allele frequency (BAF) in segments with a copy number of four in tumors classified as having undergone whole-genome doubling (WGD). **b**, BAF in segments with a copy number of four in tumors classified as being diploid. **c**, Median ploidy in diploid tumors (n = 308) and tumors with WGD (n = 54). **d**–**g**, Number of segments with copy-number loss (**d**), the fraction of the genome with a loss (**e**), the number of segments with copy-number gain (**f**) and the fraction of the genome with a gain (**g**) in diploid tumors (n = 308) and tumors with WGD (n = 54). Loss was defined as a copy number <2 in diploid tumors and <4 in tumors with WGD. Gain was defined as a copy number >2 in diploid tumors and >4 in tumors with WGD. **h**, Association between the UROMOL2021 classes and tumor ploidy status (class 2a/2b versus class 1/3). **i**, Tumor mutational burden (TMB) in diploid tumors (n = 308) and tumors with WGD (n = 54). **j**, Contribution of single-base substitutions (SBS) 96 signatures in diploid tumors (n = 206) and tumors with WGD (n = 50). Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range. Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests (**c**–**g**,**i**,**j**) or Fisher's exact test (**h**).

https://doi.org/10.1038/s41588-024-02030-z



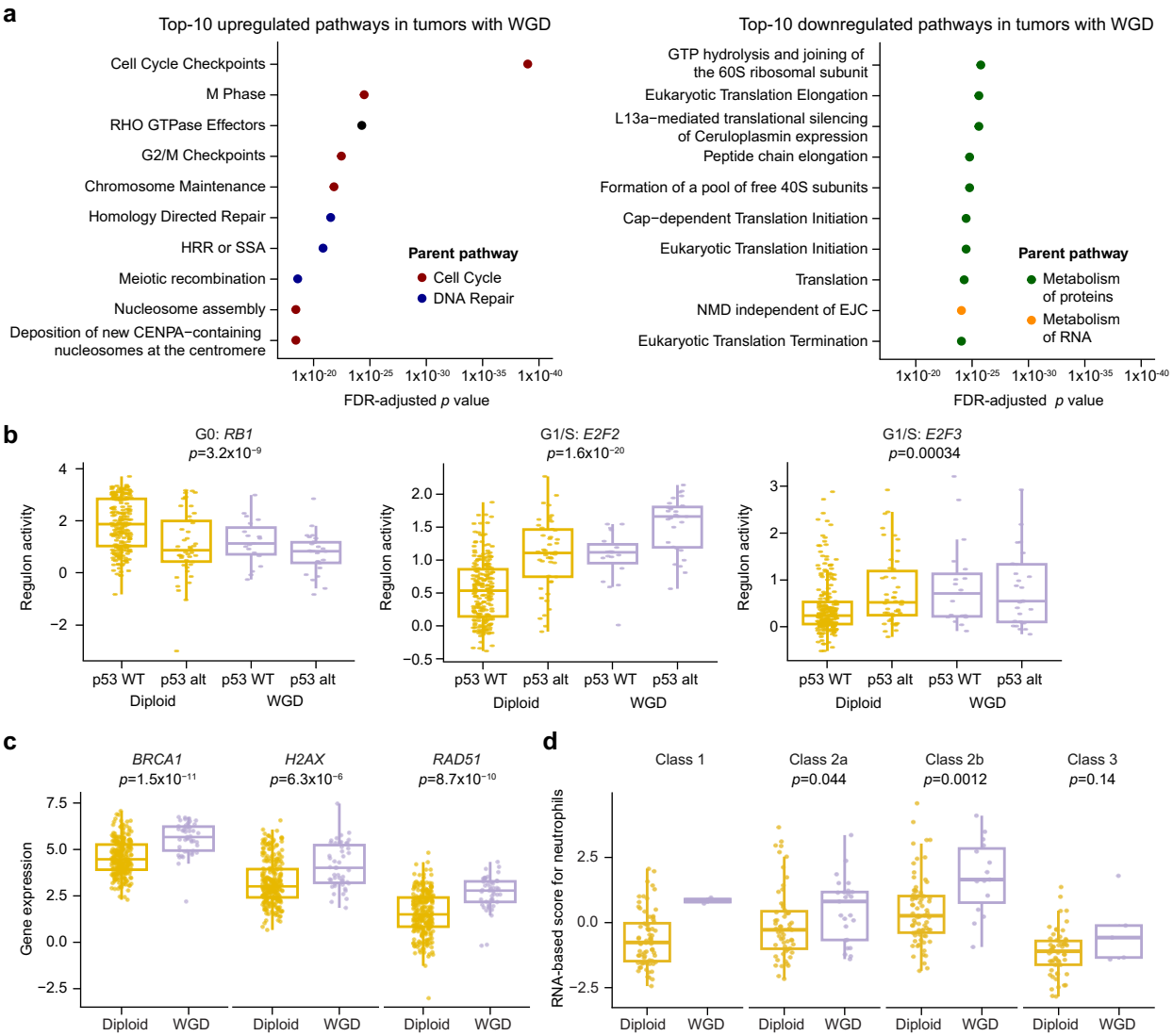
Extended Data Fig. 5 | Copy-number alterations and mutations in diploid tumors and tumors with WGD. a, Co-occurrence of mutations and gene loss in *TSC1* and *SPTAN1*, respectively, in diploid tumors. Statistically significant associations between groups were determined using chi-square test. WT = wild type. **b**, Co-occurrence of *RB1* mutations and *RB1* loss stratified by ploidy status. WGD = whole-genome doubling. Loss was defined as copy number (CN) <2 for diploid tumors and <4 for tumors with WGD. CN2 2/0 represents loss of one allele and duplication of the remaining allele. Statistically significant associations between groups were determined using a chi-square test. **c**, Genomic alterations in genes involved in cell cycle regulation and genes more frequently mutated in tumors with WGD compared with diploid tumors (Fig. 2d). Tumors are stratified by *TP53* alteration status. Statistically significant associations between groups were determined using Fisher's exact tests.

