Gene mutant dosage determines prognosis and metastatic tropism in 60,000 clinical cancer samples

Nicola Calonaci[1], Eriseld Krasniqi[2], Daniel Colic[3], Stefano Scalera[2], Giorgia Gandolfi[1], Salvatore Milite[4], Konstantin Bräutigam[5], Andrea Sottoriva[4], Trevor A. Graham[5], Leonardo Egidi[6], Biagio Ricciuti[7], Marcello Maugeri-Saccà[2] and Giulio Caravagna[1,8]

[1]*Department of Mathematics, Informatics and Geosciences, University of Trieste, Italy*
[2]*IRCCS Regina Elena National Cancer Institute, Rome, Italy*
[3]*Institute for Research in Biomedicine, Barcelona, Spain*
[4]*Centre for Computational Biology, Human Technopole, Milan, Italy*
[5]*Centre for Evolution and Cancer, Institute of Cancer Research, London, United Kingdom*
[6]*Department of Economics, Business, Mathematics and Statistics, University of Trieste, Italy*
[7]*Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts*
[8]*Area Science Park, Trieste, Italy*

Abstract

The intricate interplay between somatic mutations and copy number alterations critically influences tumour evolution and patient prognosis. However, traditional genomic analyses often treat these alterations independently, overlooking gene mutant dosage — a key emergent property of their interaction. Here, we develop an innovative computational framework that infers allele-specific copy number alterations directly from clinical targeted sequencing panels without requiring matched normal samples. Using this approach, we derived gene mutant dosage statistics for over 500,000 mutations across 60,000 clinical samples spanning 39 cancer types. By stratifying more than 20,000 patients according to mutant dosage across multiple oncogenes and tumour suppressor genes, we identified 46 tumour type–specific biomarkers predictive of overall survival. Notably, 13 of these biomarkers across 12 tumour types were undetectable using standard binary mutant/wild-type models. Additionally, 26 biomarkers were recurrently associated with metastatic spread in 10 tumour types, and 24 predicted organ-specific metastatic tropism in 6 tumour types. Alongside confirming known roles for established oncogenes and tumour suppressors, our method reveals, for the first time, gene mutant dosage patterns as independent predictors of prognosis, metastatic potential, and site-specific dissemination across diverse solid tumours. This augmented insight into genomic drivers enhances our understanding of cancer progression and metastasis, holding the potential to foster biomarker discovery significantly.

Keywords: *Gene mutant dosage, prognostic biomarker, metastatic tropism*

Corresponding author: Giulio Caravagna *E-mail address:* giulio.caravagna@units.it

This document is licensed under Creative Commons CC BY 4.0.

Contents

1 Introduction **2**

2 Results **2**

2.1 The INCOMMON mutation copy-number caller . . . 2

2.2 The prognostic effect of gene mutant dosage across 20,000 human cancers . . . 4

2.3 Gene mutant dosage predicts metastatic propensity and tropism . . . 7

3 Discussion **12**

4 Conclusion **13**

5 Extended Data Figures **24**

6 Supplementary Information **28**

6.1 Supplementary Tables . . . 28

6.2 Supplementary Figures . . . 28

1. Introduction

Cancerogenesis is a complex, multi-step process involving various genetic and epigenetic changes in tumour suppressor genes (TSGs) and oncogenes[1]. Central to this process are somatic mutations and copy number alterations (CNAs), a byproduct of genomic instability, which confer a proliferative advantage to specific cancer cell subpopulations[2], leading to the dominance of particular clones[3–6]. Genetic biomarkers derived from such driver mutations[7–9] have significantly impacted our understanding of cancer and its treatment, providing insights into cancer prognosis[10], and guiding treatment strategies against specific biological vulnerabilities. Despite several successes, however, understanding the prognostic and predictive implications of somatic mutations in key genes remains a challenge[11], promoting research towards other approaches such as that focus on epigenomic alterations[12,13].

The current approach to biomarker detection via clinical sequencing lacks of an integrative view of the tumour genome. Since mutation-derived biomarkers are usually investigated in isolation, the intricate interplay and epistasis within cancer genotypes where these alterations co-occur is potentially overlooked (Figure 1a). A clear example is the recent discovery of gene mutant dosage as a prognostic maker of pancreatic adenocarcinoma[14]. This statistic reports the number of copies of a mutation in the genome, and can be computed only by integrating mutation and allele-specific CNA data. The derivation of potent biomarkers from clinical sequencing is also complicated by other factors. For example, the number of samples available for every combination of tumour type, clinical condition and treatment regimen is still limited, given the high levels of tumour heterogeneity observed in patients[15]. Moreover, the detection of allele-specific CNAs from clinical targeted sequencing assays is complicated for experimental reasons such as the frequent lack of matched normal sample[16–19].

However, understanding and improving the detection of genetic events that can be exploited in the clinical setting remains crucial, prompting a re-evaluation of current tools and data, especially in light of the recent availability of large clinical cohorts of samples analysed for mutations in hundreds of cancer-related genes through targeted sequencing[20,21]. In this work, we developed INCOMMON, an open-source Bayesian inference tool to determine gene mutant dosage, as well as allele-specific CNAs and the mutant fraction statistics, for a set of mutations detected in a tumour-only clinical targeted sequencing assay. These statistics naturally reveal the joint effect of mutations and focal aneuploidy on TSG inactivation or oncogene activation, and are readily available for integrated biomarker discovery. To process targeted data efficiently, INCOMMON uses Bayesian priors derived from large-scale whole-genome sequencing (WGS) cohorts, together with a likelihood function for read-count data and bulk sample purity. In order to facilitate the usage of orthogonal metrics reported by pathologists in clinical tissue records, INCOMMON models the sample purity probabilistically, resisting potential discrepancies between histological and molecular purity estimates. Compared to state-of-the-art tools like ABSOLUTE[16], FACETS[17], PureCN[18], and CopyWriter[19], INCOMMON can work without raw data (FASTq/BAM files or pre-segmented depth ratios) or matched normal, facilitating its clinical implementation while lifting limitations of accessing large controlled-access datasets. .

We used INCOMMON to determine the patterns of co-occurrence of mutations and CNAs from over 500.000 mutations detected in 39 principal solid cancer types of over 60.000 clinical cancer samples from the MSK-MetTropism[22] (MSK) and AACR GENIE DFCI[23] (DFCI) cohorts. By unravelling gene mutant dosage statistics from large-scale clinical cohorts, we identified (i) 46 tumour-type specific biomarkers prognostic of overall survival across 12 tumour types (9 hotspot mutations), (ii) 26 frequently associated with metastatic spread in 10 tumour types (5 hotspot mutations) and (iii) 24 that predict organotropism in 6 tumour types (2 hotspot mutations). Besides confirming established results regarding known TSGs and oncogenes, for the first time, our method reveals specific gene mutant dosage patterns as independent predictors of patient prognosis, metastatic propensity, and organ-specific metastatic spread across a wide range of solid tumours.

2. Results

2.1. The INCOMMON mutation copy-number caller

We developed INCOMMON – *inference of copy number and mutation multiplicity for oncology* –, a Bayesian model that can infer, for a tumour mutation, (i) the total copy number at the mutant locus, and (ii) the mutation multiplicity (i.e., the number of DNA copies that harbour the mutation). This information provides the allele-specific configuration of the mutant locus, the mutant gene dosage and fraction, as canonically determined with copy number calling algorithms. INCOMMON takes input read counts data for n mutations $X = \{x_1, \dots, x_n\}$ to develop the joint likelihood

$$p(X \mid \Theta) = \prod_{x_i \in X} p(x_i \mid \Theta) = \prod_{x_i \in X} p(d_i \mid \Theta)\, p(r_i \mid d_i, \Theta)\,. \tag{1}$$

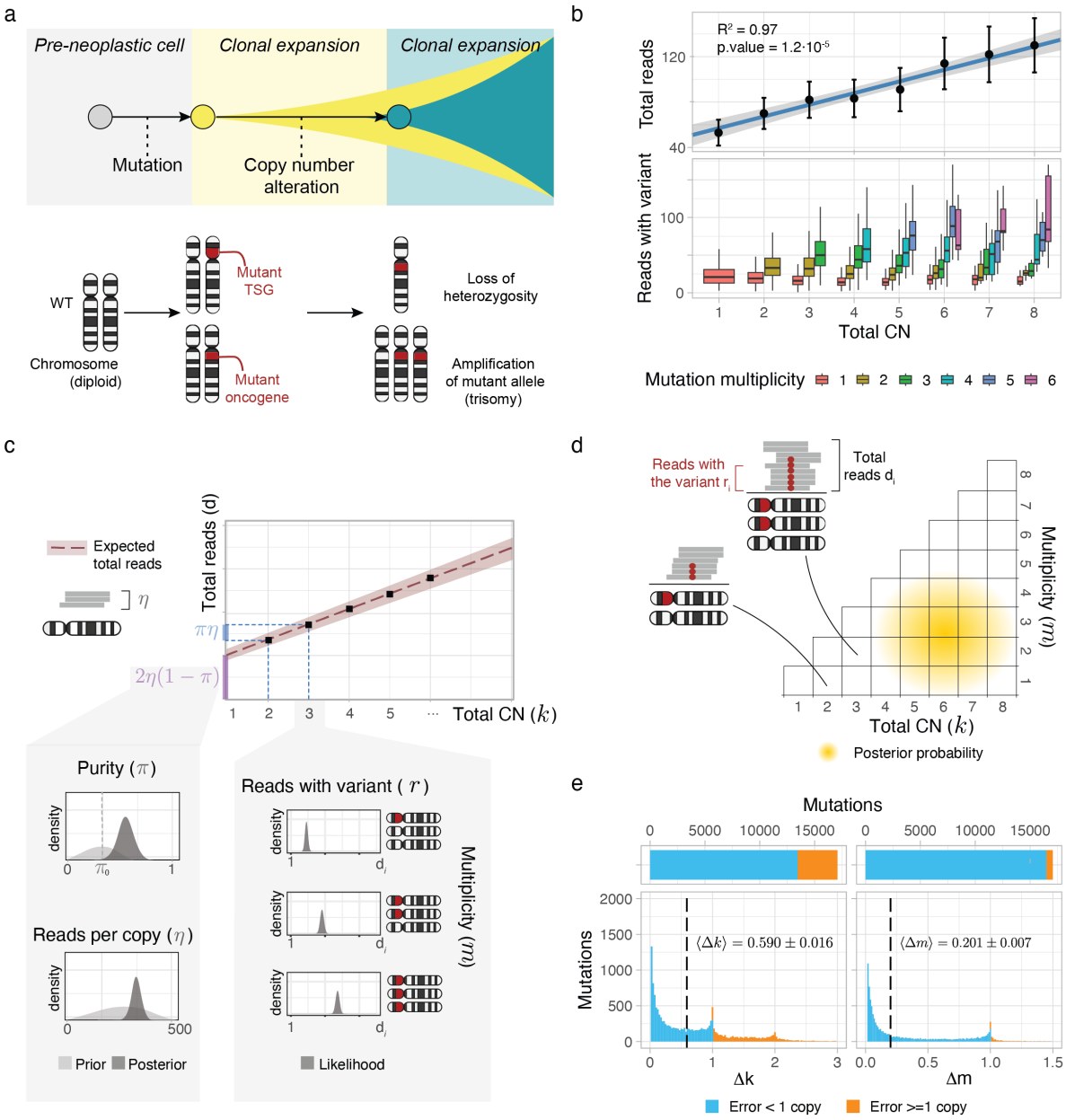


Figure 1. a. Clonal expansions from a normal cell that acquires first a mutation in a tumour suppressor gene (TSG) or an oncogene, then a copy number alteration (CNA) on the mutant locus, pictured as a loss of the wild-type (WT) allele or a gain of the mutant. b. Linear correlation between total read counts and mutation copy number (top panel), and distribution of reads with the variants (bottom panel) for increasing mutation multiplicity values, for $n = 8{,}339$ high-resolution WGS samples available from the PCAWG, HMF and TCGA cohorts. c. INCOMMON is a Bayesian model that leverages a biology-informed prior and infers the posterior distribution of two sample-level parameters (tumour purity π and the rate of reads per chromosome copy η) and two mutation level parameters (total copy number k and mutation multiplicity m) from the number of reads with the variant r and the total number of reads d of each mutation in the sample. d. INCOMMON assigns a posterior probability to any possible configuration of a mutant gene, determined by the co-occurrence of mutations and copy number alterations. e. Performance of INCOMMON across $N = 50$ cross-validation iterations of random-split cross-validation (70% training, 30% test) using ground truth data from PCAWG[24] and HMF[25]. The deviation from true values of m and k, Δk and Δm, was less than one copy on average, occurring in 78.6% $\pm$ 0.9% of k predictions and 96.4% $\pm$ 0.4% of m predictions.

For every mutation $x_i = \langle r_i, d_i \rangle$, we use the number of reads r_i with the alternative allele and the total reads d_i (depth of sequencing). The model infers two sample-specific parameters, (i) the sample purity ($0 < \pi \leq 1$), and (ii) the rate of reads per chromosome copy ($\eta \in \mathbb{R}^+$), and two mutation-specific parameters, (iii) the tumour total copy number at the locus ($k_i \in \mathbb{Z}^+$) and (iv) the mutation multiplicty ($m_i \in \mathbb{Z}^+, m_i \leq k_i$). The model adopted by INCOMMON relates copy number to coverage linearly – the expected number of reads for k chromosome copies is $k\eta$– following the observation of a significant linear correlation ($R^2 \geq 90\%$ with p-value $P < 0.05$; Figure 1b) in 8,339 samples from the large-scale Pan-Cancer Analysis of Whole Genomes[24] (PCAWG), Hartwig Medical Foundation[26] (HMF) and The Cancer Genome Atlas[27] (TCGA) cohorts, and following earlier RNA-based work[28–30].

INCOMMON links the number of reads to the multiplicity and total copy number (Figure 1c), considering tumour/ normal admixing

$$p(d_i \mid \pi, \eta, k_i) = \text{Poisson}(d_i \mid \lambda) \qquad \lambda = (1-\pi)2\eta + \pi\eta k_i \tag{2}$$

$$p(r_i \mid \pi, k_i, m_i) = \text{Binomial}(r_i \mid d_i, \varphi) \qquad \varphi = \frac{m_i \pi}{2(1-\pi) + k_i \pi} \tag{3}$$

In this model, the sequencing depth follows a Poisson distribution with an expected number of reads λ defined by combining tumour and normal readouts. Given the depth, the number of mutant reads follows a Binomial distribution with success probability φ determined by the mixing of tumour and normal success rates. This construction induces overdispersion naturally through hierarchical modelling, avoiding the need for an arbitrary overdispersion parameter—a key advantage in terms of parsimony and interpretability. A graphical representation of the model, its priors and likelihood is shown in Supplementary Figure S1.

INCOMMON uses Markov Chain Monte Carlo sampling to estimate a posterior distribution over m, k, η and π (Figure 1d). To leverage the massive amount of public WGS data of human tumours and gain precision with targeted assays, we derived biologically informed prior distributions for (k, m) configurations from the PCAWG and HMF cohorts (examples for the most common alterations in Supplementary Figure S2, full distribution in Supplementary Table S1). For frequently mutated genes (e.g., *PIK3CA* and *TP53* in breast cancer, *KRAS* in pancreatic cancer), these priors were gene- and tissue-specific (Supplementary Figure S3, Methods). Following a similar logic, the prior on η was derived empirically from the sequencing depth distribution of the analysed datasets (Supplementary Figure S4, Supplementary Table S2). Notably, we also modelled a prior for the input purity π, an essential parameter to estimate copy numbers. To support orthogonal data (e.g. from histopathological evaluation) but resist potential error in the input estimate, we centred a broad prior around the purity measurements provided with each sample (the 95% confidence interval lying within ±10-15 percentage points of the estimate). After posterior inference, INCOMMON uses posterior predictive checks to monitor the discrepancy between observed values and inferred posterior distributions on k, m, η and π. This information is used to annotate rare events (e.g. unexpectedly high copy numbers or multiplicities) and flag poor fits for the model, before assembling all data in a detailed report of the analysis (Extended Data Figure 1).

We evaluated model performance using $N = 50$ iterations of random-split cross-validation (70% training, 30% test) across combined data from the PCAWG and HMF data ($n = 6{,}492$ samples, from 46 tumour types, Figure 1e and Supplementary Figure S5). The model achieved a mean absolute error (MAE) of $\Delta k = 0.590 \pm 0.016$ for total copy number and $\Delta m = 0.201 \pm 0.007$ for mutation multiplicity, corresponding to deviations of less than one copy on average (i.e., one copy miscalled). In fact, $78.6 \pm 0.9\%$ of k predictions and $96.4 \pm 0.4\%$ of m predictions fell within ±1 copy of the true value, underscoring the model's high predictive accuracy. The narrow variability across folds further highlighted the robustness and consistency of the estimates, supporting their reliability for downstream biological interpretation. We also compared INCOMMON calls with FACETS results for lower (LCN), major (MCN), and total copy number (TCN), which were publicly available for the MSK-IMPACT[31] cohort, a subset of the whole MSK-MET cohort. While these statistics are lower-resolution than mutation multiplicities (unavailable in FACETS), we observed a good concordance from mean absolute deviations, which were 0.738, 1.00 and 1.94, respectively. Interestingly, concerning tumour purity, the two methods tended to adjust the pathologist's estimate consistently, with a mean deviation of 9.72% between the two. INCOMMON also demonstrated high computational efficiency, requiring under 12 minutes and less than 200 MB of RAM to process samples containing up to 100 mutations (Supplementary Table S3). Runtime scaled linearly with input size, supporting practical deployment in large-scale genomic pipelines.

