Mutation frequency and copy number alterations determine prognosis and metastatic tropism in 60,000 clinical cancer samples

Nicola Calonaci (1,+), Eriseld Krasniqi (2,+), Stefano Scalera (2), Giorgia Gandolfi (1), Salvatore Milite (3), Biagio Ricciuti (4), Marcello Maugeri-Saccà (2), Giulio Caravagna (1,5,*)

(1) Department of Mathematics, Informatics and Geosciences, University of Trieste, Trieste, Italy.
(2) IRCCS Regina Elena National Cancer Institute, Rome, Italy.
(3) Centre for Computational Biology, Human Technopole, Milan, Italy.
(4) Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts.
(5) Area Science Park, Trieste, Italy.

(+) These authors contributed equally
(*) Correspondence: giulio.caravagna@units.it

Abstract. The intricate interplay between somatic mutations and copy number alterations critically influences tumour evolution and patient prognosis. Traditional genomic studies often overlook this interplay by analysing these two biomarker types in isolation. Leveraging an innovative computational model capable of detecting allele-specific copy number alterations from clinical targeted panels without matched normal, we conducted a comprehensive analysis of over 500,000 mutations across 60,000 clinical samples spanning 39 cancer types. Our findings uncovered 11 genes and 6 hotspots exhibiting recurrent tumour-specific patterns of co-existing mutations and copy-number alterations across 17 tumours. By stratifying more than 24,000 patients based on these composite genotypes across multiple oncogenes and tumour suppressor genes, we identified 66 groups with distinct prognostic significance, 25% more than using a standard mutation-centric stratification. Notably, 7 groups displayed a heightened propensity for metastasis, while 16 were associated with site-specific patterns of metastatic dissemination. This augmented insight into genomic drivers enhances our understanding of cancer progression and metastasis, holding the potential to significantly foster biomarker discovery.

Statement of significance

By leveraging large datasets and new computational modelling, this study demonstrates the critical interplay between somatic mutations and copy number alterations in driving patient prognosis, tumour progression and metastatic tropism. This work implies a shift towards a more integrative and comprehensive approach in clinical sequencing, with significant implications for biomarker discovery and target identification.

Introduction

Cancerogenesis is a complex, multi-step process involving various genetic and epigenetic changes[1]. Central to this process are somatic mutations and copy number alterations (CNAs), which stochastically confer a proliferative advantage to certain cancer cell subpopulations, leading to the dominance of specific clones[2–4]. Genetic biomarkers derived from such alterations have significantly impacted our understanding of cancer and its treatment, given the expanding repertoire of actionable mutations[5,6]. Indeed, these genomic markers not only provide insights into cancer prognosis but also guide treatment strategies by indicating specific biological vulnerabilities. Despite several successes, however, understanding the prognostic and predictive implications of somatic mutations in key genes remains a challenge[7,8], promoting research towards other genomic levels, e.g. epigenomic alterations[9,10].

Many of the current drawbacks of clinical sequencing stem from a mutation-focused approach. First, biomarkers are often investigated separately for mutations and CNAs, overlooking the intricate interplay and epistasis within cancer genotypes where these alterations can co-occur (Figure 1a). Second, while the size of currently available datasets is increasing, 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[11]. Third, current targeted sequencing for clinical use is not optimised for copy number estimation, as it involves sequencing a very limited fraction of the genome and is performed only on tumour samples without the support of information from a matched normal sample.

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[12,13]. In this work, we leverage new computational modelling inspired by high-resolution whole-genome sequencing to develop INCOMMON, an open-source tool for the inference of allele-specific copy number and mutation multiplicity from clinical targeted sequencing. INCOMMON uses somatic mutation read-count data and bulk sample purity to infer, for every mutation, the number of mutant copies (mutation multiplicity) and the genome ploidy at the locus (Figure 1b). Together, these measurements provide the allele-specific copy number state of the mutation and allow for the interpretation of the joint effect of mutations and aneuploidy on suppressor inactivation or oncogene activation. Notably, INCOMMON uses conventional bulk metrics (sample purity and read counts) from the tumour and, therefore, can be applied without matched normal.

We use INCOMMON to determine the patterns of co-occurrence of mutations and CNAs from 500.000 alterations detected in 39 principal solid cancer types and over 60,000 clinical cancer samples. For the first time, we use large-scale clinical targeted panel data to determine tumour genotypes that combine mutations with CNA, unravelling states of complete suppressor inactivation or oncogene activation from thousands of samples[14]. Notably, we establish a resource that highlights 11 cancer-associated genes and 6 hotspots with mutations frequently

associated with CNAs, 66 groups based on the mutational and CNA profile of whole genes and hotspots that determine patient prognosis (e.g. *PIK3CA E542K* with amplification in breast and colorectal cancers, *KRAS G12C* with amplification, *ARID1A* and *KEAP1* with LOH in lung adenocarcinomas, *KRAS G12D* and *G12V* with amplification in pancreatic adenocarcinomas, *CDK12* with LOH in prostate adenocarcinomas) and augment the resolution of a standard mutation-centric approach (gain of 24.5% using INCOMMON), 7 groups with increased rate of metastasis (*CDH1* with LOH, *PIK3CA H1047R* and *E545K* with amplification in breast cancers, *EGFR E746_A750del* in lung adenocarcinomas, *PIK3CA* with amplification in endometrial cancers) and 16 that explain metastatic tropism (e.g. *PTPRD* with LOH diffusing from primary melanoma sites to the brain, *KRAS G12V* and *G12D* in pancreas diffusing without amplification to the lymphatic system and with amplification to the liver). Our work proposes a novel perspective on genomic alterations in cancer and how these can be used to identify new biomarkers, enhancing our understanding of key cancer biology, such as patterns of survival and metastatic spread.

Results

The INCOMMON copy-number and mutation multiplicity classifier

INCOMMON (Data and software availability) is a classifier that assigns a copy-number state and mutation multiplicity to a mutation. The model is inspired by Bayesian[15] mixtures to leverage an informative classification prior ψ (Figure 1c, Supplementary Figures S1,S2) and report a notion of classification uncertainty (Figure 1d). The prior is derived from 2,777 samples released with the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort, the largest resource of primary tumours with available WGS to date[16]. This prior is built from curated[17] CNAs and can be tumour-specific or pan-cancer (Methods, Supplementary Note; Supplementary Table S1), and serves to inject into our model existing evidence from the most advanced bulk sequencing technology.