https://doi.org/10.1038/s41588-024-02030-z



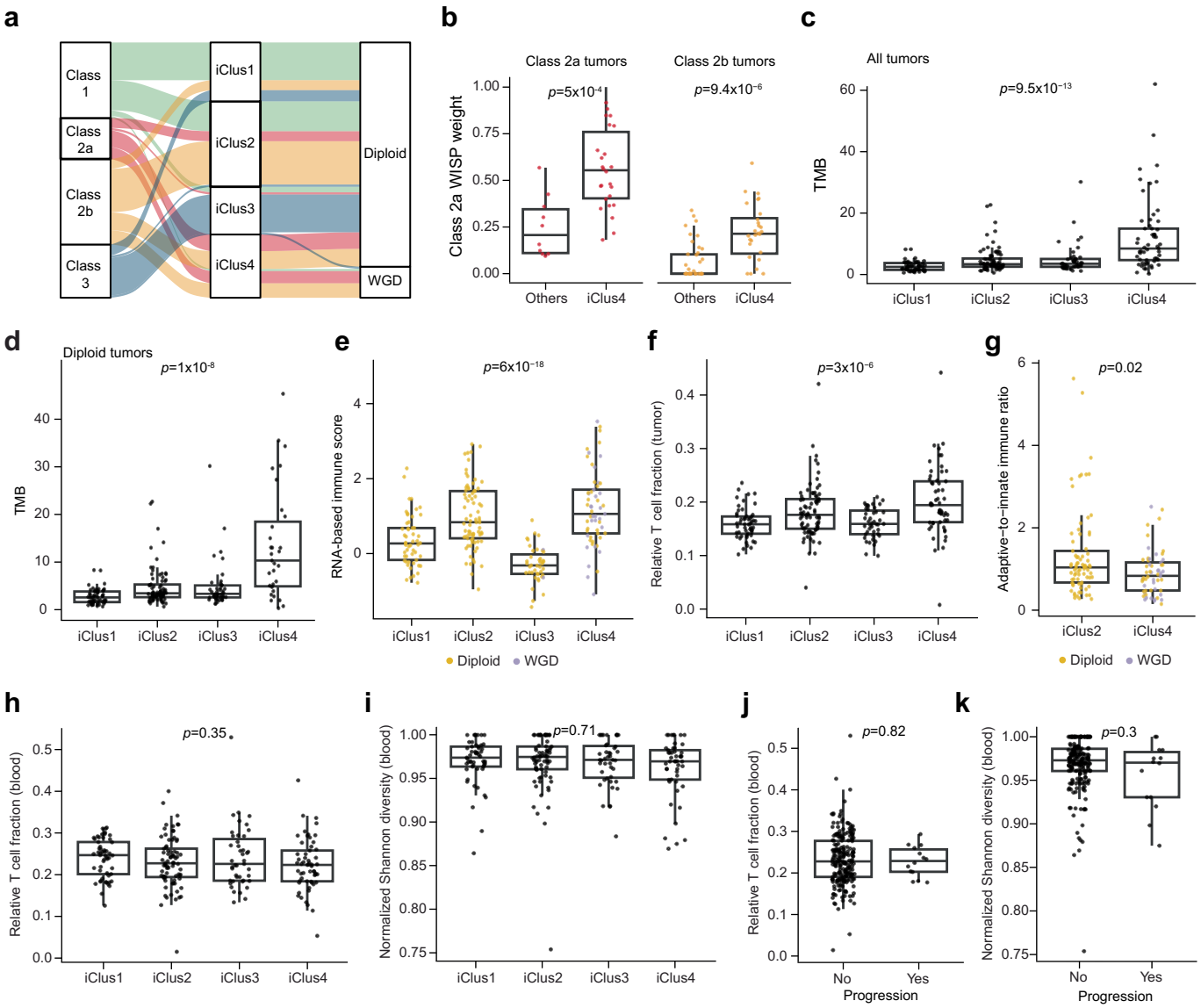
Extended Data Fig. 6 | Copy-number alterations in MIBC. a, Copy-number (CN) profile of 412 tumors from TCGA (muscle-invasive bladder cancer (MIBC)) arranged by the proportion of the genome with a CN > 2. Vertical annotations on the left: definition of chromosome arms with red lines indicating centromeres (left); chromosomal arm-level events (>70% of the chromosomal arm affected) for diploid tumors (compared to two copies) and tumors with WGD (compared to four copies; left-middle); genome-wide G scores computed by GISTIC2, red and blue bars highlight chromosomal regions that are significantly enriched for focal gains or losses, respectively (right-middle); percentage of diploid tumors with copy-neutral loss of heterozygosity (LOH; right). CNAs = copy-number alterations. **b**,**c**, GISTIC2-computed G scores of tumors from TCGA reflecting how significantly a region is affected by focal gains (**b**) and losses (**c**) in tumors with WGD and diploid tumors. **d**, Frequency of gains in chromosome (chr) 13 and gains in chr6 in diploid tumors from TCGA stratified by *TP53* mutation status.