2.2. The prognostic effect of gene mutant dosage across 20,000 human cancers

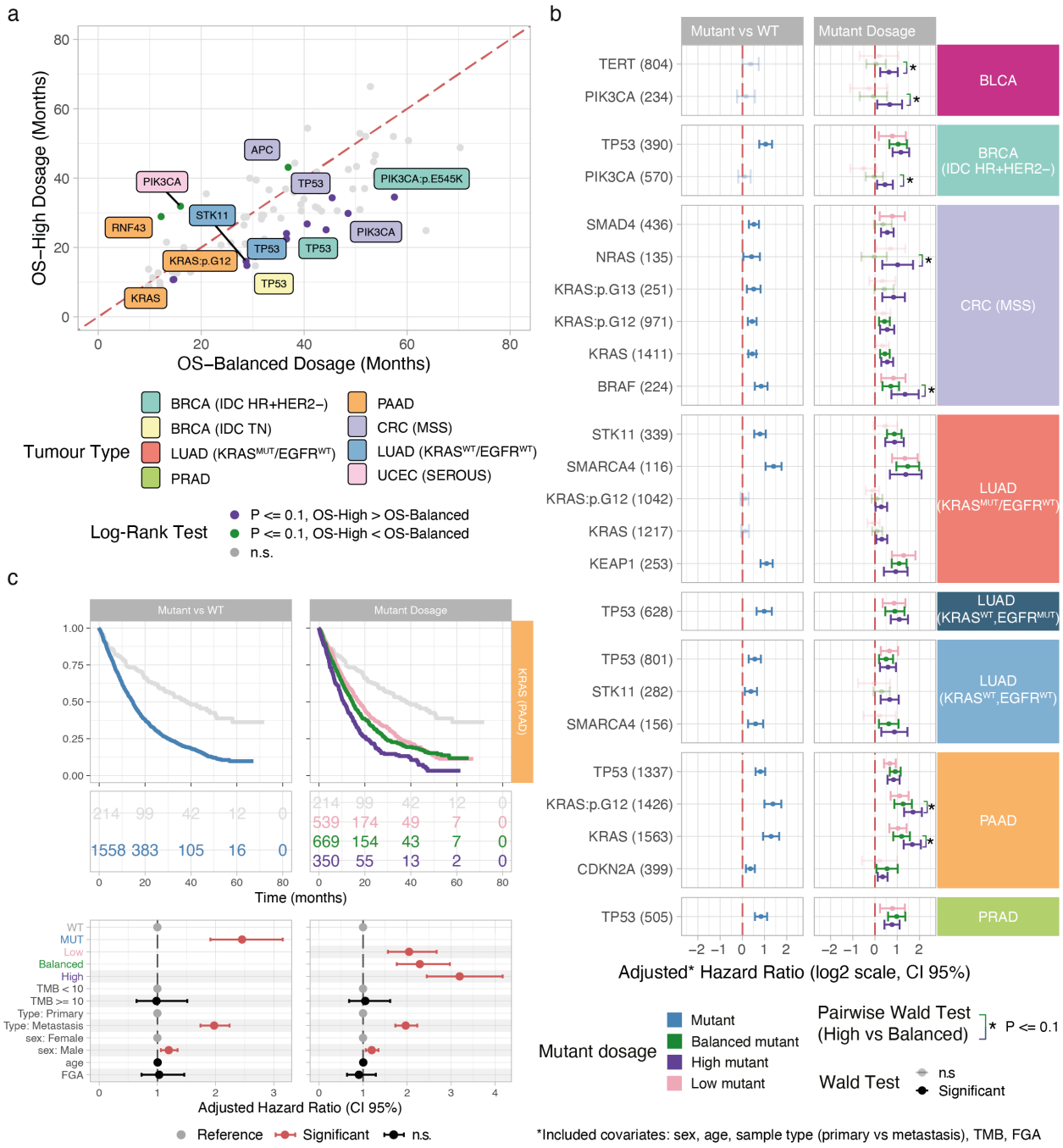


Figure 2. a. Comparison of median OS between high and balanced mutant dosage classes, highlighting markers with significant (Log-Rank test pairwise p-value $P \leq 0.1$) OS difference. b. Adjusted hazard ratios from multivariate Cox regression accounting for sex, age, sample type (primary vs metastasis), tumour mutational burden (TMB) and fragment of genome altered (FGA), comparing canonical analysis of mutant vs wild-type tumours (left panel) with mutant dosage analysis (right panel). Markers with significant (Wald test, FDR-adjusted p-value $P \leq 0.1$) negative prognostic impact associated with high mutant dosage are shown, highlighting those with significant difference relative to the balanced dosage class. c. Comparison of Kaplan–Meier curves of OS (top panels) and adjusted hazard ratios between mutant vs wild-type (left panel) and mutant dosage stratification (right panel) for KRAS in pancreatic adenocarcinoma (PAAD).

We used INCOMMON analysis from $n = 21{,}937$ tumours of the MSK-MET cohort to determine gene mutant dosage and its prognostic potential against overall survival (OS). To compare our approach with the canonical stratification into mutant (MUT) and wildtype (WT) tumours, we introduced a new statistic termed fraction of alleles carrying the mutation (FAM)

$$\mathrm{FAM}_{g,t} = \sum_{k=1}^{k_{max}} \sum_{m=1}^{k} \frac{m}{k} p(m, k \mid X_{g,t}), \tag{4}$$

which is derived from the full posterior distribution of $p(m, k \mid X_{g,t})$, computed by INCOMMON. This score, for a gene g and tumour type t, was paired with two cutoffs to determine high (FAM $\geq$ 75% for TSG, FAM $\geq$ 66% for oncogenes), low (FAM $\leq$ 25% for TSG and FAM $\leq$ 33% for oncogenes), and balanced dosage classes that maximise the rank statistics of the groups (Methods; Extended Data Figure 2). These groups are consistent with the canonical interpretation of allele-specific CNAs: the high dosage class includes, for example, deletions of the WT allele leading to loss of heterozygosity (LOH), typically occurring in TSGs[31–33], and gains of the mutant allele frequently affecting oncogenes[31,32,34,35], possibly combined as in copy-neutral LOH states; the balanced class includes mostly heterozygous diploid states, but also states of high ploidy with a similar number of mutant and WT alleles; low dosage reflects mutations following copy gains or post-mutation copy number events favoring the WT allele.

We first established the distribution of the maximum a posteriori values of read count per copy η, mutation copy number k and multiplicity m, FAM and sample purity across 286 genes and 50 tumour types in the MSK-MET dataset (Supplementary Figure S6). We then investigated the prognostic power of the classes derived from gene mutant dosage. Given the known different prognosis of cancer subtypes, we separately analysed those supported by sufficient sample numbers: mutant/wild-type *KRAS* and *EGFR* lung cancers, breast cancer molecular subtypes, micro-satellite stable (MSS) and hyper-mutant (HM) colorectal cancers, as well as endometrioid, serous and HM uterine cancers. A preliminary univariate analysis showed that the median OS was significantly different (false discovery rate, FDR, adjusted p-value $P \leq 0.1$, Log-rank test) across mutant dosage classes in 16 cases. This covered 7 mutated genes (*APC*, *KRAS*, *NRAS*, *PIK3CA*, *RNF43*, *STK11*, *TP53*) regardless of the specific mutation position, and 2 hotspot mutations (*KRAS G12* and *PIK3CA E545K*), for a total of 8 tumour types (Figure 2a). Interestingly, the lowest median OS was associated with the high dosage class in 75.0% of cases, and only rarely with the balanced and low classes (16.7% and 8.3%, respectively; Extended Data Figure 2; full analysis in Supplementary Table S4).

Next, we tested this prognostic effect in a multivariable Cox model for OS adjusted for sex, age at diagnosis, sample type (primary versus metastasis), as well as two well-known biomarkers: the tumour mutation burden (TMB)[36–38] and the overall aneuploidy[10,39–41], measured as the fraction of genome altered (FGA). Since the association between gene dosage classes and survival may be confounded by higher TMB and FGA, adjusting for these two covariates was essential to understand the actual impact of gene mutant dosage statistics (Supplementary Figure S7). Our analysis showed that high mutant dosage was a negative prognostic factor in 24 cases (11 genes and 3 hotspots) across 8 tumour types (Figure 2b; full analysis in Supplementary Table S5).

Among these, we searched for scenarios where the high dosage classes were significantly more associated with worse prognosis (higher adjusted hazard ratio, HR) compared to balanced ones. We found this strong trend for *KRAS* in pancreatic adenocarcinoma ($HR = 3.19$, $CI = 2.45$-4.17, $P < 0.0001$, Figure 2c), including G12 hotspot mutations ($HR = 3.27$, $CI = 2.48$-4.30, $P < 0.0001$), *BRAF* ($HR = 2.55$, $CI = 1.67$-3.90, $P = 0.0001$) and *NRAS* ($HR = 2.02$, $CI = 1.25$-3.27, $P = 0.0144$) in MSS colorectal cancer, *PIK3CA* ($HR = 1.58$, $CI = 1.07$-2.33, $P = 0.0945$) and *TERT* ($HR = 1.55$, $CI = 1.18$-2.03, $P = 0.0140$) in bladder cancer, and *PIK3CA* ($HR = 1.36$, $CI = 1.07$-1.74, $P = 0.0325$) in HR+HER2- invasive ductal breast cancer (Figure 3a). Notably, *PIK3CA* and *TERT* in bladder and breast cancers (Figure 3a) did not emerge as significant markers when using a canonical stratification of mutant versus wildtype patients, highlighting the enhanced resolution provided by gene mutant dosage.

A similar result occurred in specific lung cancer subtypes. For instance, high mutant *KRAS* dosage ($HR = 1.23$, $CI = 1.04$-1.46, $P = 0.0384$, Figure 3b) and its G12 hotspot mutations ($HR = 1.22$, $CI = 1.02$-1.46, $P = 0.0328$) were significanty prognostic in *EGFR*-wildtype adenocarcinomas, even if in this case we found only mild difference for the non-prognostic balanced class ($HR = 1.06$, $CI = 0.899$-1.26, $P = 0.474$).

The possibility of zooming into a mutant dosage analysis allows us to detect new prognostic classes which might be obscured in canonical analysis (mutant versus wildtype), increasing our biological understanding of tumour genomes. An exemplary case was that of *STK11* in lung adenocarcinomas that are wildtype for the major oncogenes *KRAS* and *EGFR* (Figure 3c). In this case, we observed that only high mutant dosage *STK11* was associated with negative prognosis ($HR = 1.58$, $CI = 1.19$-2.09, $P = 0.0115$). Similarly, in mutant *SMAD4* colorectal MSS tumours,

only high ($HR = 1.46$, $CI = 1.20$-1.78, $P = 0.0006$) and low ($HR = 1.71$, $CI = 1.16$-2.54, $P = 0.0431$) $SMAD4$ classes were prognostic, highlighting the strong impact of copy number alterations compared to plain $SMAD4$ mutations. Conversely, in mutant $KRAS$ tumours – particularly those with G12 hotspot mutations – a low mutant dosage was not prognostic, unlike a balanced ($HR = 1.36$, $CI = 1.17$-1.58, $P < 0.0001$) or high dosage class ($HR = 1.46$, $CI = 1.22$-1.76, $P = 0.0003$), which were associated with an increased risk of death.

For the remaining markers where the high-dosage class was prognostic, we did not observe significant differences across dosage classes in univariate or multivariate analyses. The only exceptions were $TP53$ in certain breast, lung and prostate cancer subtypes (Figure 3d), where our mutant dosage stratification was significant in univariate analysis, but lacked robustness in multivariate analysis. In these particular cases, FGA and sample type accounted for most of the prognostic effect, suggesting at least strong collinearity between the presence of genome instability (high FGA) and $TP53$ mutations. In fact, we found that the FGA was progressively larger from low to balanced and up to high gene mutant dosage, the difference across dosage classes being significant (p-value $P \leq 0.1$, Wilcoxon test) in HR+HER2- and triple-negative breast cancer, MSS colorectal cancer, hepatocellular carcinoma, $KRAS$ wiltype lung adenocarcinoma, pancreatic and prostate cancer (Extended Data Figure 2).

2.3. Gene mutant dosage predicts metastatic propensity and tropism

We next examined how gene mutant dosage impacts metastatic propensity and organotropism. First, we compared the distribution of gene mutant dosage between primary ($n = 13{,}216$) and metastatic ($n = 8,721$) tumour samples from the MSK-MET cohort.

Our analysis revealed an uneven distribution of INCOMMON gene mutant dosage classes between primary and metastatic tumors from 27 tumour types, 95 genes and 21 hotspot mutations (Figure 4), using a two-tailed Fisher's exact test at significance level $\alpha = 0.1$, together with a score r from the standardised residuals). Although the median enrichment was small for all classes ($r = 0.732$ for low mutant dosage in primary tumours, $r = 0.136$ for balanced dosage and $r = 0.391$ for high dosage in metastases), the distribution of scores was significantly different across classes (Extended Data Figure 3). As a general trend, low mutant dosage classes were more frequent in primary tumours, whereas high mutant dosage classes were enriched in the metastases (Supplementary Figure S8; full analysis in Supplementary Table S6).

We observed a consistent enrichment of high mutant dosages in metastatic samples and low mutant dosages in primary samples, for both $TP53$ and $PIK3CA$ in breast and colorectal cancers. We found a similar enrichment pattern for $TP53$ also in hepatocellular carcinomas, in pancreatic and prostate cancer, in melanoma, and in lung cancers together with $KRAS$, $EGFR$ and $STK11$; for $PIK3CA$ in uterine endometroid cancer along with $AIRD1A$. Interestingly, in hyper-mutant uterine cancers, low dosages of $POLE$ were enriched in primary samples and high dosages in the metastases.

Additionally, high mutant dosage was enriched in several metastatic contexts: APC in prostate cancer metastases; $KEAP1$ in lung cancer metastases lacking $KRAS$ or $EGFR$ mutations; $TP53$ in small cell lung cancer metastases; and $KMT2A$ in melanoma metastases. Interestingly, in primary melanoma tumours, low dosages of $KMT2D$, $NTRK3$, $PREX2$, $ROS1$, $TP63$ and $ZFHX3$ were more frequent. Low dosages of $ZFHX3$ were enriched also in primary prostate cancers, and of $KRAS$ in pancreatic cancers.

Overall, the strongest enrichment score in metastatic samples was for $TP53$ in MSS colorectal ($r = 8.52$, $P < 0.0001$) and prostate cancers ($r = 6.23$, $P < 0.0001$), and for $PIK3CA$ in HR+HER2- breast cancers ($r = 5.33$, $P < 0.0001$). We validated our results on the MSK-MET cohort using $n = 21{,}462$ primary and $n = 12{,}049$ metastasis samples from the GENIE-DFCI cohort (Extended Data Figure 3, Supplementary Figure S9). A large fraction (89.29%) of the 56 significant markers that we found in MSK-MET had a consistent enrichment score between the two cohorts (inter-chort variation lower than intra-cohort, Methods), confirming the general trend of low mutant dosage to be more frequent in primary tumours, and high mutant dosages in metastases across a total of $n = 55{,}448$ clincal samples.

To investigate the impact of gene mutant dosage on the metastatic propensity, we calculated the odds ratio (OR) of primary tumours with low, balanced or high mutant gene dosage, to metastasise relative to wild-type (WT) tumours. In a preliminary univariate analysis, we found 12 genes and 4 specific hotspots with significant OR across 10 tumour types (Figure 5a). High mutant dosage increased the metastatic propensity in $n = 9$ cases (20.5% of the significant ones), and decreased it in $n = 5$ cases (Extended Data Figure 3). Low mutant dosage was primarily associated with reduced metastatic propensity ($n = 11$ cases), with only one case having a negative impact. The balanced class had, in most cases, a similar trend of the corresponding high or low counterpart, but with a less pronounced effect. To account for potential confounders, we applied a multivariable logistic regression model incorporating sex, age at

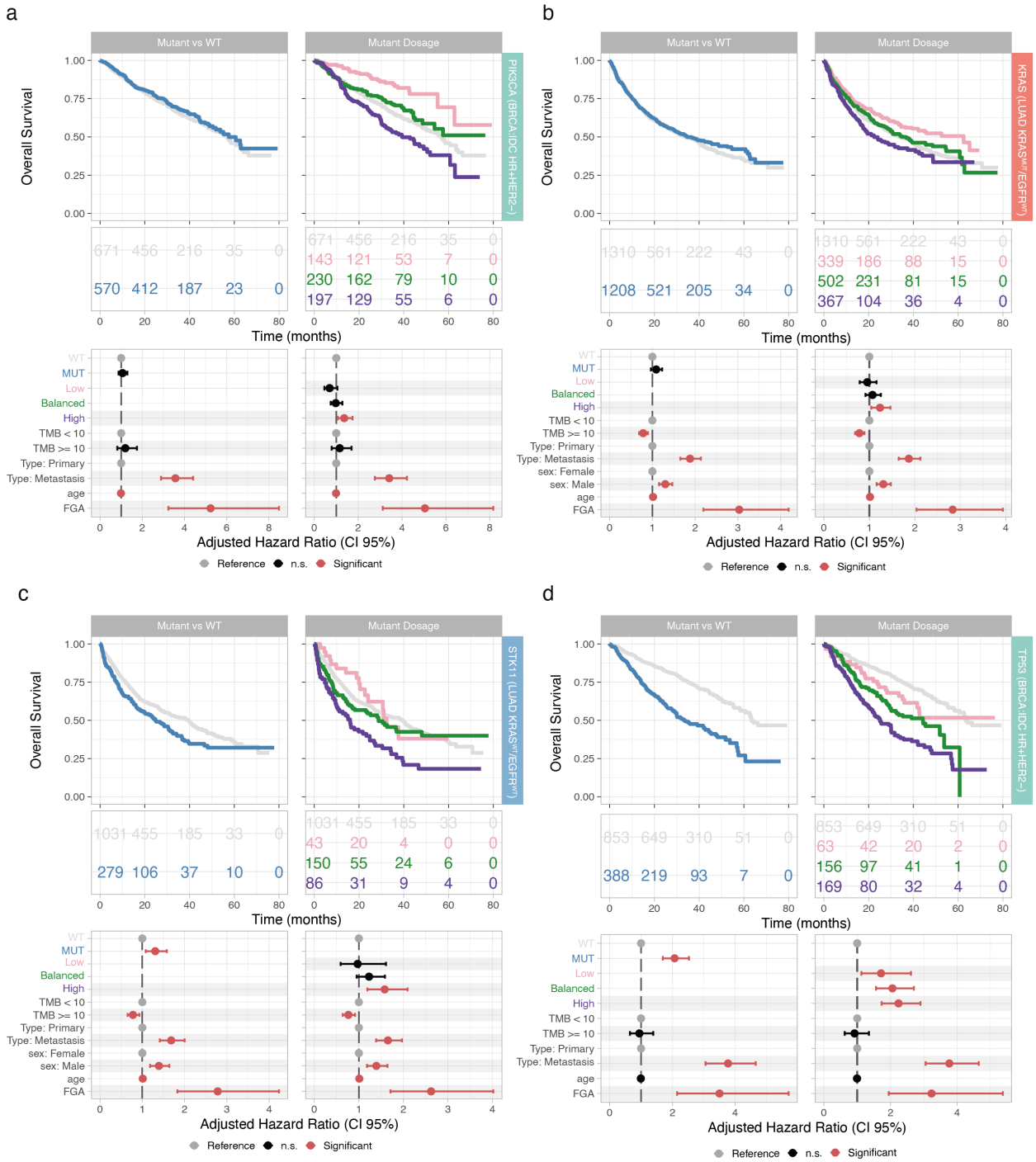


Figure 3. a. Comparison of Kaplan–Meier curves, annotated with numbers at risk, of overall survival (top panel) and adjusted hazard ratios (with 95% confidence interval, bottom panel) between mutant vs wild-type (left panel) and mutant dosage stratification (right panel), as in Figure 2e, for PIK3CA mutant patients with invasive ductal breast HR+HER2- carcinoma. For each patient group, the forest plot indicates whether the corresponding regression coefficient is significant (red) or not (black). The regression coefficient for the reference group (grey) is 1. b. Results for KRAS in EGFR wild-type lung adenocarcinoma. c. Results for STK11 in KRAS wild-type/EGFR wild-type lung adenocarcinoma. d. Results for TP53 in invasive ductal breast HR+HER2- carcinoma.