For a mutation, INCOMMON computes a classification probability from the mixture likelihood

$$(1) \qquad p\left(n \mid N, \pi, \rho\right) = \sum_{\eta} \psi_{\eta} \text{Beta-Binomial}\left(n \mid N, f_{\eta}\left(\pi\right), \rho\right)$$

which depends on four parameters linked to the sequencing assay, technology and sample quality. Here, N is the number of reads that cover that mutation locus, of which $n < N$ harbour the mutant allele. Instead, $0 < \pi \leq 1$ is the sample purity, i.e., the proportion of tumour DNA in the sequenced sample, determined either by sequencing or pathology assessment. Finally, the Beta-Binomial likelihood depends on $\rho \geq 0$, the sequencing dispersion that can be hold fixed based on sequencing features, and the known parameters of the distribution $f_{\eta}(\pi)$ (Methods).

The classes η predicted by INCOMMON are triples (n_A, n_B, m), where n_A and n_B are the major and minor allele copies of the genome at the mutation site so that the tumour genome ploidy is $p = n_A + n_B$ and the mutation is present in $m \leq \max(n_A, n_B)$ copies (Figure 1d). The multiplicity m also defines the fraction m/p of mutant alleles, and the fraction $(p - m)/p$ of wild-type (WT) alleles. INCOMMON can classify 6 configurations, of which 5 identify clonal mutations (present in 100% of the tumour) associated with CNAs characterised by a high fraction ($m/p > 1/3$) of mutant alleles. These somatic alterations were defined as Tier-1 because they emerge early in tumour development, and include heterozygous mutant diploid states without copy numbers (HMD; $n_A = n_B = m = 1$); loss-of-heterozygosity in single-copy (LOH; $n_A = m = 1$, $n_B = 0$) and copy-neutral states (CNLOH; $n_A = m = 2$, $n_B = 0$), trisomy ($n_A = 3$, $n_B = m = 2$) and tetrasomy ($n_A = n_B = m = 2$) amplifications (AM) of mutant alleles. The sixth configuration is instead Tier-2, and captures subclonal (present less than 100% of tumour cells) or clonal mutations with high-ploidy and/or low-multiplicity ($m/p \leq 1/3$). Note that Tier-1 classes associated with copy-gains (CNLOH and AM) capture temporal ordering, with mutation preceding amplification. For example, for CNLOH the copy-gain after LOH carries the mutant allele to multiplicity $m = 2$, whereas mutations arising after copy-gain CNLOH/AM are Tier-2 because the mutant is only present in one copy ($m = 1$). Our distinction is meant to prioritise mutation-focused evolutionary paths sustained by aneuploidy[3,4,18].

INCOMMON computes a posterior probability for the assignment $z_\eta = 1$ to class η as

$$(2) \qquad p(z_\eta = 1 \mid n, N, \pi, \rho) = \frac{\psi_\eta \text{Beta-Binomial}(n \mid N, f_\eta(\pi), \rho)}{\sum_\eta \psi_\eta \text{Beta-Binomial}(n \mid N, f_\eta(\pi), \rho)}$$

and from z_η it derives the uncertainty in the class prediction η using the notion of entropy[17,19]. The complete mathematical formulation of the model and its implementation is available as Methods. Mutations mapping to oncogenes or tumour-suppressor genes (TSG) classified by INCOMMON can be interpreted (Figure 1d): mutant TSG inactivation is determined by lack of WT alleles (LOH/CNLOH classes), and oncogene hyper-activation by mutant copy-gains (AM/CNLOH classes). A limitation that affects TSGs for models like INCOMMON is that, without haplotyping, inactivations by concurrent mutations are impossible to determine.

We assessed the performance of INCOMMON with 17,950 mutations detected by whole-exome sequencing in 910 samples from The Cancer Genome Atlas (TCGA)[20], with validated[17] allele-specific CNAs, and quantified model accuracy as the fraction of correct predictions. Since TCGA contains data with different quality, we tested how the performance changes if we reject classifications with excessive uncertainty or samples with low effective coverage (sample purity times sequencing depth)(Figure 1e, Supplementary Figure S3). In the former case, the model achieved a maximum accuracy of 75.4% retaining classifications with entropy below 55%. The

latter test achieved generally higher accuracy with a maximum of 86.4% for effective coverage of 420, and of 77.5% at the median effective coverage of the MSK-MetTropism (MSK) cohort. Since for high enough effective coverage (above 100) the accuracy did not change by varying values of the entropy cutoff, we decided to drop it in the classification, which allowed maximisation of the numerosity of patient groups in survival and metastatic pattern analysis. Overall, we found good agreement between the proportions of each class (Supplementary Figure S4), with a tendency to under-estimate the fraction of HMD mutations (7% decrease in TSGs, 5% in oncogenes) in favour of CNLOH events.

Pan-cancer patterns of mutation and copy numbers

We used INCOMMON (default parameters) to analyse the MSK MetTropism (MSK) cohort with 24,755 samples profiled using MSK-IMPACT panels (versions IMPACT341, IMPACT468 and IMPACT410), and the GENIE-DFCI (DFCI) cohort with 39,152 samples profiled using the DFCI-ONCOPANEL panels (versions 1.0, 2.0, 3.0 and 3.1). These panels covered a shared set of 237 genes, of which 137 were TSGs and 100 oncogenes according to the COSMIC Cancer Gene Census[21] (version 98). We classified 545,531 mutations across 39 tumour types and derived statistics for both whole genes and gene hotspots. We determined the enrichment of each gene towards one of the two interpreted genotypes (with or without LOH for TSGs and with or without amplification for oncogenes), using a one-tailed Chi-squared test, with tumour-specific expected frequencies derived by average separately for TSGs and oncogenes (Supplementary Figure S5). We corrected p-values for multiple hypotheses testing using the Benjamini-Hochberg method and considered the enrichment significant if the adjusted p-value was $p \leq 0.05$. Selected results are reported in Figure 2a, while extended results are shown in Supplementary Figures S6,S7 and the full table is in Supplementary Table S2. Our results are also available as a web-based resource that is freely available at https://ncalonaci.shinyapps.io/incommon/, that allows immediate inspection of our and new analyses.