https://doi.org/10.1038/s41588-024-02030-z



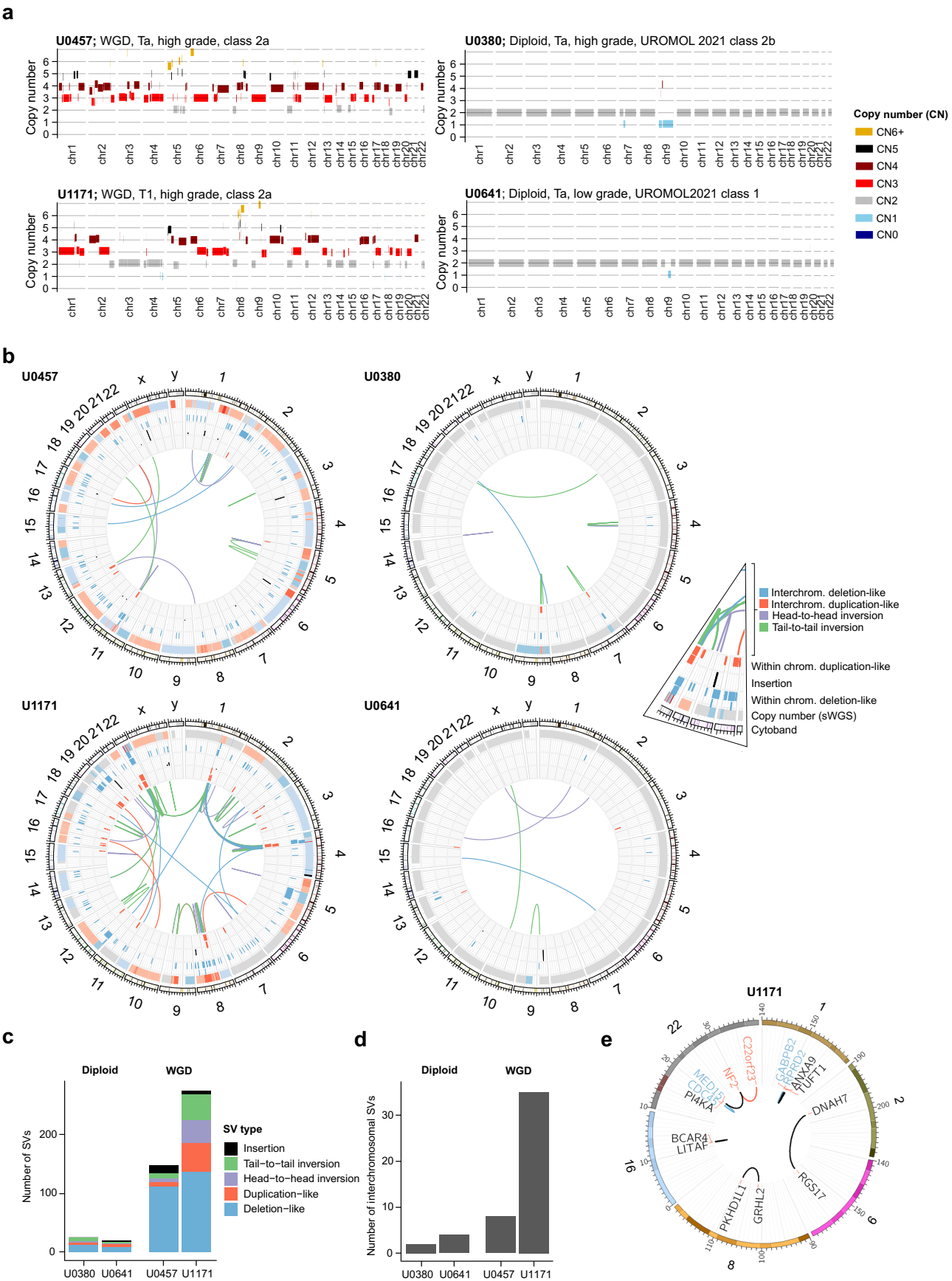
Extended Data Fig. 7 | Gene expression features associated with WGD. **a**, False discovery rate (FDR)-adjusted *p* values of the 10 most upregulated (left) and downregulated (right) reactome pathways in tumors with whole-genome doubling (WGD) as calculated by the enrichPathway function of the ReactomePA R package. HRR = homology-directed repair through homologous recombination, SSA = single-strand annealing, CENPA = centromere protein A, NMD = nonsense-mediated decay, EJC = exon junction complex. **b**, Regulon activity of genes involved in different steps of the cell cycle, stratified by ploidy status and involvement of the p53 pathway (*TP53* mutations and/or *MDM2* gain; 219 diploid, wild-type (WT) tumors; 46 diploid, p53 pathway-altered tumors; 19 WGD, WT tumors; 28 WGD, p53 pathway-altered tumors). Statistically significant differences between groups were determined using Kruskal–Wallis tests. p53 alt = altered p53 pathway. **c**, Expression of genes involved in homology-directed repair stratified in tumors with WGD (n = 47) and diploid tumors (n = 265). Statistically significant associations between groups were determined using two-sided Wilcoxon rank sum tests. **d**, RNA-based score for neutrophil infiltration stratified by ploidy status and UROMOL2021 classes (class 1, 72 diploid tumors and 2 with WGD (no statistical test performed); class 2a; 59 diploid tumors and 25 with WGD; class 2b, 75 diploid tumors and 15 with WGD; class 3, 59 diploid tumors and 5 with WGD). Statistically significant associations between groups were determined using two-sided Wilcoxon rank sum tests. Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