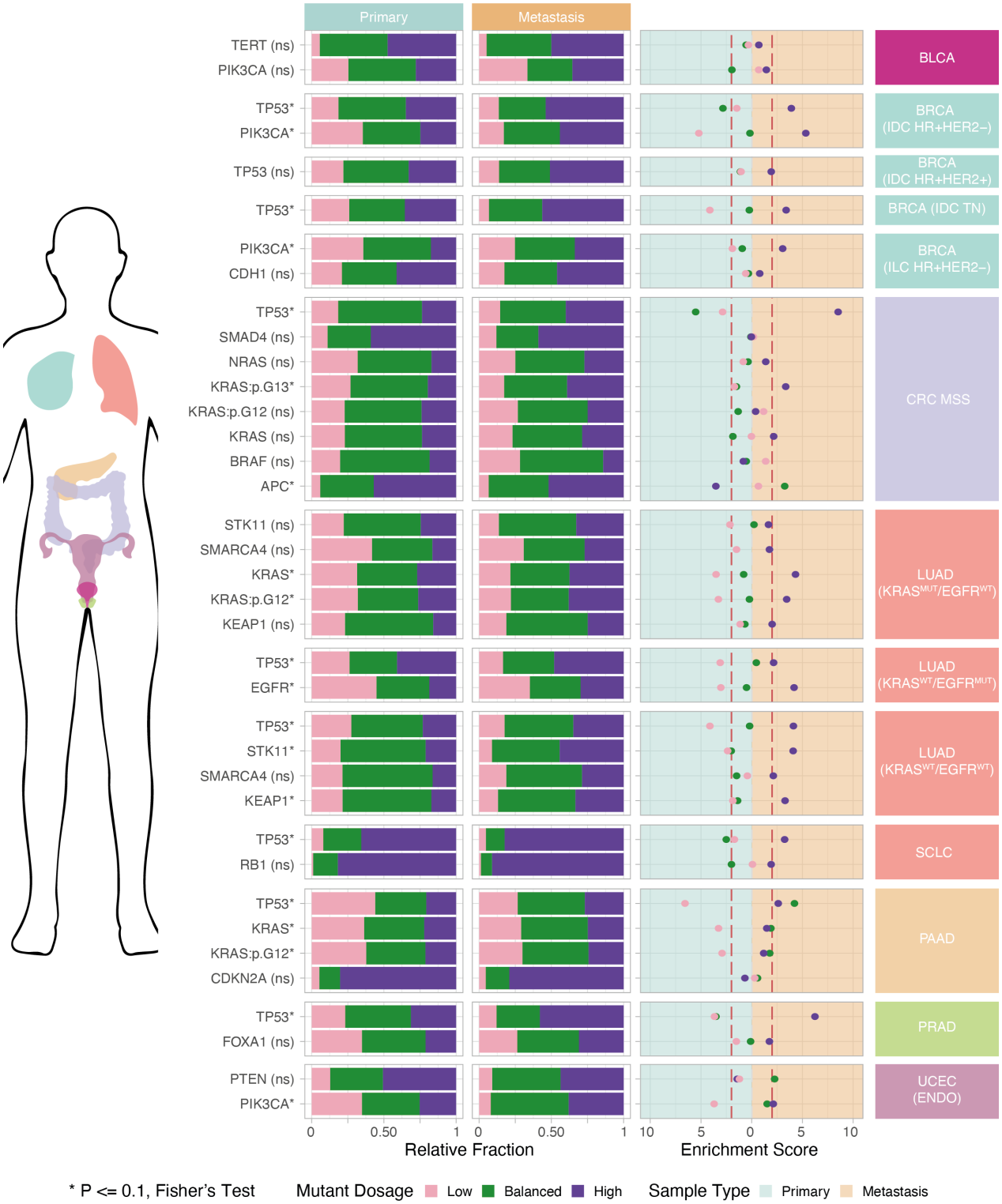


Figure 4. Distribution of low, balanced and high mutant dosage classes across 22 genes and 13 tumour subtypes, compared between primary and metastatic tumour samples (left panels). The right panel shows enrichment scores r for each class in either primary or metastatic tumours, evaluated through two-tailed Fisher's exact test with significance level $\alpha = 0.1$ and using standardised residuals. Vertical dashed lines indicate thresholds of large effect size ($r = 2$).

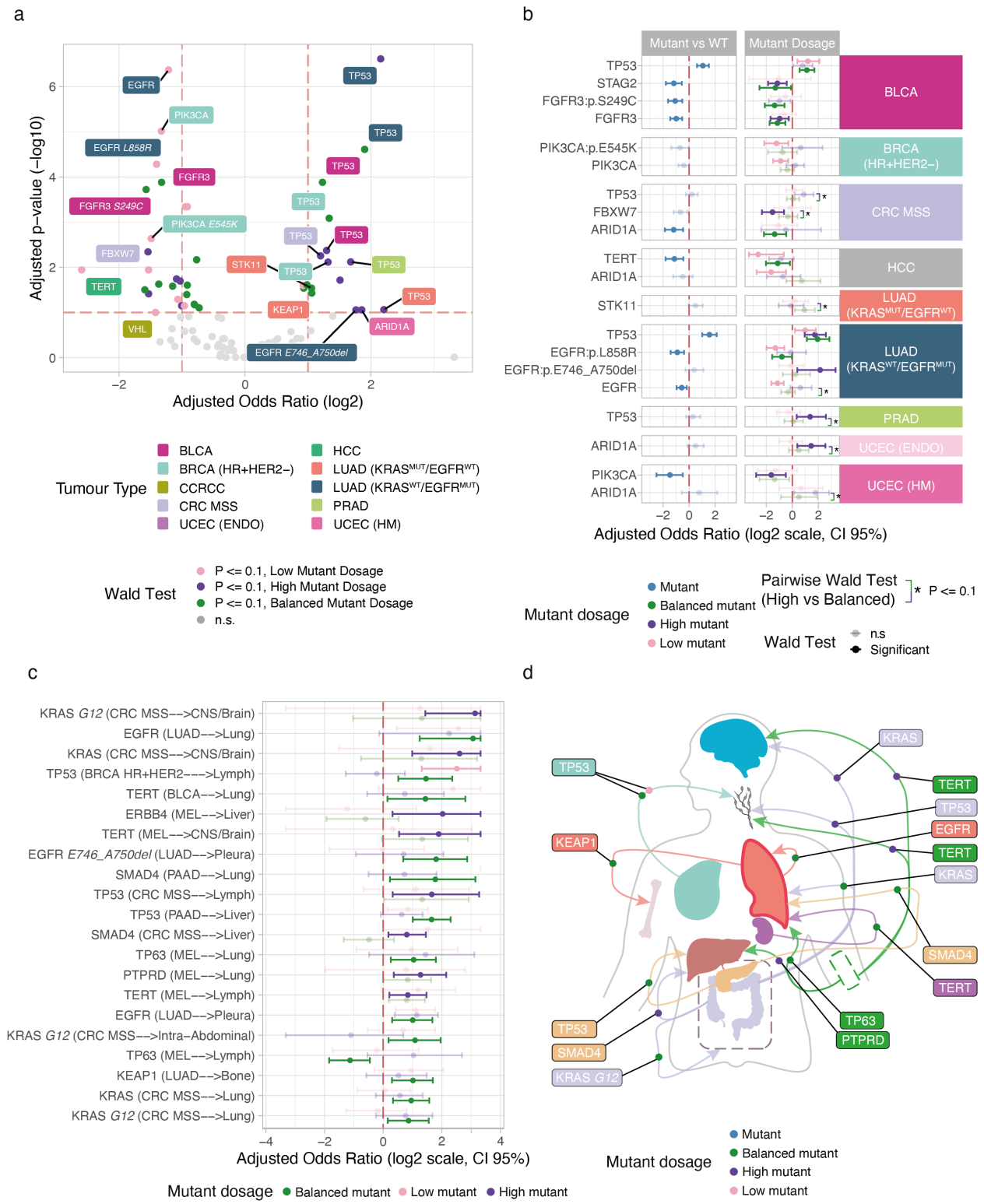


Figure 5. a. Odds ratio (OR) of primary tumours with low (pink dot), high (purple dot) or balanced (green dot) mutant gene dosage, of giving rise to metastasis. The FDR-adjusted p-value (-log10 scale) against the OR (log2 scale) is reported. Lables highlighting genes and specific hotspots, with a significant impact on metastatic propensity, are colored according to the primary tumour type. b. Adjusted OR from multivariate logistic regression accounting for sex, age, tumour mutational burden (TMB) and fragment of genome altered (FGA), comparing canonical analysis of mutant vs wild-type tumours (left panel) with mutant dosage analysis (right panel). Markers with significant (Wald test, FDR-adjusted p-value $P \leq 0.1$) impact on metastatic propensity associated with mutant dosage are shown, highlighting those with significant difference between high and balanced dosage classes. c. Adjusted OR of metastatic tropism from primary tumours to other organs associated with mutant dosage. Significant cases are colored. d. Organotropic patterns associated with gene mutant dosage, represented by arrows starting from the primary site and pointing to the metastatic site. Dots are colored accordingly to the dosage class associated with the organotropic pattern, while labels highlight significant genes and specific hotspots, and are colored accordingly to the primary tumour type.

sequencing, sample type, TMB and FGA, determining statistical significance using a Wald test with FDR-correction (Methods; Supplementary Figure S10; full analysis in Supplementary Table S7). The impact on metastatic propensity remained significant in 17 cases, including 8 genes and 4 hotspots across 8 tumour types (Figure 5b).

The high mutant dosage class was the only class significantly associated with higher metastatic propensity in several cases. For *EGFR* (*E746_A750del* mutations) in *KRAS*-wildtype lung cancers, with more than 4 times the odds to metastasise relative to wildtype tumours (adjusted odds ratio $OR = 4.42$, $CI = 1.31\text{-}27.5$, adjusted p-value $P = 0.0884$). The signal was also clear for *TP53* in prostate adenocarcinoma ($OR = 2.59$, $CI = 1.27\text{-}6.00$, $P = 0.0450$) and *ARID1A* in endometrial cancers ($OR = 2.73$, $CI = 1.31\text{-}5.83$, $P = 0.0476$). In the latter case, the increase in metastatic propensity relative to the balanced dosage class was also significant (p-value $P = 0.100$, pairwise Wald test). Remarkably, these genes would not have emerged as markers of metastatic propensity if analysed without considering mutant dosage, i.e. by only comparing mutant vs. WT tumours.

Consistent with results from the univariate analysis, a protective effect was also suggested for high mutant dosage classes associated with a low metastatic risk (OR below 1). Examples in bladder cancer were mutations in *FGFR3* ($OR = 0.511$, $CI = 0.315\text{-}0.834$, $P = 0.0629$) and *STAG2* ($OR = 0.452$, $CI = 0.276\text{-}0.743$, $P = 0.0312$). Notably, *FBXW7* mutations in colorectal MSS cancers ($OR = 0.350$, $CI = 0.199\text{-}0.644$, $P = 0.00734$) and *PIK3CA* mutations in uterine cancer ($OR = 0.327$, $CI = 0.146\text{-}0.711$, $P = 0.0381$) reduced metastatic risk by one-third.

More commonly, low dosage mutations associated with a reduced incidence of metastatic seeding. This was statistically significant for *TERT* ($OR = 0.163$, $CI = 0.0392\text{-}0.635$, $P = 0.0111$) and *ARID1A* mutations ($OR = 0.324$, $CI = 0.151\text{-}0.716$, $P = 0.0111$) in hepatocellular carcinoma, for *PIK3CA* mutations ($OR = 0.541$, $CI = 0.363\text{-}0.806$, $P = 0.0177$) including the *E545K* hotspot ($OR = 0.432$, $CI = 0.226\text{-}0.811$, $P = 0.0340$) in HR+HER- breast cancers, and for *EGFR* mutations ($OR = 0.457$, $CI = 0.329\text{-}0.636$, $P < 0.0001$) with the *L858R* hotspot ($OR = 0.406$, $CI = 0.254\text{-}0.654$, $P = 0.0004$) in *KRAS*-wildtype lung adenocarcinoma. The balanced dosage class was the only class significantly associated with reduced metastatic propensity, with a similar effect for *FGFR3* (*S249C* mutations) in bladder cancer (0.392, $CI = 0.232\text{-}0.661$, $P = 0.00271$) and *ARID1A* in MSS colorectal cancer ($OR = 0.392$, $CI = 0.222\text{-}0.726$, $P = 0.0308$). Mutant *TP53* tumours were generally more prone to metastasise than WT, irrespective of the mutant dosage, as in bladder cancer, *KRAS*-wildtype lung adenocarcinoma with *EGFR* mutations, and in MSS colorectal cancer. In the latter case, while an increase in metastatic propensity for the high dosage relative to the balanced state was observed, its statistics remained below significance.

We then analysed organ-specific patterns of metastasis (Figure 5c,d), using a similar multivariable logistic regression model incorporating sex, age at sequencing, TMB and FGA (Methods; Supplementary Figure S11; full analysis in Supplementary Table S8). We found 29 significant patterns associated with mutant dosage, covering 13 genes and 1 hotspot across 6 primary tumours and 8 metastatic sites.

Strikingly, the odds of metastatic spread from colorectal MSS cancers to the central nervous system/brain metastases were more than sixfold for high dosage *KRAS* mutations ($OR = 6.08$, $CI = 1.98\text{-}10$, $P = 0.0121$), reaching a eightfold for *G12* mutations ($OR = 8.77$, $CI = 2.70\text{-}10$, $P = 0.00403$). Conversely, lung metastases of these colorectal cancers more frequently harboured balanced *KRAS* mutant classes ($OR = 1.94$, $CI = 1.26\text{-}2.97$, $P = 0.0358$), even considering *G12* mutations ($OR = 1.82$, $CI = 1.11\text{-}2.93$, $P = 0.0788$), which were also associated with higher likelihood of intra-abdominal metastases ($OR = 2.12$, $CI = 1.14\text{-}3.88$, $P = 0.0788$). High mutant dosages of *TP53* and *SMAD4* instead led preferentially to the lymphatic system ($OR = 3.15$, $CI = 1.25\text{-}9.65$, $P = 0.0928$) and the liver ($OR = 1.75$, $CI = 1.13\text{-}2.74$, $P = 0.0574$), respectively. High mutant dosage also increased the likelihood of melanoma to metastasise, by around fourfold for *ERBB4* mutations spreading to the liver ($OR = 4.07$, $CI = 1.25\text{-}11.5$, $P = 0.0998$) and *TERT* mutations spreading to the central nervous system/brain ($OR = 3.71$, $CI = 1.46\text{-}10.8$, $P = 0.0998$), and by around twofold for *PTPRD* to the lung ($OR = 2.42$, $CI = 1.28\text{-}4.44$, $P = 0.0998$) and *TERT* to the lymphnodes ($OR = 1.79$, $CI = 1.16\text{-}2.78$, $P = 0.0998$). Lung metastases from melanoma also frequently harboured *TP63* mutations with balanced mutant dosage ($OR = 2.05$, $CI = 1.20\text{-}3.47$, $P = 0.0961$).

Interestingly, *EGFR* mutations with balanced dosage were more prone to metastasise locally in lung adenocarcinoma, with a eightfold increase in the likelihood for lung metastases ($OR = 8.35$, $CI = 2.37\text{-}30.8$, $P = 0.0470$) and a nearly fourfold increase for *E746_A750del* mutations leading to metastases in the pleura ($OR = 3.49$, $CI = 1.60\text{-}7.28$, $P = 0.0470$). Bone metastases were instead more likely from this adenocarcinoma when they harboured *KEAP1* mutations with balanced dosage ($OR = 2.02$, $CI = 1.23\text{-}3.22$, $P = 0.0887$).

Pancreatic tumours preferentially metastasised to the lung when harbouring balanced *SMAD4* mutations ($OR = 3.43$, $CI = 1.17\text{-}8.85$, $P = 0.0951$), while balanced *TP53* mutations were associated with liver metastases ($OR = 3.14$, $CI = 2.01\text{-}4.92$, $P < 0.0001$). In the context of HR+HER2- breast cancer, balanced dosage of *TP53* mutations

increased the likelihood of spread to the lymphatic system ($OR = 2.75$, $CI = 1.43\text{-}5.12$, $P = 0.0127$).

3. Discussion

In this study, we presented INCOMMON, a probabilistic framework to quantify gene mutant dosage across thousands of human cancers. Unlike existing copy-number callers[16–19], which typically require access to raw sequencing data and are limited to matched tumour–normal assays, INCOMMON operates directly on variant-level read count data, thereby avoiding reliance on controlled-access sequencing files. Moreover, INCOMMON simultaneously infers allele-specific copy number and mutation multiplicity, and uniquely incorporates informative prior distributions derived from large-scale whole-genome sequencing cohorts—a key innovation that enables robust and scalable inference across diverse datasets. By jointly modelling mutation copy number and multiplicity, our approach allows the measurement of gene mutant dosage, a statistic that provides a powerful and generalizable stratification of tumour genomes. We demonstrated the high accuracy of our method through extensive cross-validation using well-established WGS data from PCAWG and HMF, confirming the robustness of INCOMMON's dosage assignments for downstream analyses. By applying this method to more than 50,000 samples from the MSK-MET and the GENIE-DFCI cohorts, we demonstrated that gene mutant dosage is a robust biomarker for patient outcome and tumour dissemination across multiple tumour types and genomic contexts.

Our analysis of over 20,000 tumour genomes from the MSK-MET cohort uncovered 46 tumour type–specific biomarkers for which dosage-derived classes were predictive of overall survival—13 of which were not detectable using conventional binary mutant/wild-type models. In most instances (70% of biomarkers), high mutant dosage was associated with significantly poorer prognosis. An exemplary case was *KRAS*, for which we confirmed and extended prior observations linking high mutant dosage to poor prognosis in lung, pancreatic and colorectal cancers[14,42–45]. We observed a similar pattern for *NRAS* in colorectal cancers. These findings may carry important clinical implications, particularly in the context of advancing RAS-targeted therapies[46]. The adverse prognostic impact of high *PIK3CA* dosage in breast cancer aligned with and extended recent findings[47,48], while its association with poor outcome in bladder cancer, along with *TERT* dosage, also emerging as a key biomarker, represents a novel observation to the best of our knowledge. The emergence of TERT dosage as a prognostic marker aligns with prior observations that even modest increases in TERT expression can drive tumorigenesis[49]. Unlike most cancer genes, where non-coding mutations rarely have strong effects, TERT stands out as an exception—its regulation is unusually sensitive, making both promoter mutations and copy-number gains potentially impactful. In colorectal cancer, high *BRAF* mutant dosage was associated with significantly poorer overall survival, confirming the prognostic impact of mutations and *BRAF* amplifications[50,51] while highlighting mutant dosage as a potentially underappreciated factor influencing both prognosis and therapeutic response. In lung cancer, high *STK11* mutant dosage was associated with significantly poorer survival. The interaction with *KRAS* mutations seemed particularly interesting and clarified earlier observations[52]: with *KRAS* mutations, even just balanced *STK11* dosage was prognostic, whereas without *KRAS* mutations, high dosage *STK11* were required to establish prognosis. Mutant dosage of *TP53* was prognostic in our univariate models but not in multivariate analysis, where its effect was largely accounted for by overall aneuploidy, and different between primary tumours and metastases. While the association between *TP53* alteration and genomic instability is well established[53,54], our results revealed that this relationship strengthens with increasing *TP53* mutant dosage, underscoring the importance of accounting for background aneuploidy when evaluating its prognostic impact.

Altogether, these results extend prior findings that the number of mutations within single-copy regions or present in multiple copies provides stronger clinical predictive power than global measures of mutational burden or aneuploidy[55]. Importantly, they highlight how mutant dosage sharpens prognostic interpretation beyond coarse burden metrics[22,38], though such metrics remain independent predictors in specific contexts[10].

From the analysis of more than 50,000 unpaired primary and metastatic samples, we found that high mutant dosage was consistently enriched in metastatic tumours, most prominently for *TP53* and *PIK3CA* in breast and colorectal cancers. This enrichment, that we validated across the two independent MSK-MET and GENIE-DFCI datasets, was echoed by elevated metastatic odds ratios in multivariable models, supporting high dosage as a marker of metastatic fitness. For instance, *TP53* high-dosage mutations in prostate and MSS colorectal cancers showed the strongest enrichment in metastases and significantly increased metastatic propensity, even when controlling for global confounders like overall mutational burden and aneuploidy. This finding was in line with studies demonstrating that loss of the remaining wild-type *TP53* allele is a key step toward full metastatic transformation[56,57]. Building on earlier reports linking high *EGFR* mutant allele dosage to advanced disease in lung adenocarcinoma, we found a clear

dosage-dependent gradient in metastatic potential. In *KRAS* wild-type tumours, low-dosage *EGFR* mutations were associated with reduced risk of metastasis, while high-dosage mutations—particularly exon 19 deletions—marked a sharp increase in dissemination. Despite this, we did not observe a corresponding impact on overall survival. While this contrasts with earlier studies[58], it aligns with more recent findings reporting no survival difference by *EGFR* dosage[59]. Our results suggest that survival outcomes in these tumours are instead primarily shaped by background aneuploidy and metastatic stage at diagnosis, which absorb much of the prognostic signal in multivariate models. The impact of *PIK3CA* mutant dosage on tumour behaviour appeared to be highly context-dependent. While high *PIK3CA* dosage was associated with both increased metastatic propensity and poorer survival in the HR+/HER2- breast cancers, in hypermutated uterine cancers, high-dosage *PIK3CA* mutations were linked to improved survival and a reduced risk of metastasis. These findings underscore the importance of tumour context, particularly mutation burden and immune environment, in shaping the phenotypic consequences of oncogene dosage. In uterine endometrioid cancer, high *ARID1A* mutant dosage was associated with increased metastatic propensity and showed a trend toward poorer overall survival, although it did not reach significance as an independent prognostic factor in multivariate analysis. This finding supports and extends prior observations that a double-hit mechanism—particularly loss of the wild-type *ARID1A* allele—may be required to drive malignant progression and promote metastasis[60–62].