Overall, 112,169 mutations were associated with CNAs (45.5% of total): 12.9% with amplifications, 8.3% with LOH and 24.3% with CNLOH (31,790, 20,380 and 59,999 mutations). We classified 23.3% (57,552 mutations) of the total as Tier-2 (Figure 2b). The genotypes with co-existing mutations and CNAs were observed in 59% (53,258 mutations) of the cases for TSGs, and 46% (25,193 mutations) for oncogenes; the different incidence was statistically significant ($p < 0.0001$, Figure 2c). Among variants represented in at least 100 samples, we found 85 tumour-specific gene/hotspot alterations enriched toward a genotype either with (11 genes, 5 hotspots) or without (19 genes, 1 hotspot) CNA, across 17 cancer types. For a complete description of the results, including cases significant in only one cohort (MSK or DFCI), see the Supplementary Note.

Consistently with previously reported whole-exome data[22], *TP53* was the most commonly altered TSG[23,24] in both datasets and across all tumour types (10,989 samples with Tier-1 mutations in MSK; 14,481 in DFCI). These mutations were almost systematically associated with LOH in MSK (10,628 samples, 96.7%) and DFCI (13,392 samples, 92.5%) with significant

enrichment in all 9 most abundant tumour types (Figure 2a): lung adenocarcinoma (LUAD), colorectal cancer (CRC), breast cancer (BRCA), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), melanoma (MEL), endometrial cancer (UCEC), ovarian cancer (OV) and bladder cancer (BLCA).

KRAS was the most frequently mutated oncogene[25] in LUAD, CRC, PAAD and UCEC with 4,595 samples having Tier-1 mutations in MSK, and 4,715 in DFCI. *KRAS* mutations were frequently associated with amplification in MSK (2,178 samples, 47.4%) and DFCI (1,830 samples, 38.8%), although not enriched in any of the tumour types. *KRAS*-induced oncogenesis[26] is linked with activating mutations or amplification of the wild-type (WT) allele. Recently[27], a refined view underscores that *KRAS* mutant allelic imbalance is as frequent as heterozygosity, and that homozygosity is typically due to loss of WT with subsequent amplification (i.e., CNLOH) of the mutant allele. In agreement with these observations, we found heterozygous *KRAS* mutations in 52.6% of the MSK and 61.2% of the DFCI cases (2,417 and 2,885 samples), and that amplified mutations co-occur with WT loss in 40% of the cases (overall MSK and DFCI).

Tier-1 mutations of the PI3K/AKT/mTOR signalling pathway were also ubiquitous. For *PIK3CA,* these were observed in 4,108 CRC, BRCA and UCEC samples, for *PTEN* and *PIK3R1* in 2,797 CRC, BRCA, PRAD and UCEC samples. Extensive genomic, transcriptomic and proteomic analyses highlighted a recurrent alteration pattern of this pathway, with *PTEN* inactivating mutations often undergoing LOH, whereas *PIK3CA* and *PIK3R1* undergo single-hit point mutations or WT amplifications[28]. We found mutations in *PIK3CA* and *PIK3R1* systematically associated with no copy number, while in *PTEN* with LOH (largest $p = 0.006$). Related to the mTOR pathway, the association with LOH was also significant for *STK11* in LUAD (MSK and DFCI $p < 0.0001$). Our findings were consistent with previous works[28].

Besides genes with significant co-occurrence of mutations and CNA in multiple cancers, we observed additional associations at unprecedented resolution in the largest groups of cancers (LUAD, CRC, BRCA, PAAD, PRAD, MEL, UCEC, OV, BLCA). In LUAD, we found a tendency of *EGFR*[29] mutations to co-occur with amplification, with $p < 0.0001$ in MSK, but not significant (n.s.) in DFCI, and the general trend of *KRAS* without amplification also held for *G12C* hotspot mutations (n.s. in MSK, DFCI $p < 0.0001$). Mutations in *KEAP1*, *CDKN2A* and *RB1* were enriched with LOH (largest $p = 0.01$), with *ATM* and *SMARCA4* presenting a similar tendency. In CRC we found association of *APC* and *SMAD4* mutations[30] with LOH (largest $p < 0.0001$). Among the mutations most frequently co-occurring with LOH emerged the *R282W* hotspot of *TP53* (largest $p < 0.0001$), and *TCF7L2* with amplification (MSK $p < 0.0001$, n.s. in DFCI). *NOTCH1* and *ARID1B* curiously presented opposite trends in MSK and DFCI. In BRCA we found an association of *CDH1* (n.s. in MSK, DFCI $p = 0.005$) mutations with LOH, and an opposite trend for *GATA3* (larges $p < 0.0001$). In PAAD, the hotspots *G12V* and *G12D* presented an analogous trend of general *KRAS* mutations without amplification, while *CDKN2A* mutations[31] were associated (larges $p < 0.0001$) with LOH, as well as *SMAD4* (largest $p < 0.0001$). In PRAD, *APC* mutations frequently co-occurred with LOH (MSK $p = 0.006$, n.s. in DFCI), whereas *FOXA1* (n.s. in MSK, $p = 0.003$ in DFCI) and *SPOP* (largest $p < 0.0001$) occurred mostly without CNA. In MEL, *TERT* and its

promoter were associated with the absence of amplification (largest p = 0.001), *CDKN2A* with LOH (largest p < 0.0001), whereas the tendency for *NRAS* was not consistent across the two cohorts. In UCEC, mutations in *ARID1A* were associated with no LOH (larges p = 0.004), whereas *FGFR2* was associated with amplification (MSK p < 0.0001, n.s. in DFCI). In BLCA, *KDM6A* mutations were associated with LOH (largest p < 0.0001), while *TERT* and its promoter and *KMT2D* presented the opposite trend, and mutant *FGFR3* hotspot *S249C* co-occurred with amplification (MSK p < 0.0001, n.s. In DFCI).

Prognostic effects of mutations and CNAs at the single gene level

After establishing patterns of mutations and CNAs in cancer genes, we used INCOMMON to test if mutations with or without CNA added prognostic value to a baseline classification of "Mutant" (mutation regardless of its copy number state) versus wild-type (WT) (Methods). We measured as outcome the overall survival (OS) available in the MSK-MET cohort (Supplementary Note), and compared OS via the Kaplan-Meier estimator and a multivariate Cox regression adjusted for sex, age, and tumour mutational burden (TMB). We reported hazard ratios (HR) and p-values computed via the Wald test with false-discovery rate correction (cutoff $p \leq 0.05$ for the adjusted p-value) in three settings: the baseline setting (HR of mutants vs WT), the INCOMMON analysis (HR of mutants with/without LOH or amplification vs WT) and one in which we computed the HR of the mutants with LOH or amplification versus without. The output of this analysis provides, therefore, an estimation of whether mutations are prognostic *per se* or if the combination with CNAs highlight different prognosis scenarios, and of how much the presence of CNA decreases survival.