https://doi.org/10.1038/s41588-024-02030-z



Extended Data Fig. 8 | Molecular features associated with iClusters. a, Sankey plot showing the association between the UROMOL2021 classes, the iClusters (iClus) and whole-genome doubling (WGD) status of the 230 tumors included in the integrative clustering analysis. **b**, Class 2a weighted in silico pathology (WISP) weight stratified by iClusters for tumors classified as the UROMOL2021 class 2a (iClus4, n = 26; others, n = 11) and class 2b (iClus4, n = 29; others, n = 48). **c**, Tumor mutational burden (TMB) stratified by iClusters (iClus1, n = 53; iClus2, n = 77; iClus3, n = 43; iClus4, n = 57). **d**, TMB stratified by iClusters for diploid tumors (iClus1, n = 53; iClus2, n = 77; iClus3, n = 41; iClus4, n = 31). **e**, RNA-based immune infiltration score stratified by iClusters (iClus1, n = 53; iClus2, n = 77; iClus3, n = 43; iClus4, n = 57). **f**, Estimated tumor T cell fraction stratified by iClusters (iClus1, n = 53; iClus2, n = 77; iClus3, n = 43; iClus4, n = 57). The T cell fraction was estimated using whole-exome sequencing (WES) of tumor DNA. **g**, Adaptive-to-innate immune ratio for tumors in iClus2 (n = 77) and iClus4 (n = 57). Values above 1 indicate a higher adaptive component, whereas values below 1 indicate a higher innate component. **h**, Estimated blood T cell fraction stratified by iClusters (iClus1, n = 53; iClus2, n = 77; iClus3, n = 43; iClus4, n = 57). The T cell fraction was estimated using WES of germline DNA. **i**, T cell receptor diversity in blood (Shannon diversity, estimated using WES of germline DNA) stratified by iClusters (iClus1, n = 53; iClus2, n = 77; iClus3, n = 43; iClus4, n = 57). **j**, Estimated blood T cell fraction stratified by disease progression status (yes, n = 14; no, n = 215). The T cell fraction was estimated using WES of germline DNA. **k**, T cell receptor diversity in blood (Shannon diversity, estimated using WES of germline DNA) stratified by disease progression status (yes, n = 14; no, n = 215). Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range. Statistically significant associations between groups were determined using two-sided Wilcoxon rank sum tests or Kruskal–Wallis tests.