Beyond global metastatic risk, mutant dosage also predicted specific patterns of metastatic spread. We identified 24 tumour type–specific biomarkers for which mutant dosage stratified organotropic tendencies of metastasis, revealing a new layer of complexity in the molecular logic of organotropism and extending previous findings based on the independent analysis of mutations and copy number alterations[22]. For instance, *KRAS* mutations that were previously linked to lung and intra-abdominal metastases in MSS colorectal cancer showed significant tropism only for balanced dosages, whereas high-dosage *KRAS* mutations, especially *G12* hotspots, were strongly associated with brain metastases. This association is further supported by recent evidence of brain metastases exhibiting a high burden of chromosomal imbalances, including amplification of *KRAS*, accompanied by significant metabolic reprogramming characterised by increased glycolysis and upregulated cell cycle pathways[63]. High dosage *TP53* and *SMAD4* mutations preferentially seeded the lymphatic system and liver, respectively. In melanoma, high-dosage *ERBB4* and *TERT* mutations predicted metastasis to the liver and brain, respectively, and high-dosage *PTPRD* to the lung. Similarly, in lung adenocarcinoma, balanced-dosage *EGFR* mutations were linked to local spread in the lung and pleura, while balanced-dosage *KEAP1* mutations favoured dissemination to the bone. This finding aligns with additional data indicating *KEAP1* mutations as significant predictors of bone metastases and poor prognosis in lung adenocarcinoma patients[64,65].

4. Conclusion

Our findings support a model of oncogenesis in which the timing and allelic configuration of mutations—key drivers of mutant dosage changes—modulate tumour evolution, metastatic trait acquisition, and organ tropism. By highlighting gene mutant dosage as a clinically relevant axis of tumour genome interpretation, we demonstrate that not all mutations are equal: their dosage, shaped by clonal selection and copy number architecture, determines biological and clinical impact. Moving beyond binary mutation classification, our approach integrates mutation type with aneuploidy to yield a more precise risk stratification metric and a mechanistic explanation for previously unresolved clinical heterogeneity. Incorporating mutant dosage into diagnostic and prognostic models could enhance patient management, especially when treatment decisions depend on mutation status.

Despite the breadth of this new analysis, several aspects warrant further development to support broader clinical translation. Reference priors drawn from large WGS cohorts may need to be continuously refined to more closely reflect the histology, treatment exposure, and sequencing depth of the clinical populations under study. Inference of mutant dosage in tumours with pronounced intratumoral heterogeneity or very low purity remains challenging, even within our Bayesian framework, and represents an opportunity for methodological advancement. In addition, variability in gene content and coverage across targeted sequencing panels currently limits the set of alterations for which dosage can be robustly inferred. The retrospective nature of the analysed cohorts also highlights the need for prospective validation of the prognostic and metastatic associations we uncovered. From a methodological perspective, moreover, extending the model to capture wildtype oncogene amplifications – though rare in primary tumours and not detectable in the present datasets – may provide additional insights into oncogenic pathway activation. The analysis of gene mutant dosage would also benefit from alternative assays, as transcriptomic and epigenomic read-outs may reveal whether high mutant dosage consistently translates into elevated oncogene expression or is

tempered by compensatory mechanisms. Dissecting how high-dosage states modulate sensitivity or resistance to targeted agents, immunotherapies, and conventional chemotherapy should help turn dosage-based stratification into a practical precision-medicine tool. Matched longitudinal primary/metastatic samples would further refine our understanding of metastatic tropism, and survival data for the GENIE-DFCI cohort would enable validation of our MSK-MET-derived inferences.

In summary, this research advances key aspects of cancer genomics and presents INCOMMON as a transformative tool for interpreting targeted sequencing in the clinic. Our findings lay a strong foundation for future research and clinical translation, contributing both to personalised cancer therapy and to a deeper understanding of the biological complexity underlying genomic alterations. The clinical relevance of these insights underscores their potential to improve patient care.

Data and software availability

All analyses presented in this paper, from data gathering to analysis results, are available in Zenodo at

https://zenodo.org/uploads/15498338

PCAWG and TCGA data used in this article (mutations with validated matched CNAs) are available from Zenodo at

https://doi.org/10.5281/zenodo.6410935.

The MSK MetTropism data used in this article have been downloaded from the CBioPortal at

https://www.cbioportal.org,

under cohort `"msk_met_2021"`. AACR GENIE-DFCI data has been downloaded from Synapse at

https://www.synapse.org/

through access codes `syn50678641`, `syn50678411`, `syn50678410`, `syn50678644`, `syn50678531`, `syn50678642`, `syn50678530`, `syn50678532`, `syn50678295`, `syn50678640`, `syn50678653`, `syn50678296`.

HMF data (segmentation files and purity/ploidy estimates for $n = 4{,}496$ samples) were obtained from the Hartwig Medical Foundation (version 17/10/2020), using the request forms that can be found at

www.hartwigmedicalfoundation.nl

The files were generated using the PURPLE pipeline, available at

github.com/hartwigmedical/hmftools/

We then manually converted the Hartwig annotations to match those in ICGC (Supplementary Table S9).

INCOMMON is an open-source R package downloadable from Github at

https://caravagnalab.github.io/INCOMMON

The tool webpage contains RMarkdown vignettes to run analyses, visualise inputs and outputs, and parametrise the tool.

INCOMMON is also available as a ShinyApp that can be used to browse analysis results and run online similar analyses. The app is available online at

https://ncalonaci.shinyapps.io/incommon/

Author contributions

NC and GC conceptualised and formalised INCOMMON, which was implemented by NC with support from DC, SM and SS. NC, BR, SS, MMS and KB gathered real data, which was analysed by NC, EK, DC, GG, and GC, and interpreted by all authors. GC supervised the project. NC, EK and GC drafted the manuscript that all authors approved in final form.

Funding and acknowledgments

The research leading to these results has received funding from AIRC under MFAG 2020 - ID. 24913 project – P.I. Caravagna Giulio. This research was also financially supported through funding from the IRCCS Regina Elena National Cancer Institute's "Ricerca Corrente" granted by the Italian Ministry of Health. KB was funded by the Swiss National Science Foundation (P500PM_217647/1). We thank Area Science Park for computational support through the ORFEO (Open Research Facility for Epigenomics and Other) data centre. This publication and the underlying study have been made possible partly based on data that Hartwig Medical Foundation has made available to the study through the Hartwig Medical Database. We thank the reviewers and editor for providing constructive feedback to improve our manuscript.

Inclusion and ethics

All collaborators of this study who have fulfilled the criteria for authorship required by Nature Portfolio journals have been included as authors, as their participation was essential for the design and implementation of the study. Roles and responsibilities were agreed among collaborators ahead of the research.

Online Methods

Utilised datasets

To compute the empirical prior distributions of copy number configurations (k, m), we assembled a cohort of high resolution whole-genome sequencing data, collecting samples from the Pan-Cancer Analysis of Whole Genomes[24] dataset (PCAWG, $N = 2777$ samples, 40 tumour types, median depth 45 and purity 65%), and from the Hartwig Medical Foundation[26] (HMF, $N = 8616$ samples, 28 tumour types, median depth 96 and purity 55%) dataset. We retrospectively assembled CNA calls from PCAWG, which were previously obtained as part of a curated consensus[66], and CNA calls from HMF that were previously obtained using PURPLE[25]. We used the resulting dataset of mutations and CNAs as the benchmark dataset $\mathcal{B}$. Importantly, this aggregated dataset includes samples from both primary (from PCAWG) and metastatic tumours (from HMF), ensuring its usefulness in the analysis of datasets containing both sample types, such as MSK-MET. Additionally, we used data from The Cancer Genome Atlas[27] (TCGA, 1068 samples) as a validation dataset.

Bayesian inference of copy number and mutation multiplicity using INCOMMON

INCOMMON is a Bayesian framework to infer mutation multiplicity and copy number from read count data. The data consist of a set of n mutations $X = \{x_1, \dots, x_n\}$ from a tumour sample, with each mutation $x_i = \langle r_i, d_i \rangle$ characterised by the number of reads with the variant $r_i \in \mathbb{Z}^+$ and the total number of reads $d_i \in \mathbb{Z}^+$ (sequencing depth). INCOMMON jointly models the purity $0 < \pi \leq 1$ and the rate of reads per chromosome copy ($\eta \in \mathbb{R}^+$) of the sample, and the total copy number ($k_i \in \mathbb{Z}^+$) and mutation multiplicty ($m_i \in \mathbb{Z}^+$, $m_i \leq k_i$) of each mutation in the sample. The joint likelihood of the model is

$$p(X \mid \Theta) = \prod_{x_i \in X} p(x_i \mid \Theta) = \prod_{x_i \in X} p(d_i \mid \Theta)\, p(r_i \mid d_i, \Theta)\,. \tag{1}$$

where Θ is the vector of model parameters $\Theta = (\pi, \eta, \{k, m\}_{i=1}^n)$.

In INCOMMON, the total number of reads follows a Poisson distribution

$$p(d_i \mid \pi, \eta, k_i) = \text{Poisson}(d_i \mid \lambda) \qquad \lambda = (1-\pi)2\eta + \pi\eta k_i \tag{2}$$

The expected value λ is contributed by reads from the normal cells, present in the sample in a fraction $1 - \pi$, and from tumour cells present in a fraction π. The expected number of reads from normal cells, assumed to be diploid, is 2η across the whole genome. The expected number of reads from tumour cells is $k\eta$, assumed to be proportional to the unknown total copy number.

Given the total read count d_i, the number of reads with the variant r_i follows a Binomial distribution

$$p(r_i \mid \pi, k_i, m_i) = \text{Binomial}(r_i \mid d_i, \varphi) \qquad \varphi = \frac{m\pi}{2(1-\pi) + k\pi} \tag{3}$$

The probability of collecting a read with the variant (success probability) φ is equal to the variant allele frequency (VAF) expected for a mutation with copy number configuration (k, m), adjusted for the sample purity. INCOMMON embeds prior knowledge of the model parameters Θ using a prior distribution that can be factorised into

$$p(\Theta) = p(\pi)p(\eta)p(k, m) \tag{4}$$

INCOMMON uses Markov Chain Monte Carlo (MCMC) sampling to estimate the model posterior $p(\Theta \mid X)$, that would otherwise be intractable analytically.

Empirical Priors

Copy number and mutation multiplicity

A key feature of INCOMMON is the use of biologically informed prior distributions for copy number and mutation multiplicity configurations across genes and tumour types. These priors were derived from a training dataset of 4543 samples (95896 mutations with available copy number states) curated from two large high-resolution whole-genome sequencing cohorts: the Pan-Cancer Analysis of Whole Genomes (PCAWG, 1519 samples) and the Hartwig Medical Foundation (HMF, 3024 samples). A subset of samples was held out for model performance evaluation (see **??**).

We empirically estimated gene- and tumour type-specific prior distributions $p(k, m)$ over 36 possible combinations of total copy number and mutation multiplicity, with an upper bound of $k_{max} = 8$. The prior probabilities were computed based on the observed frequencies of each configuration in the training dataset for each tumour type. To ensure statistical robustness, we required that a gene be mutated in at least 50 samples within a given tumour type to construct a tumour-specific prior; otherwise, a pan-cancer prior was used by aggregating data across all tumour types.

Missing observations were imputed using pseudocounts, estimated through a Gaussian-kernel smoothing approach that incorporated copy number and mutation multiplicity distances. In particular, given observed mutation counts n_i for different (k, m) configurations, we defined a custom distance metric $d_{ij} = |k_i - k_j| + \left|\frac{m_i}{k_i} - \frac{m_j}{k_j}\right|$ and applied a Gaussian kernel $K_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)$ to weight the observations. We then computed the pseudocounts as $\tilde{n}_i = \sum_j K_{ij} n_j + 1$, where we used 1 as the base pseudocount, ensuring numerical stability. To preserve the overall mutation burden, we applied a normalisation step to get the final pseudo-count $\tilde{n}_i \leftarrow \tilde{n}_i \frac{10+n_i}{\sum_j n_j}$.

Tumour purity

The inference of copy number and multiplicity with INCOMMON critically relies on the sample purity, a piece of information that is usually accessible as provided together with genomic data. However, the available purity estimates are often pathology-based and, thus, potentially incorrect or differing from molecular estimates. The full Bayesian formulation of INCOMMON allows using an informative, yet broad, prior on sample purity π, defined starting from the available estimate π_0. INCOMMON uses a Beta distribution over purity

$$p(\pi) = \text{Beta}(\alpha_\pi, \beta_\pi) \tag{5}$$

Here, for each sample, we choose specific values of the shape parameters α_π such that the mean of the distribution is equal to the available estimate π_0, and the expected variance is fixed to σ_π^2, namely:

$$\alpha_\pi = \pi_0 \left[\frac{\pi_0 (1 - \pi_0)}{\sigma_\pi^2} - 1\right] \tag{6}$$

$$\beta_\pi = (1 - \pi_0) \left[\frac{\pi_0 (1 - \pi_0)}{\sigma_\pi^2} - 1\right] \tag{7}$$

We used $\sigma_\pi^2 = 0.05$ for all the datasets we analysed. In this way, INCOMMON accounts for uncertainty in purity estimation of around $\simeq 10\%$. Potentially, there can be cases in which the input purity estimate is not supported by the inference of INCOMMON, despite the allowed uncertainty. We control for these extreme cases through posterior predictive checks (see, for an example, Extended Data Figure 1).

Read count rate per chromosome copy

For the read count rate per chromosome copy η, we estimate the prior distribution from the data in an Empirical Bayes fashion. In principle, following Equation 2 the expected total number of reads from a mutation site with copy number k' is $\mathbb{E}(d_i \mid k = k') = (1 - \pi)2\eta + \pi\eta k'$. Thus, an empirical estimator of $\tilde{\eta}$ of the per-copy rate can be obtained from the mean total number of reads coming from all the mutations with $k = k'$ as $\tilde{\eta} = \frac{<d>_{k_i = k'}}{2(1-\pi)+k'\pi}$. Since a priori we do not know the copy number of mutations in the sample, the best we can do is to use the expected value with respect to the prior probability $p(k) = \sum_{m=1}^{k} p(k, m)$. In this way we obtain the estimator

$$\tilde{\eta} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k'} \frac{d_i p(k')}{2(1 - \pi) + k'\pi} \,. \tag{8}$$

We then use a Gamma prior distribution over η

$$p(\eta) = \text{Gamma}(\alpha_\eta, \beta_\eta) \tag{9}$$

in which we choose the values of the shape parameters such that the mean is equal to our empirical estimation $\tilde{\eta}$

and the variance is equal to the empirical variance $Var(\tilde{\eta})$ across the dataset

$$\alpha_\eta = \frac{\tilde{\eta}^2}{Var(\tilde{\eta})} \tag{10}$$

$$\beta_\eta = \frac{\tilde{\eta}}{Var(\tilde{\eta})}\,. \tag{11}$$

This prior reflects our prior knowledge over latent copy number and multiplicity configurations $p(k, m)$ and adapts it to the observed distribution of sequencing depth in the dataset, constraining posterior estimates of η to reasonable values. The prior distribution over η for all the analysed cohorts (MSK-MET, DFCI, MSK-CHORD) are reported in Supplementary Figure S4, along with the corresponding observed distributions of sequencing depth. The use of an informative prior for the estimation of the per-copy count rate is crucial, especially in settings where normal samples are not available (due to questions of data disclosure or in tumour-only assays) so that depth ratios cannot be computed.

The fraction of alleles with the mutation (FAM)

INCOMMON allows for the estimation of allele-specific copy-number configurations of mutations. In our study, utilising INCOMMON, we estimated the Fraction of Alleles carrying the Mutation (FAM), representing the relative fraction of mutant alleles of a gene within all cells of the sample. The probabilistic model enables the calculation of the expected value of FAM as:

$$\mathbb{E}(\text{FAM}) = \sum_{k=1}^{k_{max}} \sum_{m=1}^{k} \frac{m}{k} p(m, k \mid X) \tag{12}$$

where $p(m, k \mid X)$ denotes the joint posterior probability of k and m, obtained by marginalizing the full posterior over sample purity π and read count rate per chromosome copy η, $p(m, k \mid X) = \int_0^1 d\pi \int_{\mathbb{R}^+} d\eta\; p(\Theta \mid X)$. Even in cases where the exact configuration of (k, m) cannot be precisely determined, the estimator in Equation (12) leverages the full posterior distribution to provide a robust estimate of their ratio m/k, weighting contributions from different configurations according to their joint probability.

Classification based on FAM cutoffs

To determine the optimal thresholds for dichotomising FAM based on overall survival (OS), we used the maximally selected rank statistics method. We searched for the lower cutoff in the range $FAM < 0.5$ and for the upper one in the range $FAM > 0.5$. The maximally selected rank statistics method identifies the cutpoint that yields the most significant split in survival, while correcting for multiple testing. We required at least 20% of all samples in each group, and adjusted models for age, sex, tumour mutational burden, fraction of genome altered, and sample type (primary tumours vs metastases). Analyses were conducted using R version 4.3.2 with the tidyverse (v2.0.0) and survminer (v0.4.9) packages.

Posterior Predictive Checking and Bayesian p-value Calculation

To assess model fit and evaluate potential discrepancies between observed data and posterior predictive distributions, we performed posterior predictive checks and computed Bayesian p-values.

Posterior Predictive p-value

For each mutation-specific observed quantity y_i, such as reads with the variant r_i or total reads d_i, INCOMMON generates S posterior predictive samples $\{\hat{y}_i^s\}_{s=1}^S$ from the model posterior distribution, and evaluates the posterior predictive p-value as

$$p_i = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left(|\hat{y}_i^s - \mathbb{E}[\hat{y}_i]| \geq |y_i - \mathbb{E}[\hat{y}_i]|\right) \tag{13}$$

where $\mathbb{E}[\hat{y}_i^s]$ is the posterior mean of the replicated observations. The indicator function $\mathbb{I}(\cdot)$ evaluates to 1 if the absolute deviation of the posterior predictive sample from the posterior mean is at least as large as the deviation of the observed data and 0 otherwise. This p-value thus quantifies the extremity of the observed data under the model posterior, providing a measure of the goodness of fit for individual mutations.

Bayesian p-value

At the level of global (sample-wise) parameters y, such as tumour purity π and read counts per chromosome copy η, INCOMMON estimates Bayesian p-values by comparing posterior and prior distributions. Given S posterior samples $\{y^{\text{post}}\}_{s=1}^{S}$ and prior samples $\{y^{\text{prior}}\}_{s=1}^{S}$, the test statistic is based on deviations from the prior mean $\mathbb{E}[y^{\text{prior}}]$. INCOMMON computes the test statistic differently depending on the assumed prior distribution:

- For a Beta prior:
$$T_s^{\text{prior}} = \left|y_s^{\text{prior}} - \mathbb{E}[y^{\text{prior}}]\right|, \quad T_s^{\text{post}} = \left|y_s^{\text{post}} - \mathbb{E}[y^{\text{prior}}]\right| \tag{14}$$
- For a Gamma prior:
$$T^{\text{prior}} = \frac{\left|y^{\text{prior}} - \mathbb{E}[y^{\text{prior}}]\right|}{\mathbb{E}[y^{\text{prior}}]}, \quad T^{\text{post}} = \frac{\left|y^{\text{post}} - \mathbb{E}[y^{\text{prior}}]\right|}{\mathbb{E}[y^{\text{prior}}]} \tag{15}$$

The Bayesian p-value is thus computed as the proportion of posterior test statistics that exceed or equal the prior test statistics:

$$p = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left(T_s^{\text{post}} \geq T_s^{\text{prior}}\right). \tag{16}$$

This Bayesian p-value quantifies the extent to which posterior samples deviate from prior expectations, providing an additional measure of model adequacy.