Notably, out of 191 groups (considering both whole genes and gene hotspots), we found more prognostic groups compared to the baseline. In total (Figure 3a), 66 groups correlated with either a negative or a positive influence on OS after INCOMMON classification, whereas only 53 were prognostic at baseline (24.5% increase). Considering 138 cases that failed to be prognostic at baseline, we found 20 of them prognostic with INCOMMON (15% recovery), showing that our method offers more valuable insights. Only in 7 cases prognostic at baseline, the groups discriminated by INCOMMON had no prognostic power when taken individually, due to the resulting groups becoming too small to achieve statistical significance. Survival curves for any grouping can be browsed at https://ncalonaci.shinyapps.io/incommon/.

Overall, mutations with CNA systematically exacerbated the negative prognostic impact on survival. The trend was similar (Kruskal-Wallis rank sum test p = 0.62) for TSGs and oncogenes, with median HR = 1.32 for mutant TSGs with versus without LOH and HR = 1.43 for mutant oncogenes with versus without amplification (Figure 3b).

In 28 cases (Figure 3c) INCOMMON classes demonstrated statistically significant prognostic value (adjusted $p \leq 0.05$) that recapitulates two distinct scenarios: i) when the baseline test lacked prognostic capability, but INCOMMON did not and ii) when the baseline was prognostic,

and INCOMMON clarified if the presence or absence of CNAs was prognostic. All results are in Supplementary Figures S8-S9, where we also report other significant associations that however deviate from this explanation. All hazard ratios (HR) and 95% confidence intervals (CI) are available in Supplementary Table S3 and omitted (in most cases) for the sake of brevity in this text.

The baseline test lacked prognostic capability that was recovered by INCOMMON in 20 cases (Figure 3c): in BRCA for *GATA3* without LOH (HR = 0.67, p = 0.035), in CRC for *KRAS G13D* (HR = 1.61, p = 0.001, Figure 4a), *PIK3CA* (HR = 1.50, p = 0.003, Figure 3d) with its hotspot *E545K* (HR = 1.93, p = 0.004, Figure 4b) with amplification and *KDR* without (HR = 0.29, p = 0.014), in LUAD for *APC* (HR = 1.56, p = 0.03) and *ARID1A* with LOH (HR = 1.99, p = 0.007), *CTNNB1* (HR = 0.63, p = 0.034) without amplification, *ERBB2* (HR = 1.41, p = 0.04) and *KRAS* hotspots *G12C* (HR = 1.34, p = 0.007, Figure 4c), *G12D* (HR = 1.47, p = 0.036) and *G12V* (HR = 1.38, p = 0.05) with amplification and *SETD2* without LOH (HR = 0.66, p = 0.03), in MEL for *BRAF* without amplification (HR = 0.66, p = 0.025), *TERT* (HR = 0.68, p = 0.02) and its promoter (HR = 0.69, p = 0.027) without LOH, in PAAD for *KRAS G12V* (HR = 1.27, p = 0.039 with amplification, HR = 0.77, p = 0.029 without), in PRAD for *APC* (HR = 1.69 p = 0.011) and *CDK12* with LOH (HR = 1.94, p = 0.008, Figure 4d), and in UCEC for *KRAS* without amplification (HR = 0.47, p = 0.015).

In 8 cases the baseline classification was prognostic, and INCOMMON clarified the class driving the prognosis: *PIK3CA* without amplification in BRCA (HR = 0.63, p < 0.0001), *KRAS* with amplification in CRC (HR = 1.53, p < 0.0001), *KRAS* with amplification (HR = 1.40, p < 0.0001) and *KEAP1* with LOH (HR = 2.16, p < 0.0001, Figure 3e) in LUAD, *KRAS* (HR = 1.86, p < 0.0001) and its hotspot *G12D* (HR = 1.52, p < 0.0001) with amplification in PAAD, *PIK3R1* without LOH (HR = 0.36, p = 0.0002, Figure 4e) in UCEC.

In all cases where an INCOMMON class was prognostic, mutations with CNAs determined a negative prognosis, whereas we also identified mutations that, without CNAs, indicated a positive prognostic effect. The case of *KRAS G12V* in PAAD (Figure 3f) was exemplary, as the presence and absence of amplification determined negative and positive prognosis, respectively. A unique case where the mutation with CNA signalled a positive prognosis (albeit worse than without) was *EGFR* in LUAD (HR = 0.56, p < 0.0001 without amplification, HR = 0.85, p = 0.049 with, Figure 4f). This instance might reflect conditions of mutual exclusivity with other driving genomic or biological factors that are predominant in WT cases.

Prognostic signatures of mutations and CNAs from two genes

Cancer develops from the accumulation of multiple somatic mutations, so we investigated whether mutant gene pairs, with or without CNAs, could manifest some degree of interaction in determining prognosis. We chose to analyse gene pairs to avoid too small groups achieved with more complex pairings (Supplementary Figure S10).

We repeated the MSK-MET analysis with gene pairs (Methods) significantly associated with OS from the last section (Figure 3), estimating Kaplan-Meier curves and computing hazard ratio coefficients with multivariate Cox regression and Walt test for statistical significance. We selected the top 5 frequently mutated per tumour type and retained only groups with at least 30 samples using, as reference group, joint mutants without CNAs (Supplementary Note). All results are in Supplementary Table S4.

In 9 cases (60%, 8 significant) the group with the worst outcome was that with combined mutation and CNA on both genes, in 4 cases (27%, 1 significant) the one with additional CNA only on one gene and in 2 cases (13%) the group with no CNA in either genes. All the analyses are shown in Supplementary Figure S11 (multi-page figure). Overall, our INCOMMON analysis confirmed that, for almost every pair of genes considered, the joint effect of CNAs in the background of double-gene mutants generally has effects on OS, with a median hazard ratio of HR = 1.94.

In CRC the amplification of *KRAS* (Figure 5a) and *PIK3CA* (Figure 5b) mutants worsened the overall survival outcome regardless of the presence or absence of LOH in *APC*. On the other hand, contrary to what we observed at the single-gene level (Figure 3b), LOH of *APC* did not worsen survival in the context of *KRAS* or *PIK3CA* mutations. Similarly, amplifications of *KRAS* did not decrease survival outcomes when considered together with mutations in *PIK3CA* (Figure 5c), and we found an analogous effect for mutant *STK11* with LOH samples in the context of *KRAS* mutants in LUAD samples. In MEL, for *BRAF* and *TERT* the significant decrease in survival between mutants without and with CNA was confirmed also when we considered combined mutations on both genes, and similarly for *ARID1A* and *PTEN* in UCEC.