https://doi.org/10.1038/s41588-024-02030-z

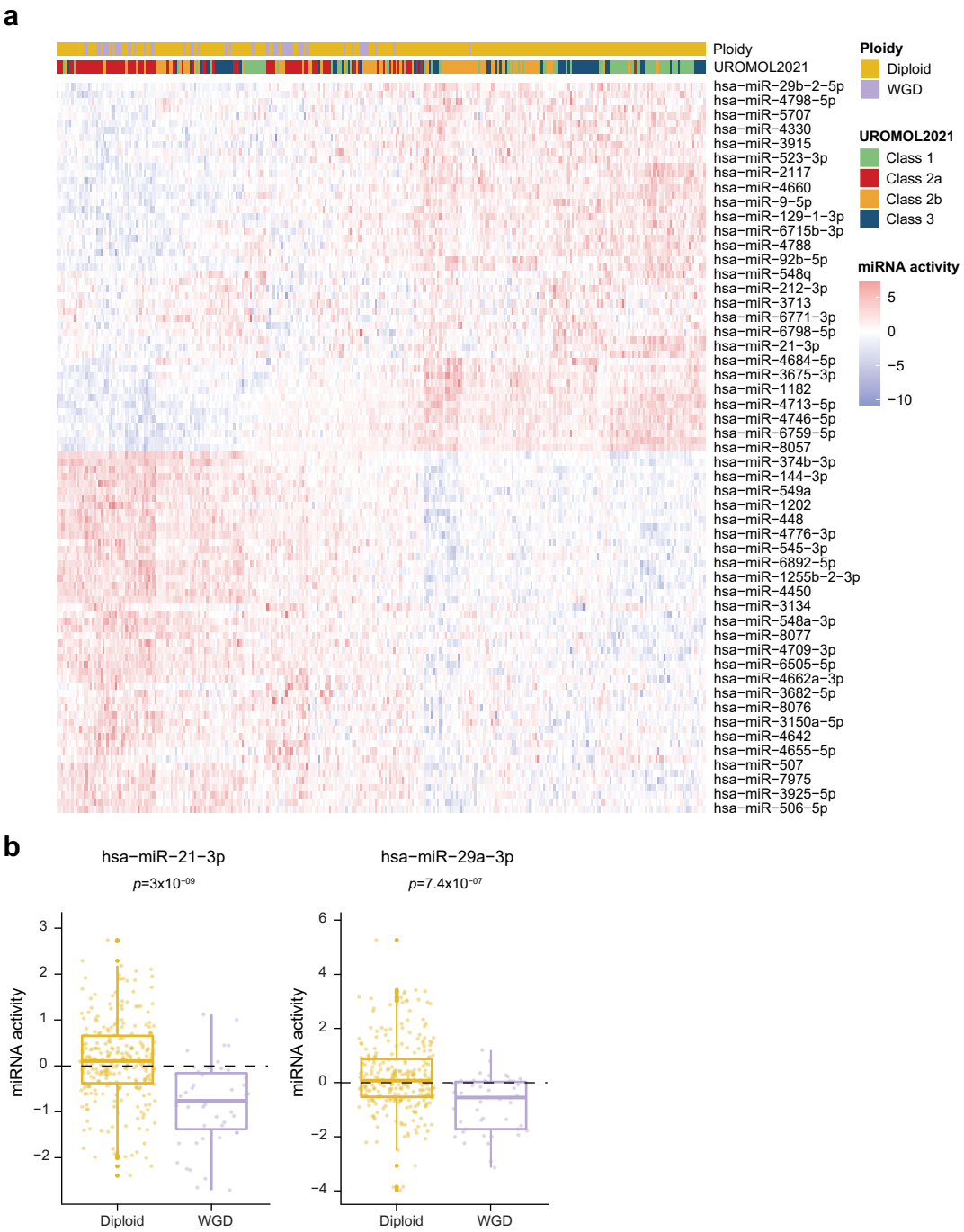


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41588-024-02030-z

Extended Data Fig. 9 | Somatic structural variants in NMIBC assessed by long-read sequencing. a, Inferred total copy-number (CN) profile estimated from shallow whole-genome sequencing (sWGS) for two tumors with whole-genome doubling (WGD; left) and two diploid tumors (right). UROMOL IDs are shown. Chr = chromosome. **b**, Circos plots showing structural variants (SVs) estimated from long-read sequencing for two tumors with WGD (U0457 and U1171) and two diploid tumors (U0380 and U0641). **c**, Number of SV types within each of the four tumors analyzed by long-read sequencing. **d**, Number of interchromosomal SVs in the four tumors. **e**, Circos plot presenting the fusion genes called from RNA-sequencing data with DNA evidence from long-read sequencing of tumor U1171.

https://doi.org/10.1038/s41588-024-02030-z



Extended Data Fig. 10 | microRNA activity in diploid tumors and tumors with WGD. a, Heatmap of cross-patient microRNA (miRNA) activities for the 101 miRNAs with differential activities between diploid tumor and tumors with whole-genome doubling (WGD; 312 tumors in total). Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests, and *p* values were adjusted using the Bonferroni method (family-wise error rate (FWER)). miRNAs with an FWER < 0.1 are shown. Hierarchical agglomerative clustering was applied to both rows and