Evaluation of model performance

We developed an iterative approach to evaluate the performance of INCOMMON and the impact of empirical priors on the model predictions. Our method involved iteratively splitting the benchmark dataset $\mathcal{B}$ into two subsets: the prior dataset $\mathcal{P}$ (70% of the samples) for prior estimation, and the evaluation dataset $\mathcal{E}$ (30% of the samples) for performance evaluation. By aggregating results over $N = 50$ iterations, we computed the mean absolute error (MAE) of INCOMMON predictions of mutation total copy number and multiplicity along with their standard deviations.

Setting an upper bound for total copy number k

In INCOMMON, the model likelihood is computed over a discrete set of combinations of total copy number k and multiplicity m. While for multiplicity, an upper bound is set naturally by the constraint that $m \leq k$, i.e. the number of mutant copies cannot exceed the total number of copies, for the total copy number, the choice of the upper bound k_{max} is arbitrary. In order to choose a reasonable value of k_{max}, we used the benchmark dataset $\mathcal{B}$ to evaluate the frequency of large (> 4) values of k and set a threshold above which the number of mutations is negligible. As shown in Supplementary Figure S12, the distribution of total copy number values significantly varied between wild-type (WT) and mutant sites, with the number of sites decreasing more rapidly with k for the mutant sites with respect to WT. In particular, we found fewer than $n = 100$ mutations per value of k for $k > 8$ across the whole benchmark dataset ($N = 2375815$ mutations in total). We thus set the upper bound to $k_{max} = 8$ for all our analyses.

Survival Analysis

Overall survival (OS) time (months) and status (living vs dead) for the MSK-MET were taken from the publicly available clinical release. OS curves and OS values were calculated using the univariate Kaplan–Meier estimator. All p-values and adjusted hazard ratios were evaluated through multivariable Cox proportional hazards models accounting for covariates including age at sequencing, sex, tumour mutational burden per megabase (TMB) and fragment of genome altered (FGA). Multivariable Cox proportional hazards models were constructed for each gene, and p-values of mutant dosage classes (WT, MUT, Low, Balanced, High) and all covariates were adjusted by FDR independently for each tumour type using the number of genes that were analysed simultaneously. For each tumour type, all genes that were mutated in less than 150 samples were excluded from the analysis. The same threshold was used for specific recurrent amino-acid substitutions (hotspot mutations). Analyses were conducted using R version 4.3.2 with the tidyverse (v2.0.0), survival (v3.7-0), survminer (v0.4.9) and multcomp (v1.4-26).

Enrichment in primary tumours vs metastases

Sample type (primary tumour/metastasis) annotations for the MSK-MET were taken from the publicly available clinical release. To evaluate whether any of the three gene mutant dosage classes were enriched in either primary

tumours or metastases, we performed an exact Fisher's test. For each gene and tumour type, we constructed a two-way contingency table of sample counts between class and sample type, and applied Fisher's test to assess deviation from independence under the null hypothesis. We evaluated an enrichment score from the standardised residuals obtained from the contingency table, which measures the deviation of observed from expected counts, normalised by the standard error. P-values for all the tests were FDR-adjusted through the Benjamini-Hochberg method, independently for each tumour type, using the number of genes that were analysed simultaneously. Analyses were conducted using R version 4.3.2 with the tidyverse (v2.0.0) and stats (v4.3.3).

Metastatic propensity and organotropism

Patient status (metastatic/non-metastatic) and site of metastasis annotations for the MSK-MET were taken from the publicly available clinical release. To assess associations between gene mutant dosage class and sample type (metastatic propensity), and between gene mutant dosage and patient status (organotropism), we fitted univariate and multivariate logistic regression models. The univariate model included only gene mutant dosage as a predictor. The multivariate model included covariates for age at sequencing, sex, tumour mutational burden per megabase (TMB) and fragment of genome altered (FGA). All models were specified as generalised linear models using the binomial family with a logit link function. From the fits, we extracted coefficient estimates for all variables and 95% confidence intervals. A two-sided Wald test for statistical significance was performed for each regression coefficient. P-values for all the covariates were FDR-adjusted through the Benjamini-Hochberg method, independently for each tumour type, using the number of genes that were analysed simultaneously. Wald pairwise comparisons between mutant dosage classes were performed using general linear hypothesis testing. Analyses were conducted using R version 4.3.2 with the tidyverse (v2.0.0), stats (v4.3.3) and multcomp (v1.4-26).

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. en. Nature **481,** 306–313. https://www.nature.com/articles/nature10762 (2012).
2. Watson, E. V. et al. Chromosome evolution screens recapitulate tissue-specific tumor aneuploidy patterns. en. Nat. Genet., 1–13. https://www.nature.com/articles/s41588-024-01665-2 (2024).
3. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. en. Nat. Rev. Genet. **20,** 404–416. http://dx.doi.org/10.1038/s41576-019-0114-6 (2019).
4. Sansregret, L. & Swanton, C. The role of aneuploidy in cancer evolution. en. Cold Spring Harb. Perspect. Med. **7,** a028373. https://pubmed.ncbi.nlm.nih.gov/28049655/ (2017).
5. Lukow, D. A. & Sheltzer, J. M. Chromosomal instability and aneuploidy as causes of cancer drug resistance. en. Trends Cancer **8,** 43–53. https://pubmed.ncbi.nlm.nih.gov/34593353/ (2022).
6. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. en. Nature **587,** 126–132. https://www.nature.com/articles/s41586-020-2698-6 (2020).
7. Henry, N. L. & Hayes, D. F. Cancer biomarkers. en. Mol. Oncol. **6,** 140–146. https://pubmed.ncbi.nlm.nih.gov/22356776/ (2012).
8. Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S. & Swisher, E. Cancer biomarkers: a systems approach. en. Nat. Biotechnol. **24,** 905–908. https://pubmed.ncbi.nlm.nih.gov/16900126/ (2006).
9. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. en. Nat. Rev. Cancer **20,** 555–572. https://pubmed.ncbi.nlm.nih.gov/32778778/ (2020).
10. Spurr, L. F., Weichselbaum, R. R. & Pitroda, S. P. Tumor aneuploidy predicts survival following immunotherapy across multiple cancers. en. Nat. Genet. **54,** 1782–1785. https://www.nature.com/articles/s41588-022-01235-4 (2022).
11. Nass, Sharyl J and Moses, Harold L. The promises and challenges of improving detection and treatment. Committee on Developing Biomarker-Based Tools for Cancer Screening, Diagnosis, and Treatment. https://www.academia.edu/download/47366363/Cancer_Biomarkers-The_Promises_and_Challenges_of_Improving_Detection_and_Treatment.pdf (2007).
12. Mulero-Navarro, S. & Esteller, M. Epigenetic biomarkers for human cancer: the time is now. en. Crit. Rev. Oncol. Hemato **68,** 1–11. https://pubmed.ncbi.nlm.nih.gov/18430583/ (2008).

13. Costa-Pinheiro, P., Montezuma, D., Henrique, R. & Jerónimo, C. Diagnostic and prognostic epigenetic biomarkers in cancer. en. Epigenomics **7,** 1003–1015. https://pubmed.ncbi.nlm.nih.gov/26479312/ (2015).
14. Varghese, A. M. et al. Clinicogenomic landscape of pancreatic adenocarcinoma identifies KRAS mutant dosage as prognostic of overall survival. en. Nat. Med. **31,** 466–477. https://www.nature.com/articles/s41591-024-03362-3 (2025).
15. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. en. Nat. Rev. Cancer **21,** 379–392. https://www.nature.com/articles/s41568-021-00336-2 (2021).
16. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. en. Nat. Biotechnol. **30,** 413–421. http://dx.doi.org/10.1038/nbt.2203 (2012).
17. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. en. Nucleic Acids Res. **44,** e131. http://dx.doi.org/10.1093/nar/gkw520 (2016).
18. Riester, M. et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. en. Source Code Biol. Med. **11,** 13. https://scfbm.biomedcentral.com/articles/10.1186/s13029-016-0060-z (2016).
19. Kuilman, T. et al. CopywriteR: DNA copy number detection from off-target sequence data. en. Genome Biol. **16,** 49. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0617-1 (2015).
20. Cheng, D. T. et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. en. J. Mol. Diagn. **17,** 251–264. https://pubmed.ncbi.nlm.nih.gov/25801821/ (2015).
21. Ptashkin, R. N. et al. Abstract 3409: MSK-IMPACT Heme: Validation and clinical experience of a comprehensive genome in Molecular and Cellular Biology / Genetics **79** (American Association for Cancer Research, 2019), 3409–3409. https://aacrjournals.org/cancerres/article/79/13_Supplement/3409/542857/Abstract-3409-MSK-IMPACT-Heme-Validation-and.
22. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. en. Cell **185,** 563–575.e11. http://dx.doi.org/10.1016/j.cell.2022.01.003 (2022).
23. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. en. Cancer Discov. **7,** 818–831. http://dx.doi.org/10.1158/2159-8290.CD-17-0151 (2017).
24. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. en. Nature **578,** 82–93. http://dx.doi.org/10.1038/s41586-020-1969-6 (2020).
25. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. en. Nature **575,** 210–216. https://pubmed.ncbi.nlm.nih.gov/31645765/ (2019).
26. Martínez-Jiménez, F. et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. en. Nature, 1–9. https://www.nature.com/articles/s41586-023-06054-z (2023).
27. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. en. Nat. Genet. **45,** 1113–1120. http://dx.doi.org/10.1038/ng.2764 (2013).
28. Campbell, K. R. et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. en. Genome Biol. **20,** 54. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1645-z (2019).
29. Milite, S., Bergamin, R., Patruno, L., Calonaci, N. & Caravagna, G. A Bayesian method to cluster single-cell RNA sequencing data using copy number alterations. en. Bioinformatics **38,** 2512–2518. http://dx.doi.org/10.1093/bioinformatics/btac143 (2022).
30. Patruno, L. et al. A Bayesian method to infer copy number clones from single-cell RNA and ATAC sequencing. en. PLoS Comput. Biol. **19,** e1011557. http://dx.doi.org/10.1371/journal.pcbi.1011557 (2023).
31. Bielski, C. M. et al. Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. en. Cancer Cell **34,** 852–862.e4. http://dx.doi.org/10.1016/j.ccell.2018.10.003 (2018).
32. Park, S., Supek, F. & Lehner, B. Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers. en. Nat. Commun. **12,** 7051. http://dx.doi.org/10.1038/s41467-021-27242-3 (2021).

33. Ciani, Y. et al. Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. en. Cell Syst. **13,** 183–193.e7. http://dx.doi.org/10.1016/j.cels.2021.10.001 (2022).

34. Yu, C.-C. et al. Mutant allele specific imbalance in oncogenes with copy number alterations: Occurrence, mechanisms, and potential clinical implications. en. Cancer Lett. **384,** 86–93. http://dx.doi.org/10.1016/j.canlet.2016.10.013 (2017).

35. Bielski, C. M. & Taylor, B. S. Mutant allele imbalance in cancer. en. Annu. Rev. Cancer Biol. **5,** 221–234. http://dx.doi.org/10.1146/annurev-cancerbio-051320-124252 (2021).

36. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor mutational burden and response rate to PD-1 inhibition. en. N. Engl. J. Med. **377,** 2500–2501. http://dx.doi.org/10.1056/nejmc1713444 (2017).

37. Strickler, J. H., Hanks, B. A. & Khasraw, M. Tumor mutational burden as a predictor of immunotherapy response: Is more always better? en. Clin. Cancer Res. **27,** 1236–1241. http://dx.doi.org/10.1158/1078-0432.CCR-20-3054 (2021).

38. Aggarwal, C. et al. Assessment of tumor mutational burden and outcomes in patients with diverse advanced cancers treated with immunotherapy. en. JAMA Netw. Open **6,** e2311181. http://dx.doi.org/10.1001/jamanetworkopen.2023.11181 (2023).

39. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. en. N. Engl. J. Med. **376,** 2109–2121. http://dx.doi.org/10.1056/nejmoa1616288 (2017).

40. Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. en. Nature **616,** 525–533. http://dx.doi.org/10.1038/s41586-023-05783-5 (2023).

41. Spurr, L. F. & Pitroda, S. P. Exploiting tumor aneuploidy as a biomarker and therapeutic target in patients treated with immune checkpoint blockade. en. NPJ Precis Oncol **8,** 1. http://dx.doi.org/10.1038/s41698-023-00492-8 (2024).

42. Modrek, B. et al. Oncogenic activating mutations are associated with local copy gain. en. Mol. Cancer Res. **7,** 1244–1252. http://dx.doi.org/10.1158/1541-7786.MCR-08-0532 (2009).

43. Chiosea, S. I., Sherer, C. K., Jelic, T. & Dacic, S. KRAS mutant allele-specific imbalance in lung adenocarcinoma. en. Mod. Pathol. **24,** 1571–1577. http://dx.doi.org/10.1038/modpathol.2011.109 (2011).

44. Hartman, D. J., Davison, J. M., Foxwell, T. J., Nikiforova, M. N. & Chiosea, S. I. Mutant allele-specific imbalance modulates prognostic impact of KRAS mutations in colorectal adenocarcinoma and is associated with worse overall survival. en. Int. J. Cancer **131,** 1810–1817. http://dx.doi.org/10.1002/ijc.27461 (2012).

45. Fung, A. S. et al. Prognostic and predictive effect of KRAS gene copy number and mutation status in early stage non-small cell lung cancer patients. en. Transl. Lung Cancer Res. **10,** 826–838. https://pmc.ncbi.nlm.nih.gov/articles/PMC7947394/ (2021).

46. Punekar, S. R., Velcheti, V., Neel, B. G. & Wong, K.-K. The current state of the art and future trends in RAS-targeted cancer therapies. en. Nat. Rev. Clin. Oncol. **19,** 637–655. http://dx.doi.org/10.1038/s41571-022-00671-9 (2022).

47. Migliaccio, I. et al. PIK3CA co-occurring mutations and copy-number gain in hormone receptor positive and HER2 negative breast cancer. en. NPJ Breast Cancer **8,** 24. http://dx.doi.org/10.1038/s41523-022-00382-5 (2022).

48. Correia, L. et al. Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. en. NPJ Breast Cancer **8,** 71. https://www.nature.com/articles/s41523-022-00435-9 (2022).

49. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. en. Nature **578,** 102–111. http://dx.doi.org/10.1038/s41586-020-1965-x (2020).

50. Sanz-Garcia, E., Argiles, G., Elez, E. & Tabernero, J. BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. en. Ann. Oncol. **28,** 2648–2657. http://dx.doi.org/10.1093/annonc/mdx401 (2017).

51. Corcoran, R. B. et al. Combined BRAF, EGFR, and MEK inhibition in patients with BRAFV600E-mutant colorectal cancer. en. Cancer Discov. **8,** 428–443. http://dx.doi.org/10.1158/2159-8290.CD-17-1226 (2018).

52. Ricciuti, B. et al. Diminished efficacy of programmed death-(ligand)1 inhibition in STK11- and KEAP1-mutant lung adenocarcinoma is affected by KRAS mutation status. en. J. Thorac. Oncol. **17,** 399–410. http://dx.doi.org/10.1016/j.jtho.2021.10.013 (2022).

53. Hanel, W. & Moll, U. M. Links between mutant p53 and genomic instability. en. J. Cell. Biochem. **113,** 433–439. http://dx.doi.org/10.1002/jcb.23400 (2012).

54. Wang, H., Guo, M., Wei, H. & Chen, Y. Targeting p53 pathways: mechanisms, structures, and advances in therapy. en. Signal Transduct. Target. Ther. **8,** 92. http://dx.doi.org/10.1038/s41392-023-01347-1 (2023).

55. Niknafs, N. et al. Persistent mutation burden drives sustained anti-tumor immune responses. en. Nat. Med. **29,** 440–449. http://dx.doi.org/10.1038/s41591-022-02163-w (2023).

56. Nakayama, M. et al. Loss of wild-type p53 promotes mutant p53-driven metastasis through acquisition of survival and tumor-initiating properties. en. Nat. Commun. **11,** 2333. http://dx.doi.org/10.1038/s41467-020-16245-1 (2020).

57. Powell, E., Piwnica-Worms, D. & Piwnica-Worms, H. Contribution of p53 to metastasis. en. Cancer Discov. **4,** 405–414. http://dx.doi.org/10.1158/2159-8290.CD-13-0136 (2014).

58. Oakley 3rd, G. J. & Chiosea, S. I. Higher dosage of the epidermal growth factor receptor mutant allele in lung adenocarcinoma correlates with younger age, stage IV at presentation, and poorer survival. en. J. Thorac. Oncol. **6,** 1407–1412. http://dx.doi.org/10.1097/JTO.0b013e31821d41af (2011).

59. Malapelle, U. et al. EGFR mutant allelic-specific imbalance assessment in routine samples of non-small cell lung cancer. en. J. Clin. Pathol. **68,** 739–741. http://dx.doi.org/10.1136/jclinpath-2015-203101 (2015).

60. Wilson, M. R. et al. ARID1A and PI3-kinase pathway mutations in the endometrium drive epithelial transdifferentiation and collective invasion. en. Nat. Commun. **10,** 3554. http://dx.doi.org/10.1038/s41467-019-11403-6 (2019).

61. Yachida, N. et al. ARID1A protein expression is retained in ovarian endometriosis with ARID1A loss-of-function mutations: implication for the two-hit hypothesis. en. Sci. Rep. **10,** 14260. http://dx.doi.org/10.1038/s41598-020-71273-7 (2020).

62. Suryo Rahmanto, Y. et al. Inactivation of Arid1a in the endometrium is associated with endometrioid tumorigenesis through transcriptional reprogramming. en. Nat. Commun. **11,** 2717. http://dx.doi.org/10.1038/s41467-020-16416-0 (2020).