Copy number-associated metastatic propensity and tropism

Last, we investigated if INCOMMON can be used to determine trends of metastatic propensity and organotropism (Methods). By comparing primary tumour samples from metastatic (10,345 samples) versus non-metastatic (3,372 samples) patients (Supplementary Note), we computed the Odds Ratio (OR) via logistic regression for the likelihood of metastasis associated with mutations having CNA (control group: mutations without CNA; p-values computed via the Wald test corrected for false discovery rate).

Overall, across 3 tumour types, we found 7 statistically significant (adjusted p-value ≤ 0.05) mutant cases (3 whole genes and 3 hotspots; Figure 6a,b) that, with CNAs, had a significantly increased risk of metastasis (OR > 2, $p \leq 0.05$). In one case, instead, the presence of mutations and CNAs determined a significant decrease in risk (OR < 0.5, $p \leq 0.05$). All results are reported in Supplementary Table S5, summarised in Supplementary Figure S12 and available for browsing at https://ncalonaci.shinyapps.io/incommon/.

The most prominent case was that of *E746_A750del* exon 19 deletions of *EGFR* in LUAD (OR = 3.55, p = 0.015, Figure 6a,b), with a more than three-fold metastasis risk increase by

amplification. Also non-hotspot mutations showed a similar effect, albeit with a lower impact (OR = 1.98, p = 0.002). Interestingly, *PIK3CA* mutations with amplification were associated with increased metastatic risk both in BRCA (OR = 2.12, p = 0.0002), particularly for the hotspots *E545K* (OR = 2.47, p = 0.039) and *H1047R* (OR = 2.93, p = 0.0021), and in UCEC (OR = 2.26, p = 0.014), reflecting the decrease in overall survival (Figure 3c and Supplementary Figure S8). In BRCA, also *CDH1* mutations with LOH showed an increase in metastatic propensity (OR = 2.45, p = 0.032).

We then analysed organ-specific patterns of metastasis (Figure 6c,d) from 7,829 metastasis samples. For each gene (whole-genes and hotspots) and tumour type we computed the OR of mutants with CNA to metastasise to a specific organ with respect to mutants without CNAs, using logistic regression and Wald test. All results are reported in Supplementary Table S5, summarised in Supplementary Figure S13 and available for browsing at https://ncalonaci.shinyapps.io/incommon/.

Overall, we found 16 significant associations, 6 of which involving mutations with CNA, and 10 without. Notably, tropism of mutants without CNA tended to target sites physically located close to the spreading tumour, like intra-abdominal tissues, the lymphatic system and bones, or pleura for lung cancer, whereas mutants with CNA showed diffusion to further organs. In contrast, mutants with CNA demonstrated a broader dispersion to distant organs.

The strongest association was for amplified *KRAS* mutations and CNS/Brain metastases in CRC (OR = 7.87, p = 0.049), with a similar tendency, with lower impact, for *APC* mutants with LOH (OR = 2.84, p = 0.11). *KRAS* mutations with amplification were associated with metastatic spread to the liver in PAAD (OR = 2.48, p < 0.0001), particularly for the hotspots *G12D* (OR = 2.45, p < 0.0001) and *G12V* (OR = 3.32, p < 0.0001), and LUAD (OR = 2.84, p = 0.027), reflecting the association of these genotypes to worse survival outcomes. In PAAD, *KRAS* mutations with amplifications were associated with reduced spread to the intra-abdominal area, with respect to non-amplified mutants (OR = 0.59, p = 0.048) and the lymphatic system (OR = 0.16, p = 0.0003). We observed a similar trend for *PTPRD* and *NF1*, for which mutations with LOH in MEL were associated with diffusion to the CNS/Brain (*PTPRD* OR = 5.70, p = 0.027, *NF1* HR = 3.39, p = 0.056) whereas they showed lower rates of diffusion to the lymphatic system with respect to mutations without LOH (*PTPRD* OR = 0.43, p = 0.033, *NF1* HR = 0.44, p = 0.046). Also for *STK11* and bone metastases (OR = 0.40, p = 0.024) and *EGFR* and pleura metastases (OR = 0.55, p = 0.024), in particular for its hotspot mutation *L858R* (OR = 0.16, p = 0.0004), the presence of LOH and amplification, respectively, were associated with decreased tropism from primary LUAD cancers. Conversely, we observed a tendency, albeit over the threshold of significance, to metastasise in the liver for *STK11* (OR = 4.89, p = 0.13) and *KEAP1* (OR = 2.97, p = 0.15) mutants with LOH, and for *EGFR* mutant hotspot *E746_A750del* with amplification (OR = 4.71, p = 0.14), all genotypes that were associated with negative prognosis. In BRCA, *TP53* mutants with LOH were associated with lower rates of skin metastasis (OR = 0.16, p = 0.03), and we found a tendency of amplified *PIK3CA* mutants, which significantly increased the risk to metastasise, to diffuse to the CNS/Brain (OR = 3.48, p = 0.059).

Discussion

We have presented INCOMMON, a computational tool which advances the analysis of clinical cancer genomic data by elucidating the interplay between somatic mutations and CNAs. INCOMMON is particularly effective with targeted panel sequencing from clinical samples, which are routinely not matched with a normal sample. We used our tool to process publicly available targeted data for 62,548 patients and characterised prognostic biomarkers from complex genomic interactions in various cancer types. First, we identified 11 genes and 6 hotspots whose mutations were frequently associated with CNAs in at least one tumour type. We found 66 groups that were prognostic, making INCOMMON classification 24.5% more informative than standard analyses that use just mutations. Overall, we retrieved 20 prognostic alterations missed when one tests for mutations without considering CNAs.