columns for visualization purposes. **b**, miRNA activity of miR-21 and miR-29a for diploid tumors (n = 265) and tumors with WGD (n = 47). Statistically significant differences between groups were determined using two-sided Wilcoxon rank sum tests. Boxplots represent the median, lower and upper quartiles, and whiskers correspond to the 1.5× interquartile range.

nature portfolio

Corresponding author(s): Lars Dyrskjøt

Last updated by author(s): Oct 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	Data analysis and figures: R v4.1 Preprocessing of WES data: cutadapt v3.7 bwa v0.7.17 Picard tool MarkDuplicates v2.27.00 Preprocessing of long read sequencing data: minimap2 v2.24 samtools v1.15.1 Downstream data analysis of genomic data: Mutect2 v2.2 Strelka v2.9.10 SnpEff v4.3t MutSigCV v1 GenomicRanges v1.44.0 rtracklayer v1.52.00 sigprofilerassignment v0.0.11 sigprofilerextractor v1.1.10

sigprofilermatrixgenerator v1.2.9
sigprofilerplotting v1.2.2
PurBayes v1.3
ichorCNA v.0.3.2
Battenberg v2.2.10
GISTIC2.0
SAVANA v0.2.3
CIRCOS v1.3.5
LOHHLA v1
POLYSOLVER v4
survminer v0.4.9
survival v3.7.0

Preprocessing and analysis of RNA sequencing data:
Salmon v1.4
tximport v1.20.0
edgeR v3.34.1
classifyNMIBC v1.1.0
ReactomePA v1.47.0
ARACNe-AP v1
VIPER v1.22.0
STAR v2.7.6a
Arriba v2.1.0
Star-fusion v1.10.0
Fusion-inspector v2.5.0
miReact v1.0.0
miRBase v20
g:Profiler version e110_eg57_p18_4b54a898
TcellExTRECT v1.0.1
MiXCR v3.0.13

Integrative clustering was performed using iClusterBayes.