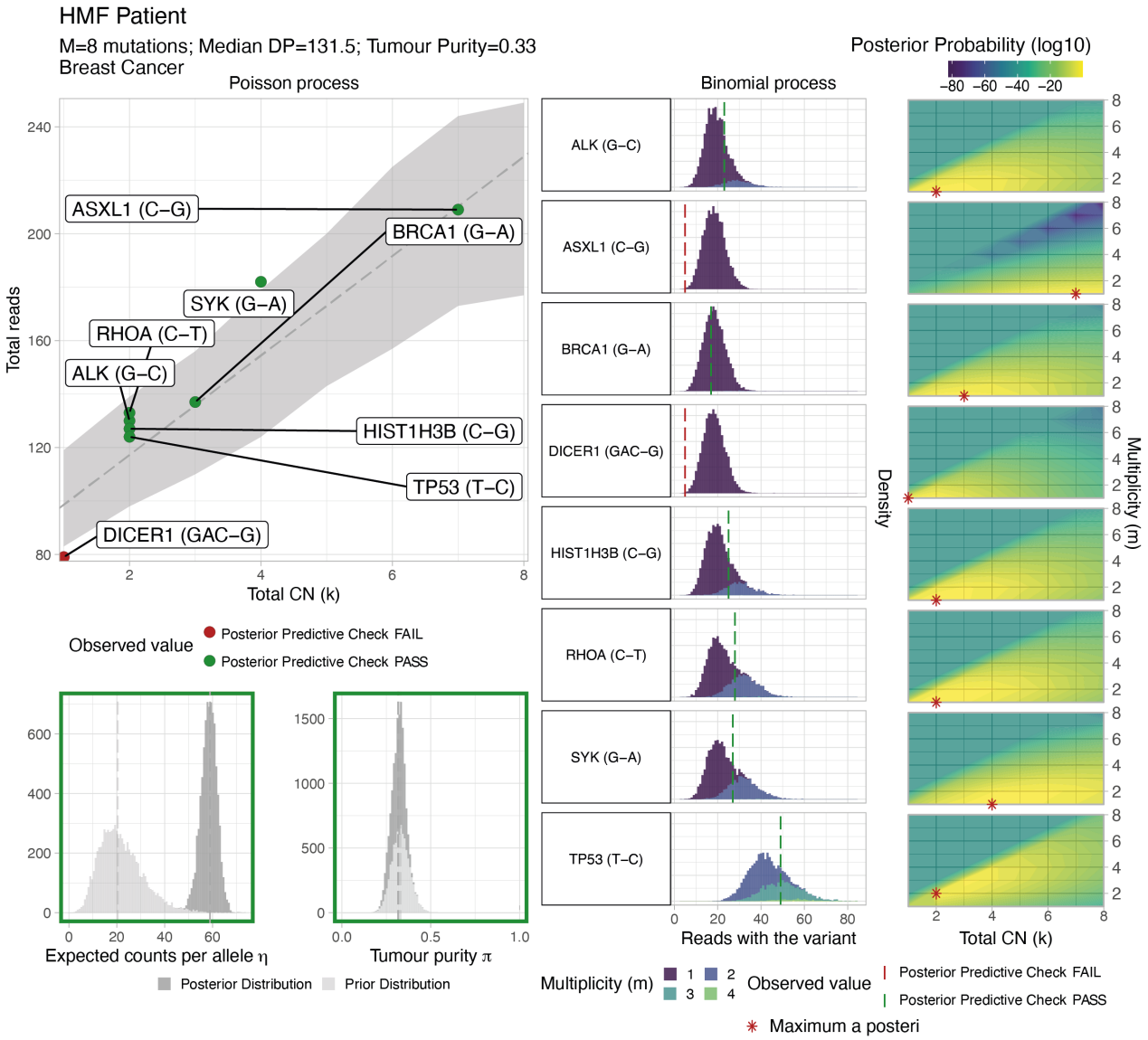
63. Golas, M. M. et al. Cytogenetic signatures favoring metastatic organotropism in colorectal cancer. en. Nat. Commun. **16,** 3261. http://dx.doi.org/10.1038/s41467-025-58413-1 (2025).

64. Sharie, A. H. A. et al. Lung adenocarcinoma with bone metastases: Clinicogenomic profiling and insights into prognostic factors. en. Cancer Control **32,** 10732748251325587. http://dx.doi.org/10.1177/10732748251325587 (2025).

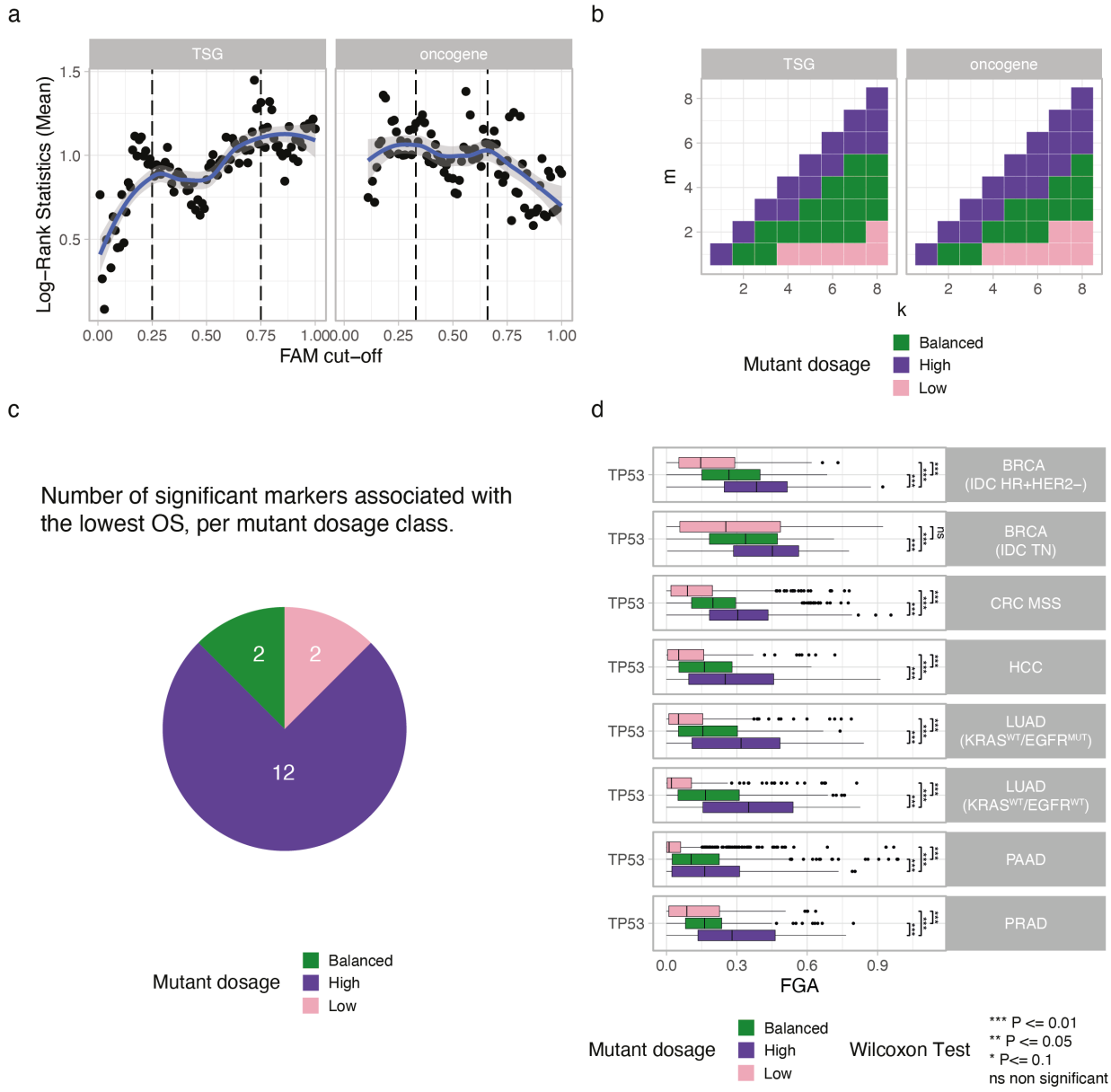
65. Scalera, S. et al. Clonal KEAP1 mutations with loss of heterozygosity share reduced immunotherapy efficacy and low immune cell infiltration in lung adenocarcinoma. en. Ann. Oncol. **34,** 275–288. http://dx.doi.org/10.1016/j.annonc.2022.12.002 (2023).

66. Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. en. Cell **184,** 2239–2254.e39. https://pubmed.ncbi.nlm.nih.gov/33831375/ (2021).

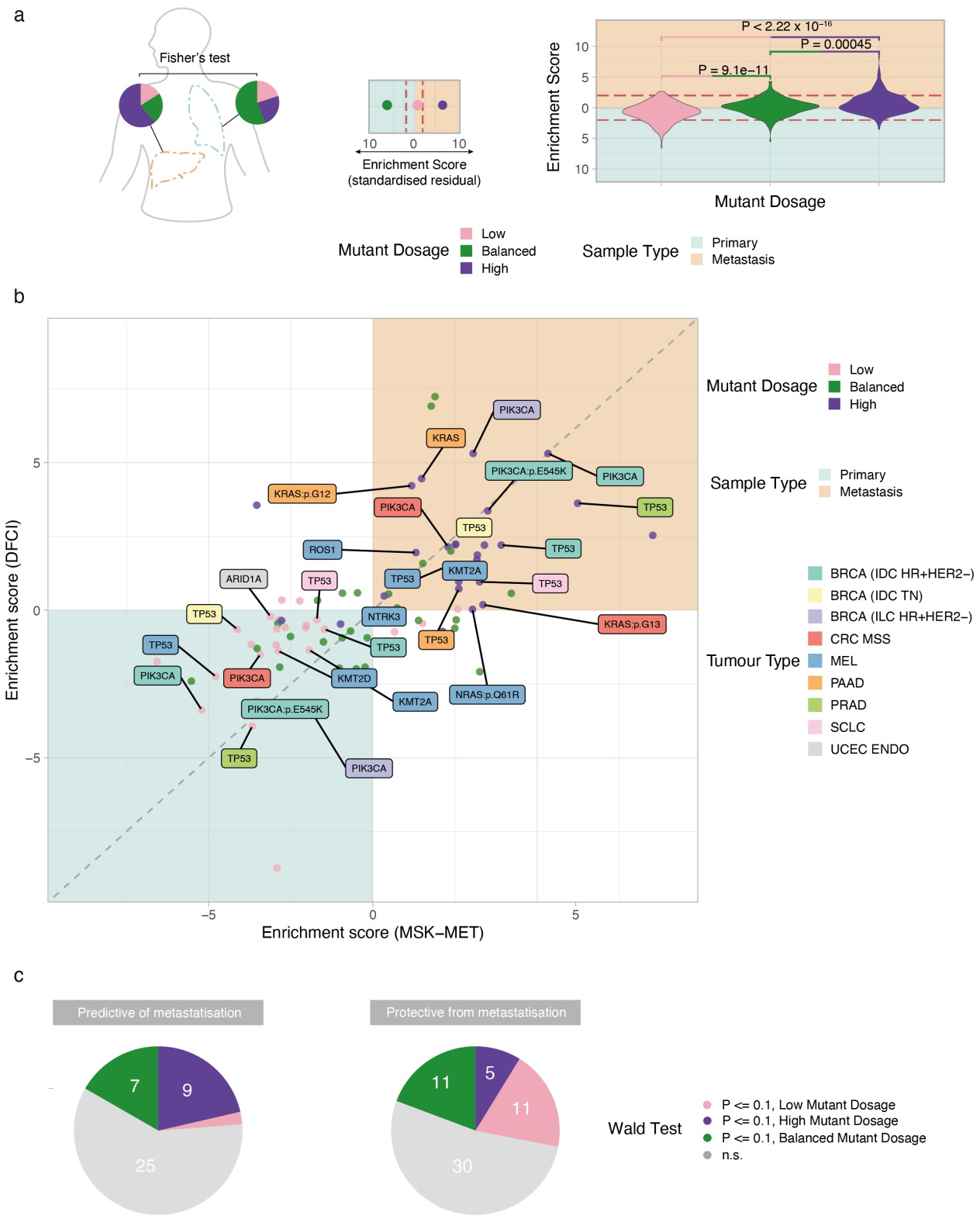
5. Extended Data Figures



Extended Data Figure 1. Report of the INCOMMON fit of a breast cancer (BRCA) sample from HMF, containing 8 mutations sequenced at medain depth 131.5 and with pathologist's tumour purity estimate of 0.33. In the top left panel, the report displays the Poisson likelihood of total read counts parametrised by mutations' total copy number. The grey line and ribbon represent the expected number of reads per k chromosome copies-$k\eta$, and the 95% confidence interval of the posterior distribution of total reads per value of k. Red dots highlight data failing the posterior predictive check (PPC). The central panel displays the posterior distribution of reads with the variant for all sampled values of multiplicity m, with the vertical line indicating the observed value, in red if failing the PPC. The bottom left panels show the comparison between the prior and posterior distributions for the expected count per allele η and the tumour purity π. The rightmost panels display the posterior distribution over all the possible combinations (k, m), with the red marker indicating the maximum a posteriori value.



Extended Data Figure 2. a. Maximally selected rank statistics quantifies the optimal lower and upper cut-points of the fraction of allele with the mutation (FAM) in overall survival (OS) analysis stratifying mutant vs wild-type patients across 84 tumor suppressor genes (TSG, optimal values 25% and 75%), 35 oncogenes (optimal values 33% and 66%) and 18 tumour types. b. Classification of mutation total copy number and multiplicity combinations (k, m) into Low, Balanced and High mutant dosage classes. c. Relative fraction of prognostic gene mutant dosage classes across significant markers. d. Distribution of the fraction of genome altered (FGA) across gene mutant dosage classes of TP53 in tumour types, compared through pariwise Wilcoxon test.



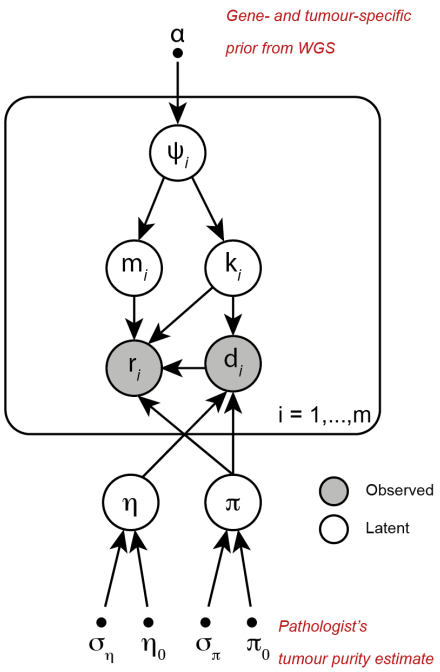
Extended Data Figure 3. a. Distribution of enrichment scores in primary tumours vs metastases, defined as the standardised residuals of Fisher's exact test, across mutant dosage classes. b. Comparison of enrichment scores between the MSK-MET and the DFCI cohorts. Significant markers with consistent enrichment scores (distance lower than independent expected noise) across the two cohorts are highlighted. c. Distribution of mutant gene dosage classes across markers of increased (left) and decreased (right) metastatic propensity using a univariate logistic model.

6. Supplementary Information

6.1. Supplementary Tables

- Supplementary Table S1: Priors on (k, m) from WGS data.
- Supplementary Table S2: Priors on the read count rate per chromosome copy.
- Supplementary Table S3: Assessment of INCOMMON computational efficiency.
- Supplementary Table S4: Survival analysis (Kaplan-Meier method).
- Supplementary Table S5: Survival analysis (Cox regression).
- Supplementary Table S6: Enrichment analysis.
- Supplementary Table S7: Metastatic propensity analysis.
- Supplementary Table S8: Metastatic tropsim analysis.
- Supplementary Table S9: Conversion between HMF and INTOGEN tumour type annotations.

6.2. Supplementary Figures



$$\psi_i \sim \text{Dirichlet}(\alpha_i)$$

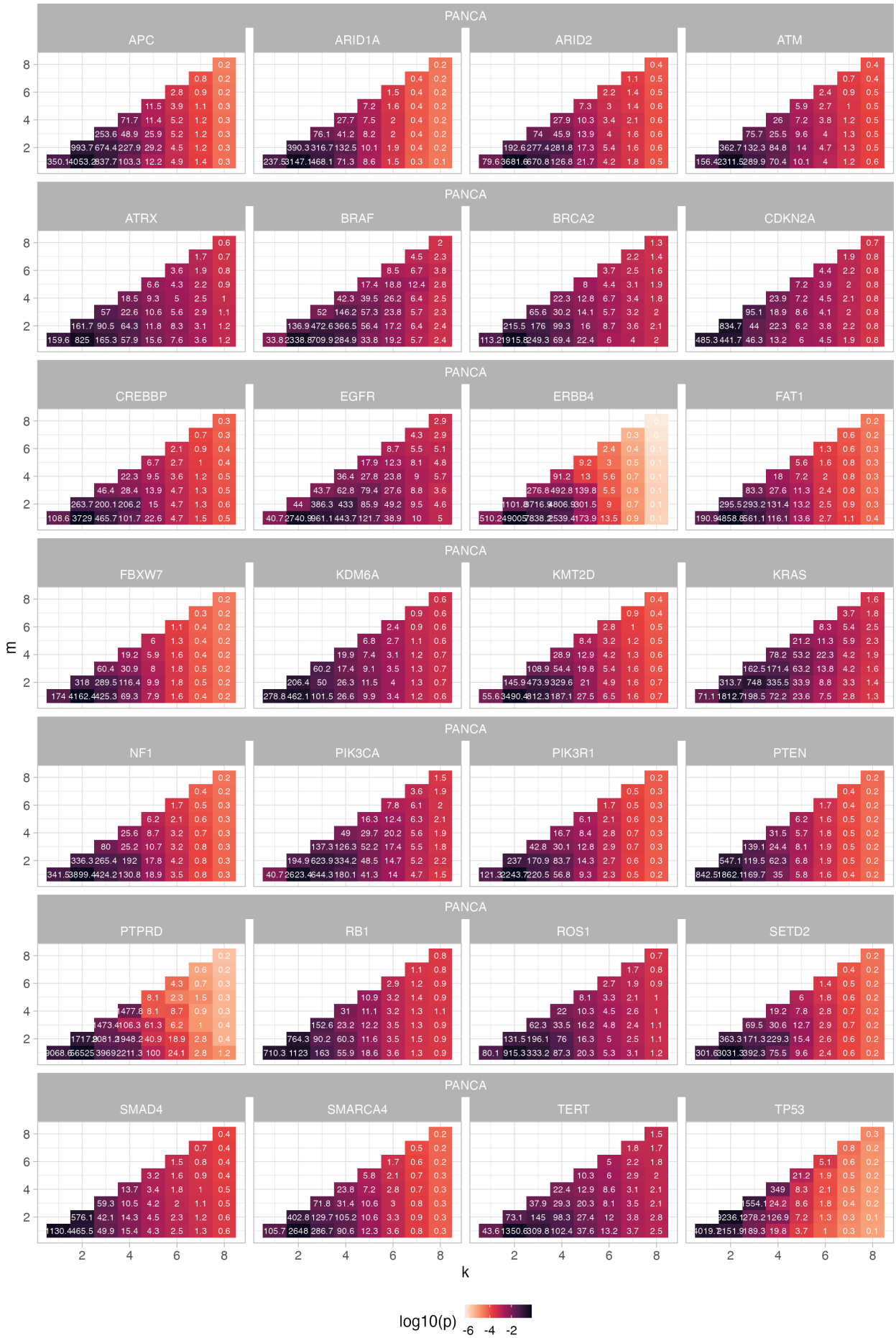
$$d_i \mid \pi, \eta, k_i \sim \text{Poisson}(\lambda_i) \qquad \lambda_i = (1-\pi)2\eta + \pi k_i \eta$$

$$r_i \mid \pi, k_i, m_i \sim \text{Binomial}(\varphi_i) \qquad \varphi_i = \frac{m_i \pi}{2(1-\pi) + k_i \pi}$$