From a functional perspective, we observed an enhanced potential of mutations with co-occurrent CNAs, a phenomenon evident for both oncogenes and TSGs. Previous research has highlighted a pattern where activating mutations of *KRAS* and *EGFR* in NSCLC, and *BRAF* in MEL are often associated with copy-gain events[32]. For *KRAS,* an emblematic oncogene, this pattern was additionally linked to increased aggressiveness[33]. We found analogous patterns, extending the observations concerning *KRAS* to CRC, PAAD and UCEC also highlighting the hotspots where this trend is exacerbated. We observed a similar influence for *PIK3CA* mutations in BRCA, a phenomenon previously documented in the literature[34], extending this finding to UCEC and CRC. When considering TSGs, the accompanying CNAs typically involved CNLOH or LOH events. The effect of such a condition on prognosis was found to be detrimental in our study, as observed for *KEAP1, ARID1A* and *APC* in LUAD, *TERT* in MEL, *CDK12* and *APC* in PRAD and *PIK3R1* in UCEC. Similar to our findings, previous studies have also reported worse outcomes associated with LOH juxtaposed to inactivating mutations of *CDK12* in PRAD[35] and *KEAP1* in LUAD[36], offering initial insights into their biological mechanisms. Despite these hints, a comprehensive understanding of these mechanisms remains elusive.

Recent research has suggested that chromosomal imbalance inherent in aneuploidy can propel tumour heterogeneity and adaptability without the direct influence of specific driver gene mutations[37]. Complementing this perspective, further experiments of unbiased aneuploidy screens in normal human epithelial cells have identified a repeated selection of CNAs linked to cancer in a tissue-specific manner, in the absence of classic driver mutations[38]. The contrast between the direct impact of mutations and the more systemic influence of aneuploidy highlights the complexity of cancer evolution. Moreover, efforts have been recently directed towards modelling the selection dynamics impacting double-hit events involving mutations and CNAs, within both tumour suppressor genes and oncogenes[39,40]. In our study, we bridge these insights by examining the effect of CNAs in the presence of clonal driver mutations in target genes. Our findings suggest that CNAs not only contribute to the genomic landscape shaped by aneuploidy but also interact with mutations in a manner that significantly influences cancer prognosis. The emergence of joined mutations and CNAs as novel factors offers innovative

insights into the prognostic impact of known genomic alterations, and helps *de novo* biomarkers discovery. This also extends to other aspects of tumorigenesis, such as metastatic potential and tropism. In lung cancers, *EGFR* mutations usually associated with favourable outcomes due to the availability of targeted therapies[29,41], presented a contrasting prognostic value when accompanied by CNAs. The concurrent presence of amplifications indicated a higher tendency to metastasise, with a distinct pattern of diffusion toward the liver, and a decreased overall survival. These observations potentially reflect reduced treatment sensitivity and, therefore, also a predictive value. Similarly in BRCA, concurrent amplifications of mutant *PIK3CA* increased the metastatic propensity and the rate of brain metastases, offering possible explanations for the negative prognostic impact of these conditions. Among the novelties, we observed tumour-specific biomarkers like *CDH1* mutations with LOH in BRCA, increasing the metastatic propensity, and amplified *PIK3CA* mutations in CRC leading to decreased survival outcomes.

Conclusion

The practical application of INCOMMON in clinical settings could significantly enhance the management of cancer patients, informing personalised therapeutic strategies by identifying genomic signatures associated with metastatic risk, organ tropism, and response to therapies. Our approach therefore carries profound clinical significance to enhance precision cancer medicine. By revealing the interplay of complex genetic alterations, INCOMMON enables patient stratifications that can potentially guide more tailored treatments and unravel mechanisms of intrinsic or acquired resistance to target therapies. Indeed, the co-existence of an actionable mutation with CNAs might identify patients with oncogene-addicted tumours as characterised by lower sensitivity to matched inhibitors.

While we use INCOMMON to analyse the largest clinical cohorts available to date, our study acknowledges intrinsic limitations linked to data and methodologies. The adopted cohorts present mostly unmatched primary or metastatic samples, whereas longitudinal data would better capture trends of metastatic tropism. Moreover, the lack of survival data available for the GENIE-DFCI cohort prevents the validation of our groups derived from the MSK-MET cohort. At the methodological level our tool is not designed to process longitudinal data, whereas a better classification could be achieved once longitudinal data are popular and our algorithms updated. Moreover, our model neglects events of copy numbers independent of mutations, such as amplifications of WT oncogenes. While these are not ubiquitous across primary tumours (Supplementary Figure 14) they might have an important role in determining altered pathway functions. However, to detect these events one requires depth-of-sequencing data through the genome[42], an information that we did not find available in targeted panels from our clinical cohorts.

In summary, this research highlights key aspects of cancer genomics and introduces INCOMMON as a transformative tool for interpreting clinical targeted sequencing data. Our insights, encompassing both established and novel biomarkers, lay a robust foundation for

future investigations and clinical applications, contributing to advanced personalised cancer therapy and our understanding of the unknown biology behind these genomic alterations. The clinical relevance of our findings suggests a promising potential to improve patient care.

Methods

1 Bayesian inference using INCOMMON

INCOMMON is a mixture model[15] to classify single mutations, in terms of their copy number and multiplicity status. The read counts $(n,\ N)$ of a mutation, characterised by the total number of reads covering the mutation site N and the number of reads carrying the variant are assumed to be distributed according to a Beta-Binomial mixture

$$(1) \qquad p\left(n \mid N, \pi, \rho\right) = \sum_{\eta} \psi_{\eta} \text{Beta-Binomial}\left(n \mid N, f_{\eta}\left(\pi\right), \rho\right)$$

.

It is necessary to know a priori the purity of the sample π, for instance from sequencing or pathology assessment. The mixture components (classes), weighted by the mixing proportions ψ_{η} (XXX sum to 1) are identified by triples $\eta = (n_A, n_B, m)$, where n_A and n_B are the major and minor allele copies so that the tumour genome ploidy is $p = n_A + n_B$ and the mutation is present in $m \leq max(n_A, n_B)$ copies (multiplicity). The over-dispersion parameter ρ models the sequencing noise, and $f_{\eta}(\pi)$ is the probability of success of the binomial read counting process. INCOMMON supports K = 6 classes, identified by triples $\eta = (n_A, n_B, m)$. Among these classes, 5 determine clonal mutations (Tier-1) with the most common copy number configurations:

- Loss of heterozygosity in monosomy LOH: $(n_A = 1,\ n_B = 0, m = 1)$
- Copy-neutral loss of heterozygosity CNLOH: $(n_A = 2,\ n_B = 0, m = 2)$
- Amplification AM: $(n_A = 3,\ n_B = 2, m = 2)$ or $(n_A = 4,\ n_B = 2, m = 2)$
- Heterozygous mutant diploid HMD: $(n_A = 2,\ n_B = 1, m = 1)$.