Quantification of protein markers was carried out using Visiopharm version 2018.9.5.5952 software (Visiopharm A/S, Hørsholm, Denmark).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw sequencing data are deposited at The European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG). The data are available under controlled access at EGA due to privacy laws and legal restrictions associated with sharing sensitive data under the General Data Protection Regulation (GDPR). Access to data that is under controlled access requires that the data requestor (the legal entity) enter into Collaboration and Data Processing Agreements with the Central Denmark Region (the legal entity controlling and being responsible for the data). Request to access data furthermore requires that the purpose of data re-analysis is approved by The Danish National Committee on Health Research Ethics. Upon request, the authors, on behalf of the Central Denmark Region, will enter into a collaboration with the data requestor to apply for approval. Any requests will be reviewed within a time frame of 2–3 weeks by the data assessment committee. This applies to the following datasets:
sWGS data are available under accession number: EGAS50000000513; WES data are available under accession number: EGAS50000000511; RNA-seq data are available under accession number: EGAS50000000512; Nanopore long-read sequencing data are available under accession number: EGAS50000000510.
Clinical information, molecular tumor characteristics, annotated somatic variant calls, copy number estimations and gene expression levels are provided in the Source data.
The Reactome Pathway database65 was accessed via the function enrichPathway from the R package ReactomePA v1.47.081.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Patients included in the study were not selected based on sex, but resembled the general sex discrepancy of bladder cancer cases (around 75% of diagnosed cases are males) 346 males and 92 females were included. Information on gender was not collected.
Reporting on race, ethnicity, or other socially relevant groupings	Patients were not selected based on race, ethnicity or other socially relevant groupings and information on this was not collected.

Population characteristics	Characteristics and demographics of the patients included in this study are summarized in Supplementary Table 1 and a per-patient table of clinical information can be found in the Source Data.
Recruitment	Patient recruitment was carried out in each involved clinical center following national ethics guidelines (see below). As frozen materials were needed for analysis, a bias towards patients with larger tumors where adequate materials for both research and pathological evaluation is expected. Patients were not requiered to take active part in the study, but should donate their tissue to research. The inactive participance limits patient bias.
Ethics oversight	This study complies with all relevant ethical regulations. Informed written consent to take part in research projects was obtained from all patients, and all ethical regulations for work with human participants were followed according to national guidelines. The study was approved by the Central Denmark Region Committees on Biomedical Research Ethics (#1994/2920; Skejby, Aalborg, Frederiksberg); the Danish National Committee on Health Research Ethics (#1906019; #1708266), the ethics committee of the University Hospital Erlangen (#3755); the ethics committee of the Technical University of Munich (#2792/10); Medical Ethics Committee of Erasmus MC (MEC#168.922/1998/55; Rotterdam); the Uppsala Region Committee on Biomedical Research Ethics (#2008/252); the Ethical Committee of Faculty of Medicine, University of Belgrade (#440/VI-7); the Ethics Committee (CEIC) of Institut Municipal d'Assistència Sanitària/Hospital del Mar (2008/3296/I); the ethics committee of the University Hospital Jena (#4774-4/16). Patients did not receive any compensation for participation.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size and samples were included based on availability.
Data exclusions	Patients were excluded from WES or sWGS if the available tumor sample were not suitable for the specific platform (low carcinoma cell fraction, low DNA concentration, fresh frozen material not available, no clinical information). This resulted in the partially overlapping multi-omics analyses.
Replication	As this is an exploratory study and patient material is limited, no replication was performed.
Randomization	As this is an exploratory study, randomization is not relevant. It is not a clinical trial and no treatment is given based on this study.
Blinding	This is an exploratory study and no treatment is given to the patients based on this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	PD-1; Clone:EP239; dilution:1:100; incubation:32 min; Cell Marque (Cat # 315R). Pan Cytokeratin; Clone:AE1/3; dilution:1:100; incubation:16min; Dako, Agilent (Cat # GA05361-2).

Validation

Staining was performed at The Department of Pathology, Aarhus University Hospital on the Ventana Benchmark Ultra instrument by a trained technician.
We have provided a link for the relevant data sheet for each antibody. The data sheet includes the manufacturer's validation statements, quality control procedures and relevant citations:
PD-1: https://www.cellmarque.com/antibodies/CM/3562/PD-1_EP239
Pan Cytokeratin: https://www.agilent.com/cs/library/packageinsert/public/107609005.PDF

Plants

Seed stocks

NA

Novel plant genotypes

NA

Authentication

NA