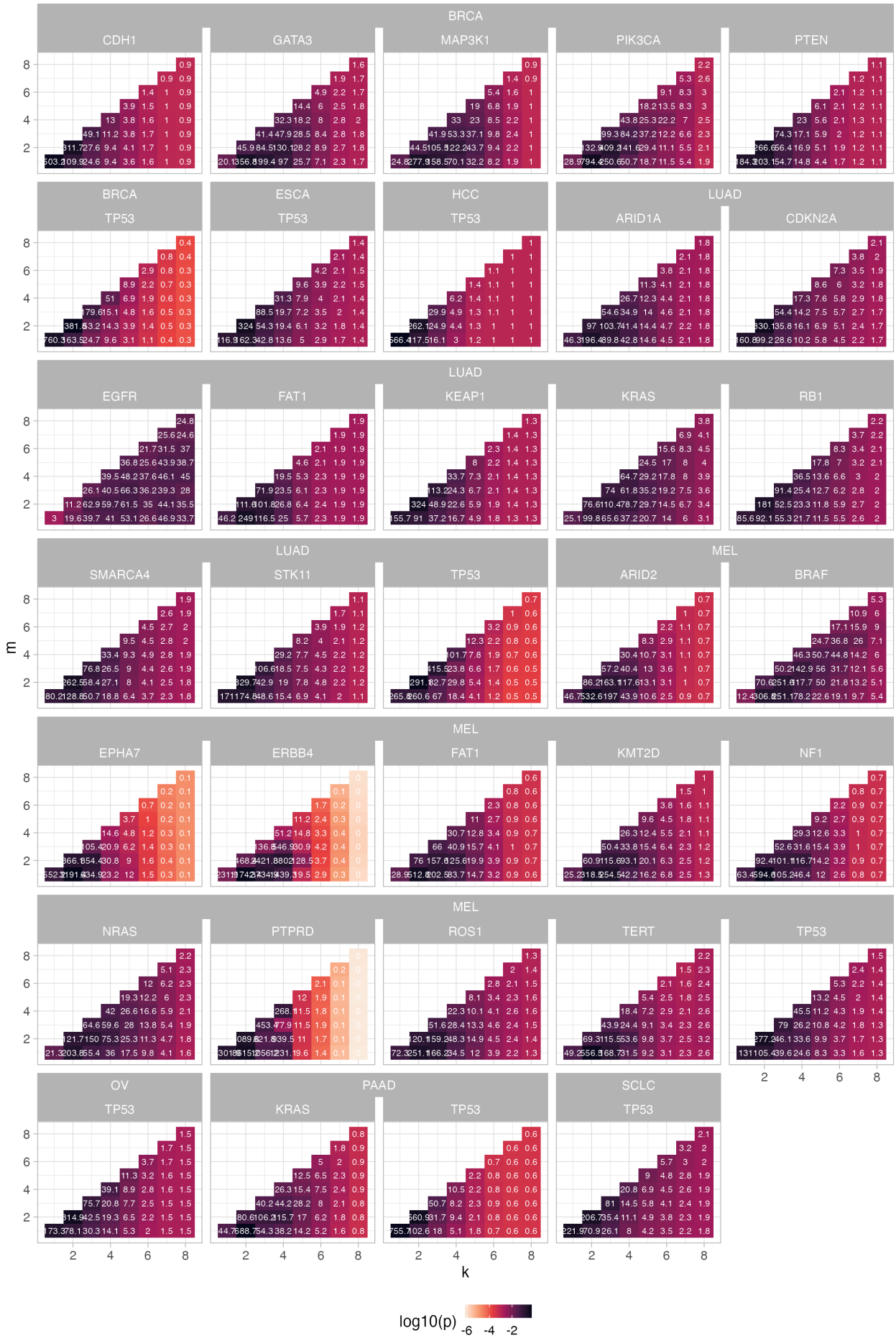
$$\pi \mid \pi_0, \sigma_\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi) \qquad \begin{aligned} \alpha_\pi &= \pi_0\left[\tfrac{\pi_0(1-\pi_0)}{\sigma_\pi^2} - 1\right] \\ \beta_\pi &= (1-\pi_0)\left[\tfrac{\pi_0(1-\pi_0)}{\sigma_\pi^2} - 1\right] \end{aligned}$$

$$\eta \mid \eta_0, \sigma_\eta \sim \text{Gamma}(\alpha_\eta, \beta_\eta) \qquad \begin{aligned} \eta_i &= \sum_k \tfrac{p(k) d_i}{2(1-\pi_0) + k\pi_0} \\ \eta_0 &= \tfrac{1}{m} \sum_{i=1}^{m} \eta_i \\ \sigma_\eta^2 &= \tfrac{1}{m-1} \sum_{i=1}^{m} (\eta_i - \eta_0)^2 \\ \alpha_\eta &= \frac{\eta_0^2}{\sigma_\eta^2} \qquad \beta_\eta = \frac{\eta_0}{\sigma_\eta^2} \end{aligned}$$

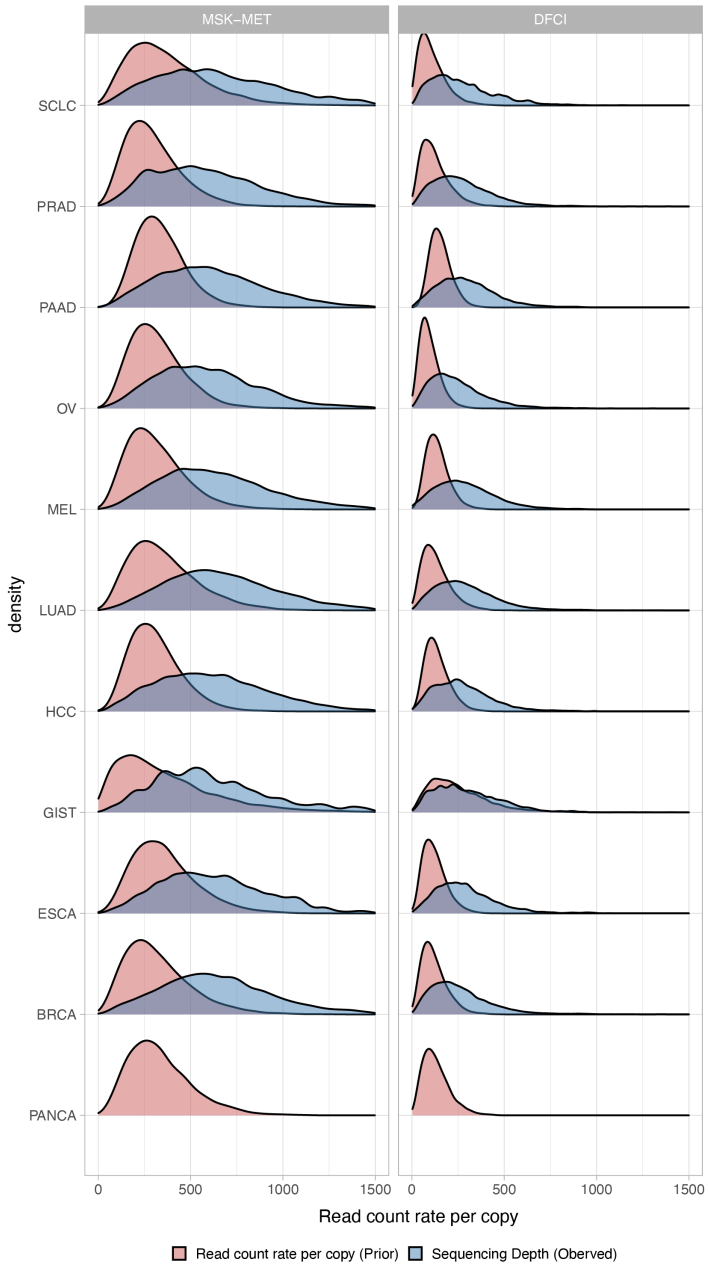
Supplementary Figure S1. Probabilistic graphical model of INCOMMON. Unobserved (latent) variables are indicated as white circles, observed variables as grey circles and fixed hyperparameters as dots. Prior probabilities ψ for each combination of mutation total copy number k and multiplicity m are sampled from a Dirichlet distribution, with concentration parameters α reflecting the relative fraction observed from large-scale WGS cohorts. The prior probability for tumour purity π is centred at the pathologist's estimate π_0 with fixed variance σ_π, whereas the mean η_0 and variance σ_η of the prior probability for the read count rate per chromosome copy η are estimated from the data. The total number of reads at each site d follows a Poisson distribution with rate λ contributed by normal ($(1-\pi)2\eta$) and tumour ($\pi k \eta$) cells. The number of reads with the variant r follows a Binomial distribution with success probability ϕ defined as the expected variant allele frequency $m\pi/(2(1-\pi) + k\pi)$.



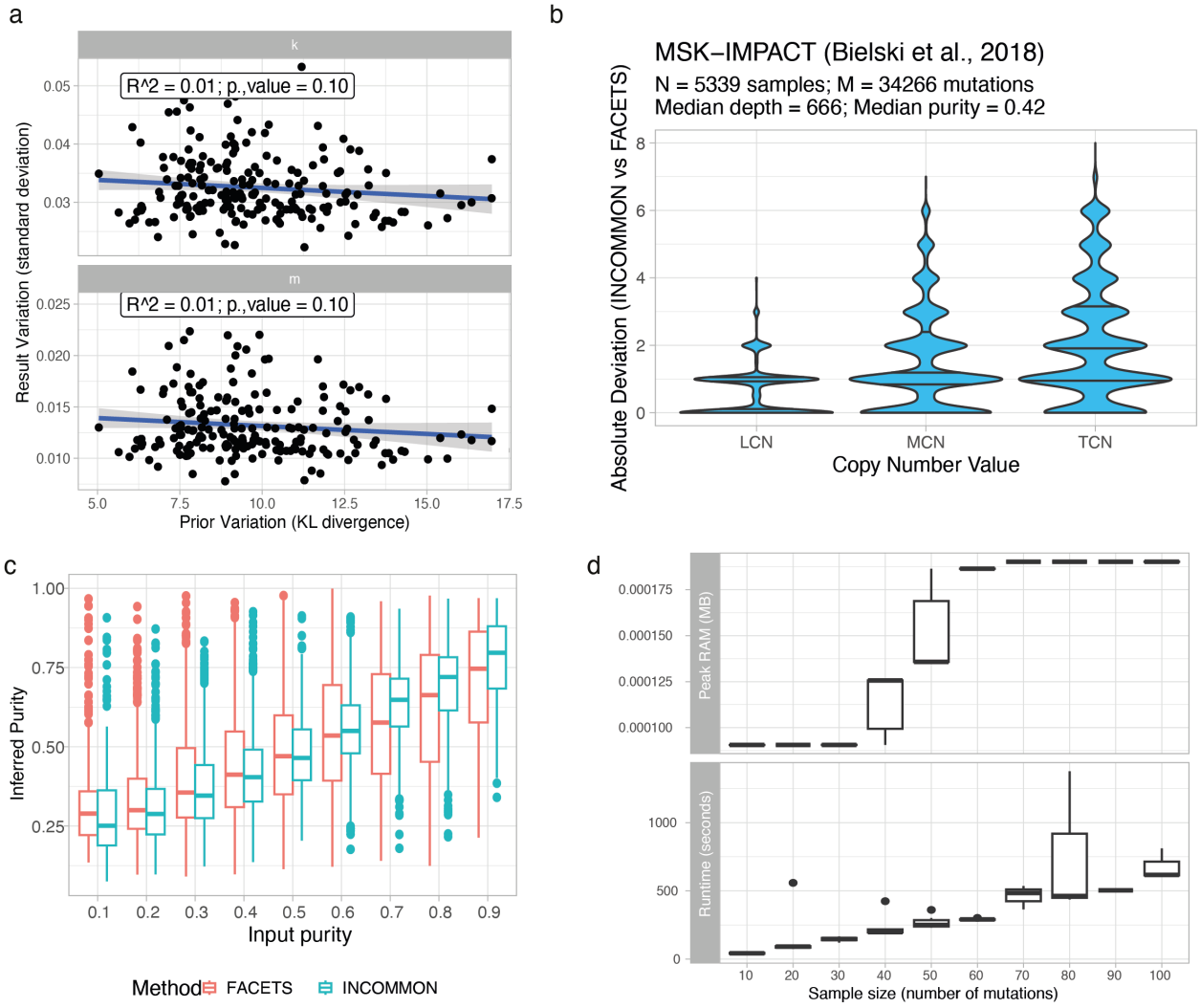
Supplementary Figure S2. Pan-cancer (PANCA) prior distribution on the copy number and multiplicity configurations (k, m) obtained from the PCAWG and HMF whole genome sequencing cohorts. The color of each cell corresponds to the relative fraction of the configuration in the dataset, while the number label inside is equal to the actual count (or pseudo-count, see Methods), used as Dirichlet concentration parameter.



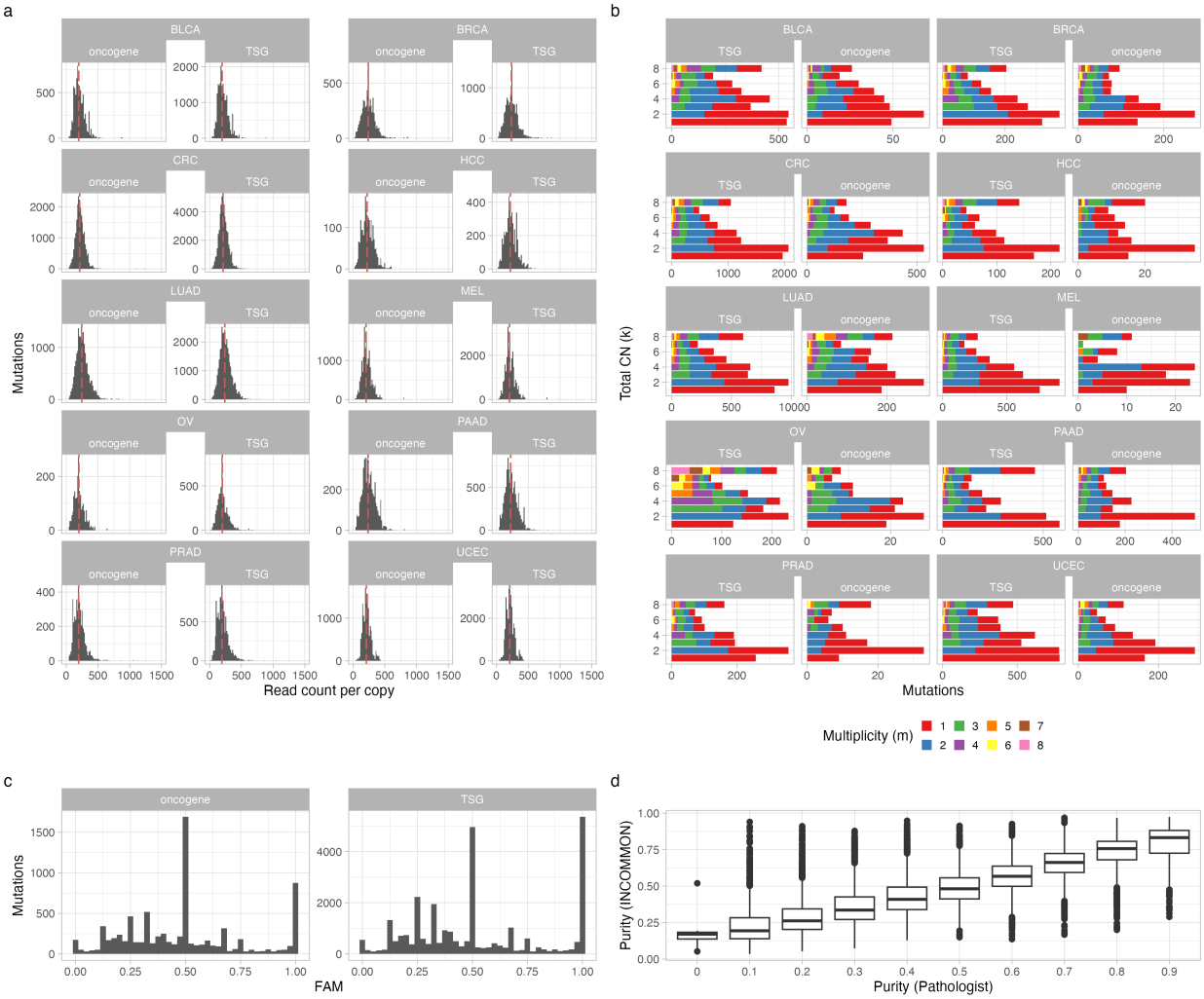
Supplementary Figure S3. Tumour type-specific prior distribution on the copy number and multiplicity configurations (k, m) obtained from the PCAWG and HMF whole genome sequencing cohorts. The color of each cell corresponds to the relative fraction of the configuration in the dataset, while the number label inside is equal to the actual count (or pseudo-count, see Methods), used as Dirichlet concentration parameter.



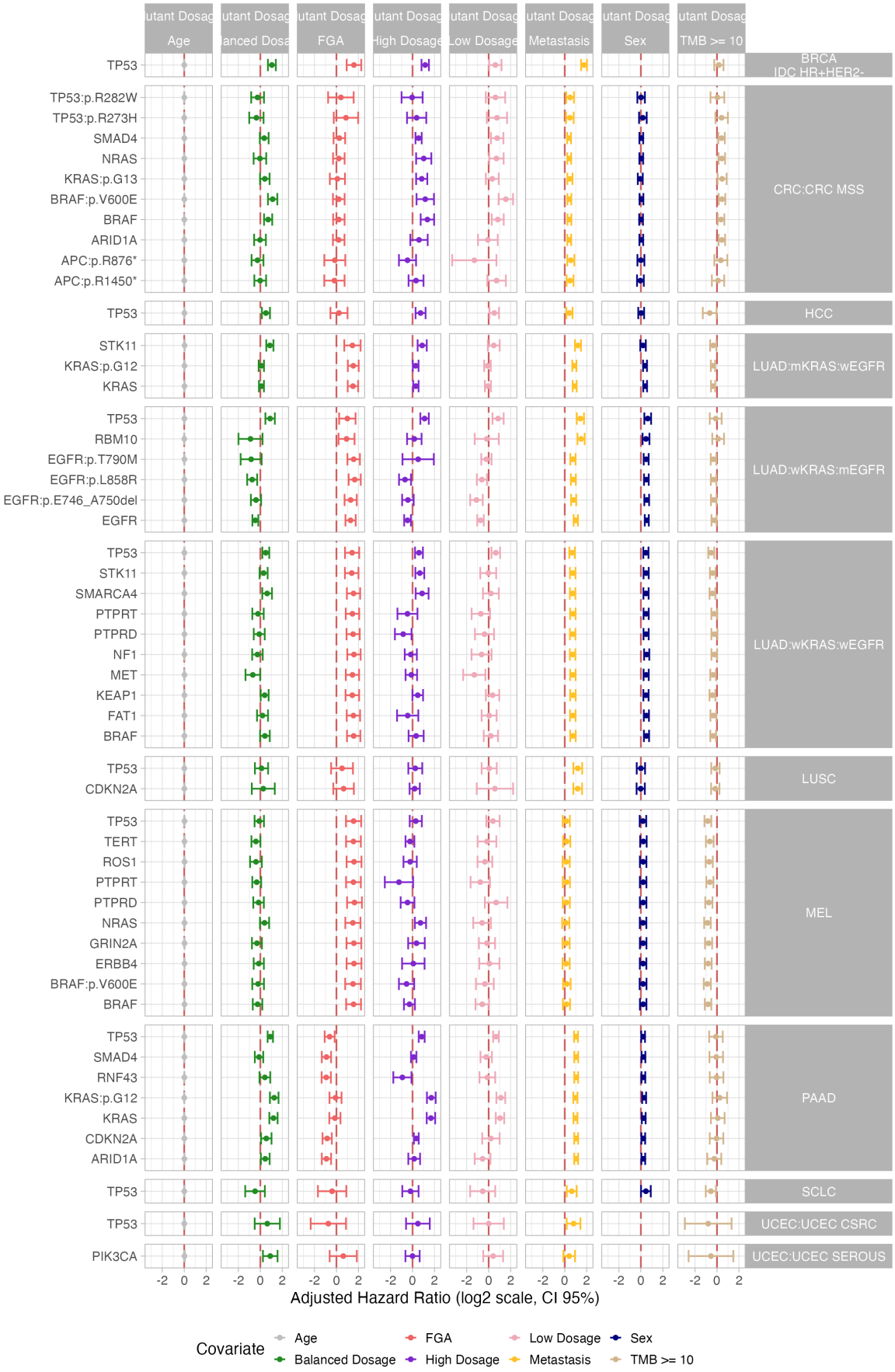
Supplementary Figure S4. Gamma prior distribution (red) over the read count per chromosome copy, compared with the observed distribution of sequencing depth (blue), across the MSK-MET and GENIE-DFCI cohorts. Notice that the empirical distribution of sequencing depth is confounded by mutation copy number and tumour purity.