The sixth configuration identifies Tier-2 mutations, which contain subclonal (present less than 100% of tumour cells) or clonal mutations with high-ploidy and/or low-multiplicity $(m/p \leq 1/3)$.

The assignment of a mutation to one of the K classes η is modelled by the latent random variable ***z*** that has a 1-of-K representation, i.e. a single element z_{η} of the vector is equal to 1 and all the other elements are equal to 0, $z_{\eta} \in \{0, 1\}$ and $\sum_{\eta} z_{\eta} = 1$.

The posterior probability of the assignment $z_\eta = 1$ conditioned on the observed read counts (n, N) is, by Bayes' theorem

$$(2) \qquad p(z_\eta = 1 \mid n, N, \pi, \rho) = \frac{\psi_\eta \text{Beta-Binomial}(n \mid N, f_\eta(\pi), \rho)}{\sum_\eta \psi_\eta \text{Beta-Binomial}(n \mid N, f_\eta(\pi), \rho)}.$$

Given that the copy number and multiplicity of a mutation is $\eta = (n_A, n_B, m)$, the probability of collecting a read with the variant is equal to its expected variant allele frequency, which is known to be

$$(3) \qquad f_\eta(\pi) = \frac{m\pi}{2(1-\pi) + \pi(n_A + n_B)}.$$

The mixing proportions ψ_η can be specified in terms of the prior probability of the assignment $z_\eta = 1$ such that

$$(4) \qquad p(z_\eta = 1) = \psi_\eta$$

Consequently, they must satisfy $0 \leq \psi_\eta \leq 1$ and $\sum_\eta \psi_\eta = 1$.

1.1 Empirical prior from whole-genome sequencing

In general, it can be very hard to infer the copy number and mutation multiplicity of a mutation, especially when the sequencing depth is low. In order to improve our classification and inject biological prior knowledge into the model, we relied on empirical priors. We thus built ψ_η by considering the frequency distribution of the supported copy number configurations in the large-scale (2778 samples) PCAWG cohort of primary tumours[16]. This whole genome sequencing dataset comprises driver gene and tumour type annotations, as well as copy number calls and phased mutations generated by a multi-cohort effort and recently validated computationally[17].

We empirically estimated the prior distribution over the K classes supported by INCOMMON specific to genes and tumour types, by estimating the frequency at which a gene is mutant in a tumour type from PCAWG. We selected only cases with a reasonable number of observations: if a gene was mutated in at least 5% of samples from a tumour type, and at least in 20 samples, we built a tumour-specific prior, otherwise we pooled from all tumour types a pan-cancer prior.

To build the prior we used the copy number calls validated by quality control and converted to INCOMMON classes. To deal with missing observations of specific INCOMMON classes, we

first initialised the prior to a reasonable configuration. Using the heterozygous mutant diploid class (HMD) as the baseline, we assigned this class the maximum probability 25%, and 12.5% to all the other classes. Then we replaced the probability of any observed class with its empirical frequency in the dataset, and normalised the final distribution to ensure that $\sum_{\eta} \psi_{\eta} = 1$.

The prior parameters estimated from PCAWG are available as Supplementary Table S1.

1.2 Classification

The posterior probability of the assignment variable ***z*** is proportional to

$$(5) \qquad p\left(\mathbf{z} \mid n, N, \pi, \rho\right) \propto \prod_{\eta} \left[\psi_{\eta} \text{Beta-Binomial}\left(n, N, f_{\eta}\left(\pi\right), \rho\right)\right]^{z_{\eta}}$$

INCOMMON classifies mutations by assigning it to the class with the maximum a posteriori (MAP) probability, thus estimating

$$(6) \qquad \hat{z}_{\eta} = \arg\max_{\eta} \{\psi_{\eta} \text{Beta-Binomial}\left(n, N, f_{\eta}\left(\pi\right), \rho\right)\}$$

In other words, INCOMMON estimates the most likely class given both the data and prior knowledge from a higher-resolution (WGS) validated dataset. The probabilistic model has the advantage of returning a probability of assignment to each one of the tested classes, a piece of information that we use to derive a confidence metric for the prediction.

1.3 Prediction uncertainty

INCOMMON quantifies the uncertainty associated with each classification. This is particularly valuable for cases in which the read counts do not strongly and uniquely indicate a specific copy number configuration, which tends to occur especially for tumour samples with low purity or low sequencing depth.

We compute uncertainty by examining the variability in the posterior probabilities assigned to different classes for each mutation. High variability indicates that the mutation could reasonably belong to multiple classes, while low variability suggests a more confident classification. INCOMMON adopts the concept of entropy to quantitatively measure the uncertainty associated with the classification of a data point. The entropy associated with read counts (n, N) is computed using the posterior probabilities

$$(7) \qquad H\left(n, N, \pi, \rho\right) = -\sum_{\eta} p\left(z_{\eta} = 1 \mid n, N, \pi, \rho\right) \cdot \log_2 p\left(z_{\eta} = 1 \mid n, N, \pi, \rho\right)$$

A higher entropy value indicates greater uncertainty, as the probabilities are spread out across multiple classes. Conversely, lower entropy implies more certain classification, with one class having a significantly higher probability. This is inspired by popular clustering methodologies that seek to induce sparsity in their inferences [19,43,44].

Data and software availability

Software availability

INCOMMON is an open-source R package at https://caravagnalab.github.io/INCOMMON. The tool webpage contains RMarkdown vignettes to run analyses, visualise inputs and outputs, and parametrise the tool.

All analyses presented in this paper, from data gathering and analysis results, are available in Zenodo at https://zenodo.org/records/10927218.

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/.

Data availability

PCAWG and TCGA data used in this paper (mutations with matched validated CNAs) are available at https://doi.org/10.5281/zenodo.6410935, following [42].

MSK MetTropism data has been downloaded from cohort “msk_met_2021” at the CBioPortal, following link https://www.cbioportal.org/study/summary?id=msk_met_2021.

AACR GENIE-DFCI data has been downloaded at https://www.synapse.org/# through access codes syn50678641, syn50678411, syn50678410, syn50678644, syn50678531, syn50678642, syn50678530, syn50678532, syn50678295, syn50678640, syn50678653, syn50678296.

Supplementary Tables

- Supplementary table S1: Empirical prior distributions obtained from PCAWG.
- Supplementary table S2: Summary statistics and enrichment of INCOMMON classes.
- Supplementary table S3: Survival Analysis (multivariate Cox regression) with INCOMMON classes, single gene-level.
- Supplementary table S4: Survival Analysis (multivariate Cox regression) with INCOMMON classes, two genes-level.
- Supplementary table S5: Metastatic propensity with INCOMMON classes.
- Supplementary table S6: Metastatic tropism with INCOMMON classes.