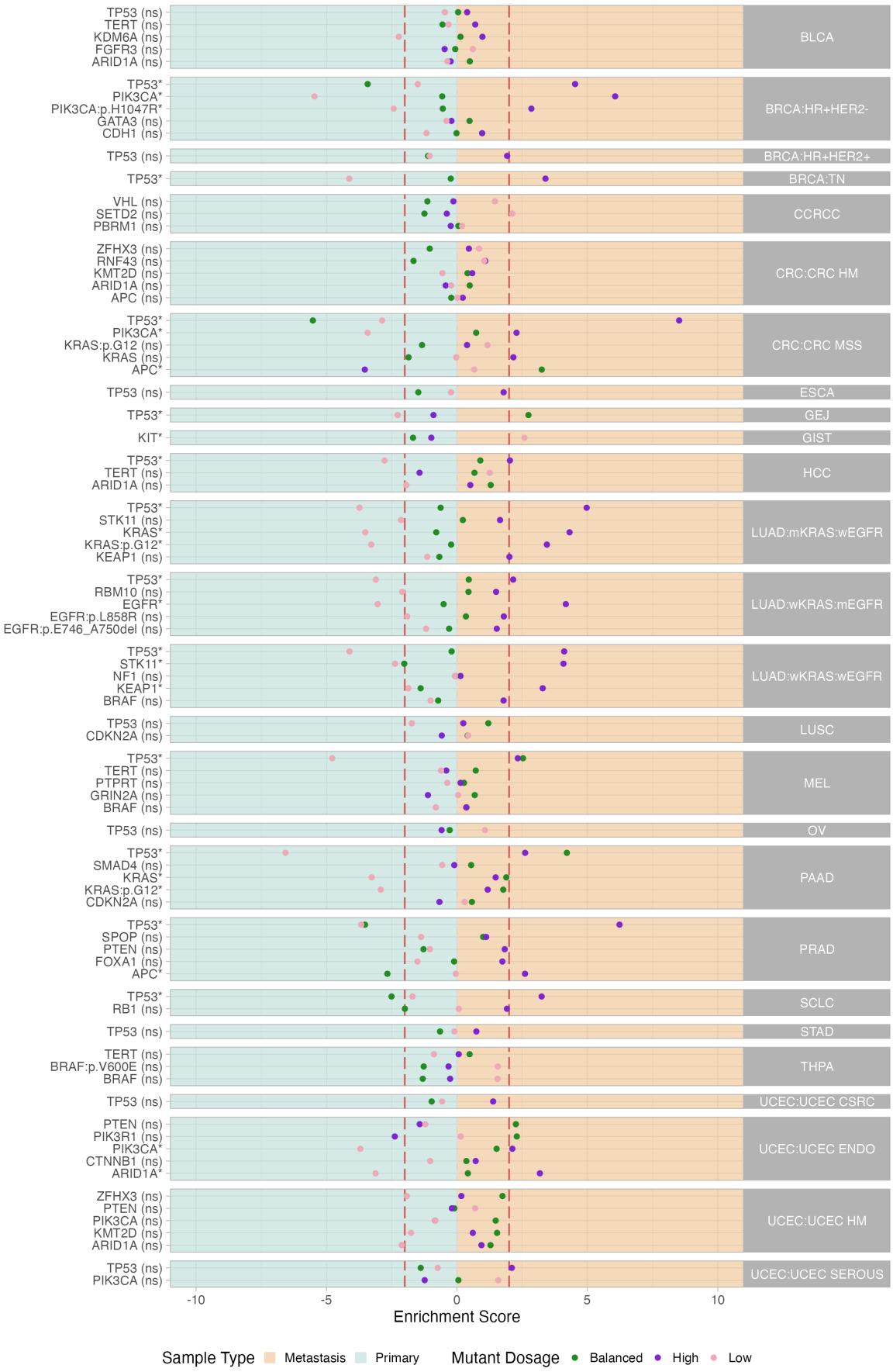
Supplementary Figure S5. a. Standard deviation of the mean absolute error (MAE) estimates against prior variation measured by Kullback-Leibler divergence, across $N = 50$ iterations of random-splitting (70% training, 30% test) of the combined PCAWG and HMF data. b. Comparison of INCOMMON and FACETS copy-number calls across the MSK-IMPACT dataset published in Bielski et. al.[31] c. Comparison of INCOMMON and FACETS inference of tumour purity. d. The low peak RAM usage (top panel) and runtime required for the classification of samples of increasing size (number of mutations) demonstrate the high computational efficiency of INCOMMON.



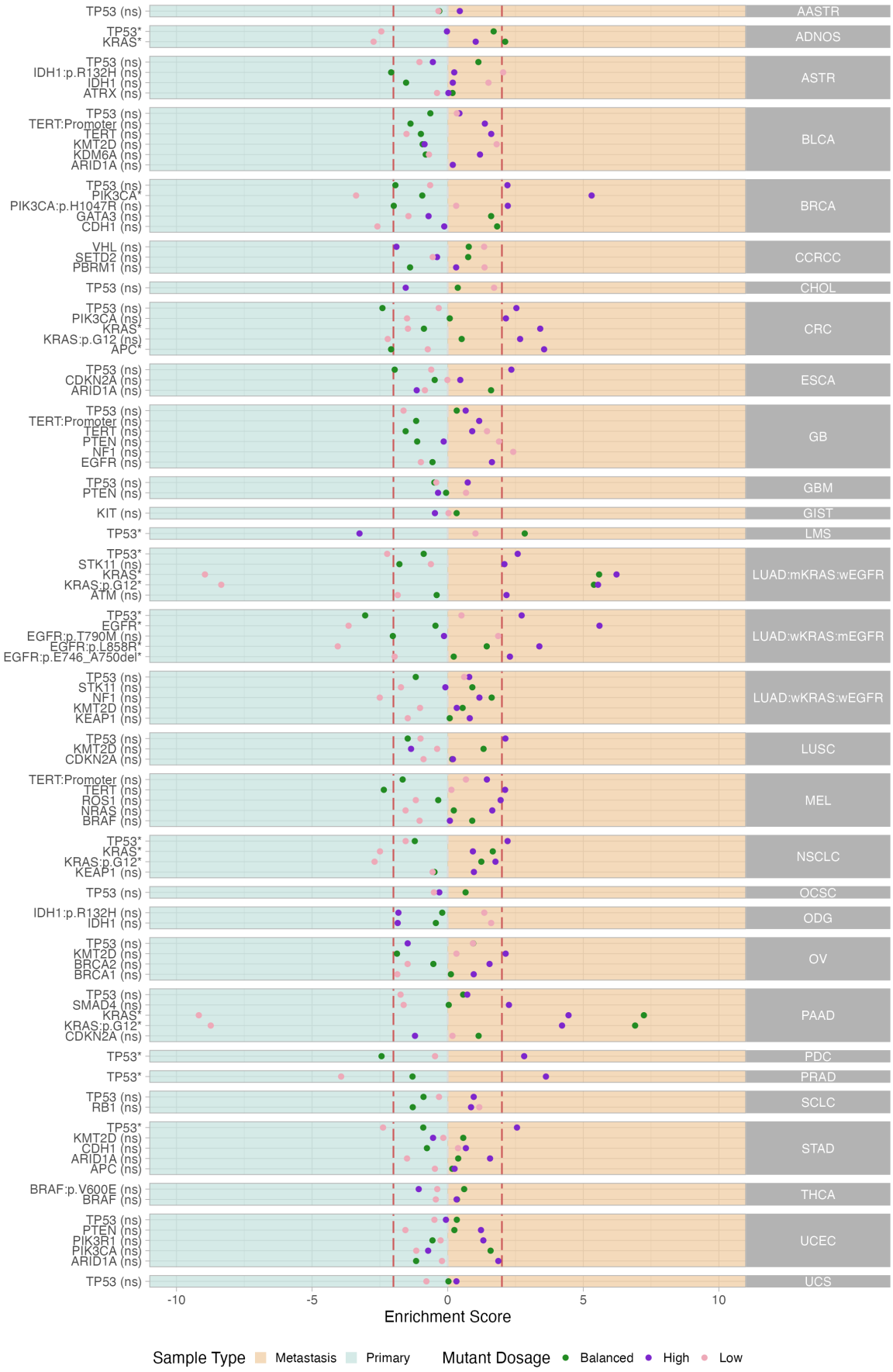
Supplementary Figure S6. a. Distribution of maximum a posteriori (MAP) values of read count rate per copy η, across the main tumour types of the MSK-MET cohort. Red lines indicate median values. b. Distribution of MAP values of total copy number k and multiplicity m. c. Distribution of the MAP values of FAM. d. Comparison of MAP values of the sample purities with the pathologists' estimates.



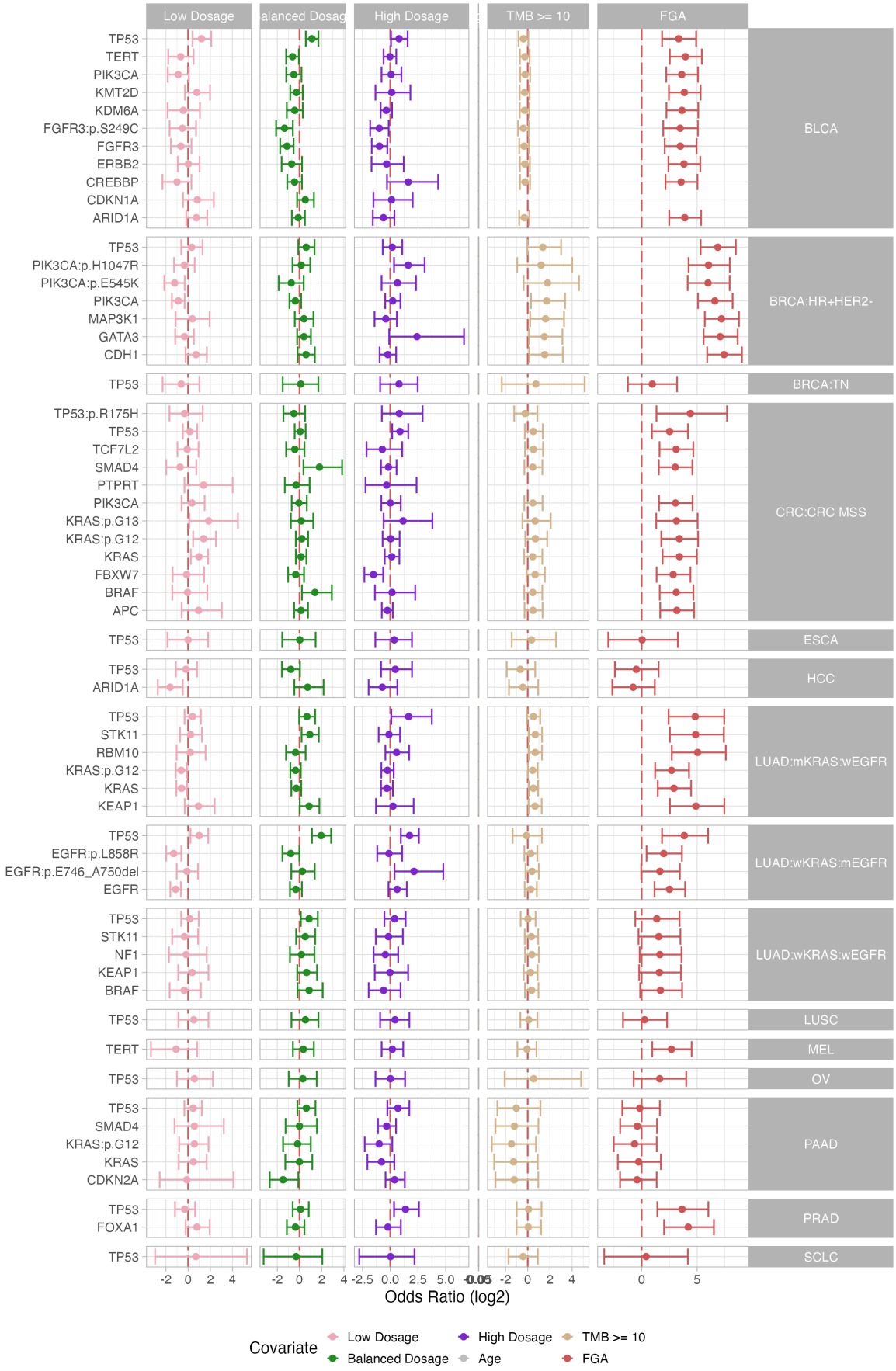
Supplementary Figure S7. Adjusted hazard ratios (HR, log2 scale) with 95% confidence interval for all the covariates included in the multivariate Cox regression models of overall survival. The included covariates are mutant dosage class, age, sex, TMB, FGA and sample type. Red vertical lines indicate HR = 1. For each tumour type, only the most frequently mutated genes are shown.



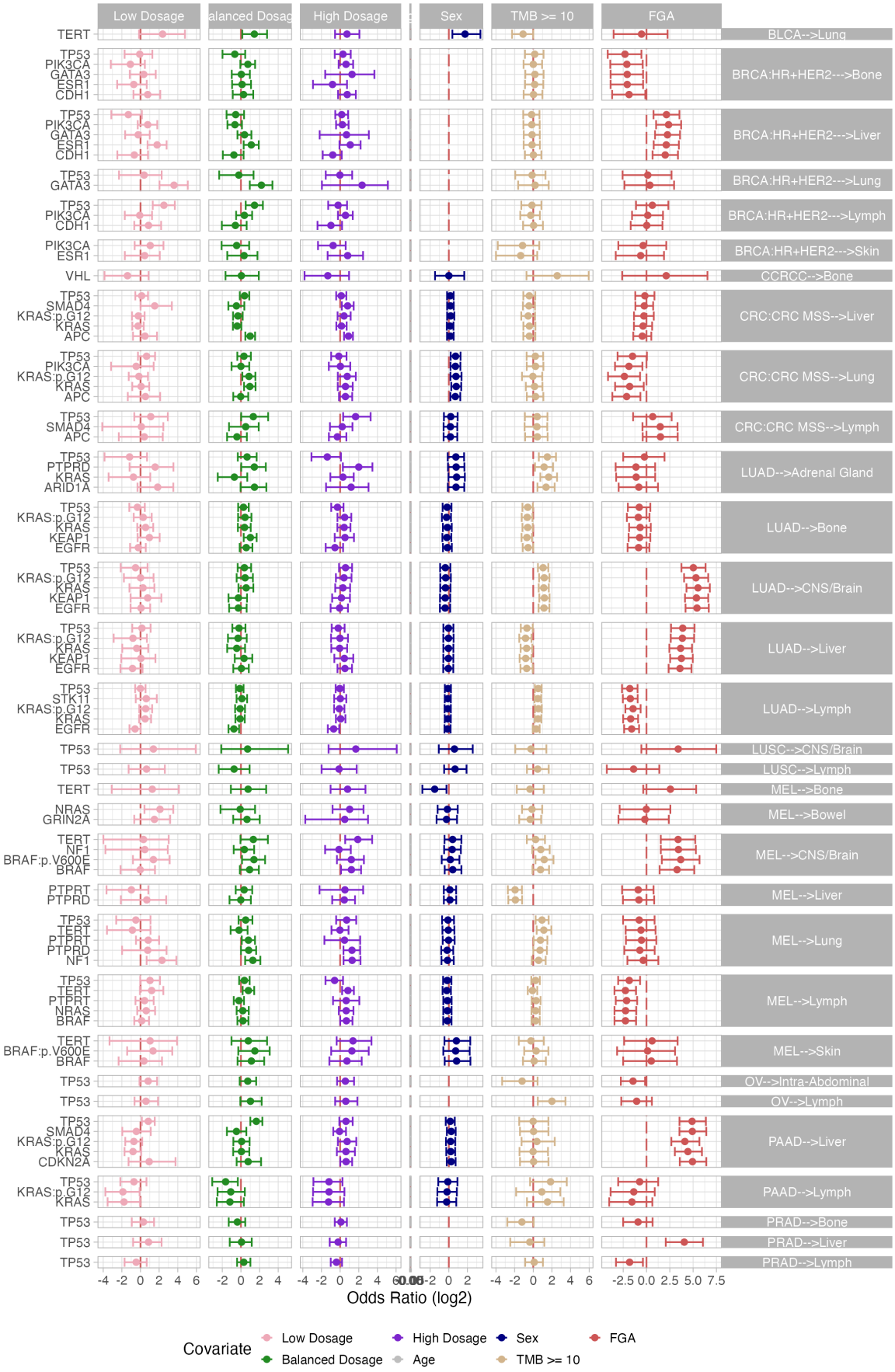
Supplementary Figure S8. Enrichment scores r in primary tumours vs metastases of low, balanced and high mutant dosage classes, across the most frequently mutated genes per tumour type in the MSK-MET cohort. The enrichment score is evaluated through a two-tailed Fisher's exact test with a significance level $\alpha = 0.1$ and using standardised residuals. Vertical dashed lines indicate thresholds of large effect size ($r = 2$).



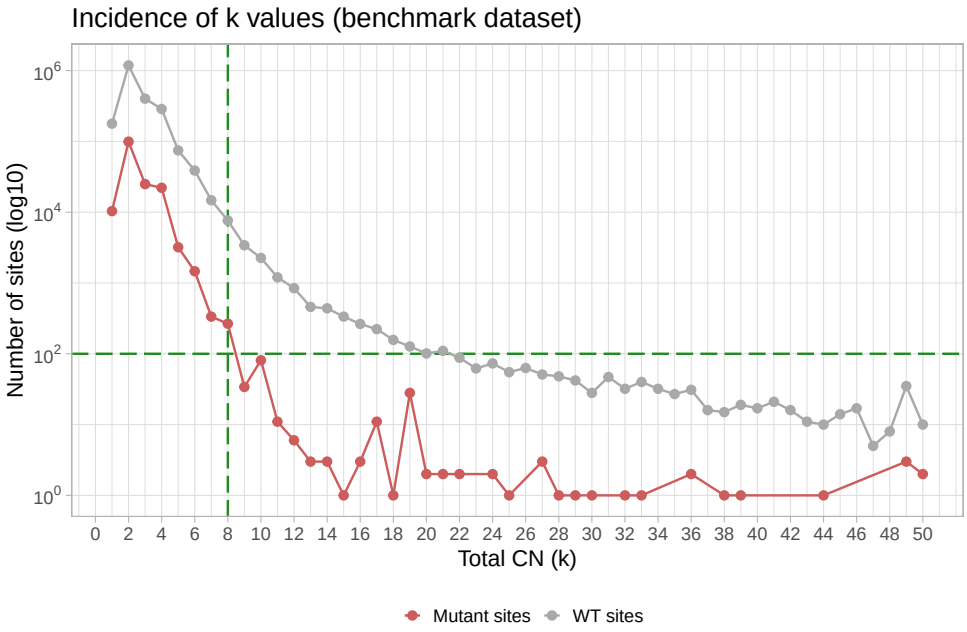
Supplementary Figure S9. Enrichment scores r in primary tumours vs metastases of low, balanced and high mutant dosage classes, across the most frequently mutated genes per tumour type in the GENIE-DFCI cohort. The enrichment score is evaluated through a two-tailed Fisher's exact test with a significance level $\alpha = 0.1$ and using standardised residuals. Vertical dashed lines indicate thresholds of large effect size ($r = 2$).



Supplementary Figure S10. Adjusted OR of metastatisation (log2 scale) with 95% confidence interval, from multivariate logistic regression models. Covariates include low, balanced and high mutant dosage classes, age, tumour mutational burden (TMB) and fragment of genome altered (FGA). Red vertical lines indicate OR = 1. For each tumour type, only the most frequently mutated genes are shown.



Supplementary Figure S11. Adjusted OR of organotropic trajectories (log2 scale) with 95% confidence interval, from multivariate logistic regression models. Covariates include low, balanced and high mutant dosage classes, age, sex, tumour mutational burden (TMB) and fragment of genome altered (FGA). Red vertical lines indicate OR = 1. For each tumour type, only the most frequently mutated genes are shown.



Supplementary Figure S12. Number of mutant (red) and wild-type (gray) sites per total copy number across 8,339 samples from PCAWG, HMF and TCGA. In the analyses reported in this work, we set the maximum total copy number at $k_{max} = 8$ to reduce the computational burden. High copy gains were more frequent in wild-type regions, and for copy number values $k > k_{max}$ we observed a significant drop in the number of mutant sites per copy-number (less than 100).