Contributions

NC and GC conceptualised and formalised INCOMMON, which was implemented by NC with support from SM and SS. NC, BR, SS and MMS gathered real data, which was analysed by NC, EK, 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. We wish to thank Area Science Park for computational support through the ORFEO (Open Research Facility for Epigenomics and Other) data centre.

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
3. Sansregret, L. & Swanton, C. The Role of Aneuploidy in Cancer Evolution. *Cold Spring Harb. Perspect. Med.* **7**, (2017).
4. Lukow, D. A. & Sheltzer, J. M. Chromosomal instability and aneuploidy as causes of cancer drug resistance. *Trends Cancer Res.* **8**, 43–53 (2022).
5. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol. Oncol.* **6**, 140–146 (2012).
6. Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S. & Swisher, E. Cancer biomarkers: a systems approach. *Nat. Biotechnol.* **24**, 905–908 (2006).
7. Brooks, J. D. Translational genomics: the challenge of developing cancer biomarkers. *Genome Res.* **22**, 183–187 (2012).
8. Committee on Developing Biomarker-Based Tools for Cancer Screening, Diagnosis, and Treatment & Institute of Medicine. *Cancer Biomarkers: The Promises and Challenges of Improving Detection and Treatment*. (National Academies Press, 2007).
9. Mulero-Navarro, S. & Esteller, M. Epigenetic biomarkers for human cancer: the time is now. *Crit. Rev. Oncol. Hematol.* **68**, 1–11 (2008).
10. Costa-Pinheiro, P., Montezuma, D., Henrique, R. & Jerónimo, C. Diagnostic and prognostic epigenetic biomarkers

in cancer. *Epigenomics* **7**, 1003–1015 (2015).

11. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).
12. 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. *J. Mol. Diagn.* **17**, 251–264 (2015).
13. Ptashkin, R. N. *et al.* Abstract 3409: MSK-IMPACT Heme: Validation and clinical experience of a comprehensive molecular profiling platform for hematologic malignancies. *Cancer Res.* **79**, 3409–3409 (2019).
14. Lee, E. Y. H. P. & Muller, W. J. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.* **2**, a003236 (2010).
15. Bishop, C. M. Pattern recognition. *Mach. Learn.* **128**, (2006).
16. ICGC/TCGA PCAWG Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
17. Antonello, A. *et al.* Computational validation of clonal and subclonal copy number alterations from bulk tumour sequencing. *bioRxiv* 2021.02.13.429885 (2023) doi:10.1101/2021.02.13.429885.
18. Cross, W. *et al.* Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer. *biorXiv* 2020.03.26.007138 (2020) doi:10.1101/2020.03.26.007138.
19. Caravagna, G. *et al.* Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.* **52**, 898–907 (2020).
20. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
21. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
22. Donehower, L. A. *et al.* Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. *Cell Rep.* **28**, 1370–1384.e5 (2019).
23. Nguyen, B. *et al.* Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* **185**, 563–575.e11 (2022).
24. Kaur, P., Porras, T. B., Ring, A., Carpten, J. D. & Lang, J. E. Comparison of TCGA and GENIE genomic datasets for the detection of clinically actionable alterations in breast cancer. *Sci. Rep.* **9**, 1–15 (2019).
25. Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **72**, 2457–2467 (2012).

26. Zhu, G., Pei, L., Xia, H., Tang, Q. & Bi, F. Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. *Mol. Cancer* **20**, 143 (2021).
27. Timar, J. & Kashofer, K. Molecular epidemiology and diagnostics of KRAS mutations in human cancer. *Cancer Metastasis Rev.* **39**, 1029–1038 (2020).
28. Zhang, Y. *et al.* A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* **31**, 820–832.e3 (2017).
29. Peng, D. *et al.* Effect of EGFR amplification on the prognosis of EGFR-mutated advanced non-small-cell lung cancer patients: a prospective observational study. *BMC Cancer* **22**, 1323 (2022).
30. Heide, T. *et al.* The co-evolution of the genome and epigenome in colorectal cancer. *Nature* **611**, 733–743 (2022).
31. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
32. Modrek, B. *et al.* Oncogenic activating mutations are associated with local copy gain. *Mol. Cancer Res.* **7**, 1244–1252 (2009).
33. 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. *Transl Lung Cancer Res* **10**, 826–838 (2021).
34. Migliaccio, I. *et al.* PIK3CA co-occurring mutations and copy-number gain in hormone receptor positive and HER2 negative breast cancer. *NPJ Breast Cancer* **8**, 24 (2022).
35. Antonarakis, E. S. *et al.* CDK12-Altered Prostate Cancer: Clinical Features and Therapeutic Outcomes to Standard Systemic Therapies, Poly (ADP-Ribose) Polymerase Inhibitors, and PD-1 Inhibitors. *JCO Precis Oncol* **4**, 370–381 (2020).
36. Scalera, S. *et al.* Clonal KEAP1 mutations with loss of heterozygosity share reduced immunotherapy efficacy and low immune cell infiltration in lung adenocarcinoma. *Ann. Oncol.* **34**, 275–288 (2023).
37. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
38. Watson, E. V. *et al.* Chromosome evolution screens recapitulate tissue-specific tumor aneuploidy patterns. *Nat. Genet.* (2024) doi:10.1038/s41588-024-01665-2.
39. Bielski, C. M. *et al.* Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell* **34**, 852–862.e4 (2018).
40. Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver events of human cancer. *bioRxiv* (2023) doi:10.1101/2023.08.05.552104.

41. Nyati, M. K., Morgan, M. A., Feng, F. Y. & Lawrence, T. S. Integration of EGFR inhibitors with radiochemotherapy. *Nat. Rev. Cancer* **6**, 876–885 (2006).
42. Antonello, A. *et al.* Computational validation of clonal and subclonal copy number alterations from bulk tumor sequencing using CNAqc. *Genome Biol.* **25**, 38 (2024).
43. Côme, E., Jouvin, N., Latouche, P. & Bouveyron, C. Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. *Adv. Data Anal. Classif.* **15**, 957–986 (2021).
44. Caravagna, G., Sanguinetti, G., Graham, T. A. & Sottoriva, A. The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. *BMC Bioinformatics* **21**, 531 (2020).

Main Figures

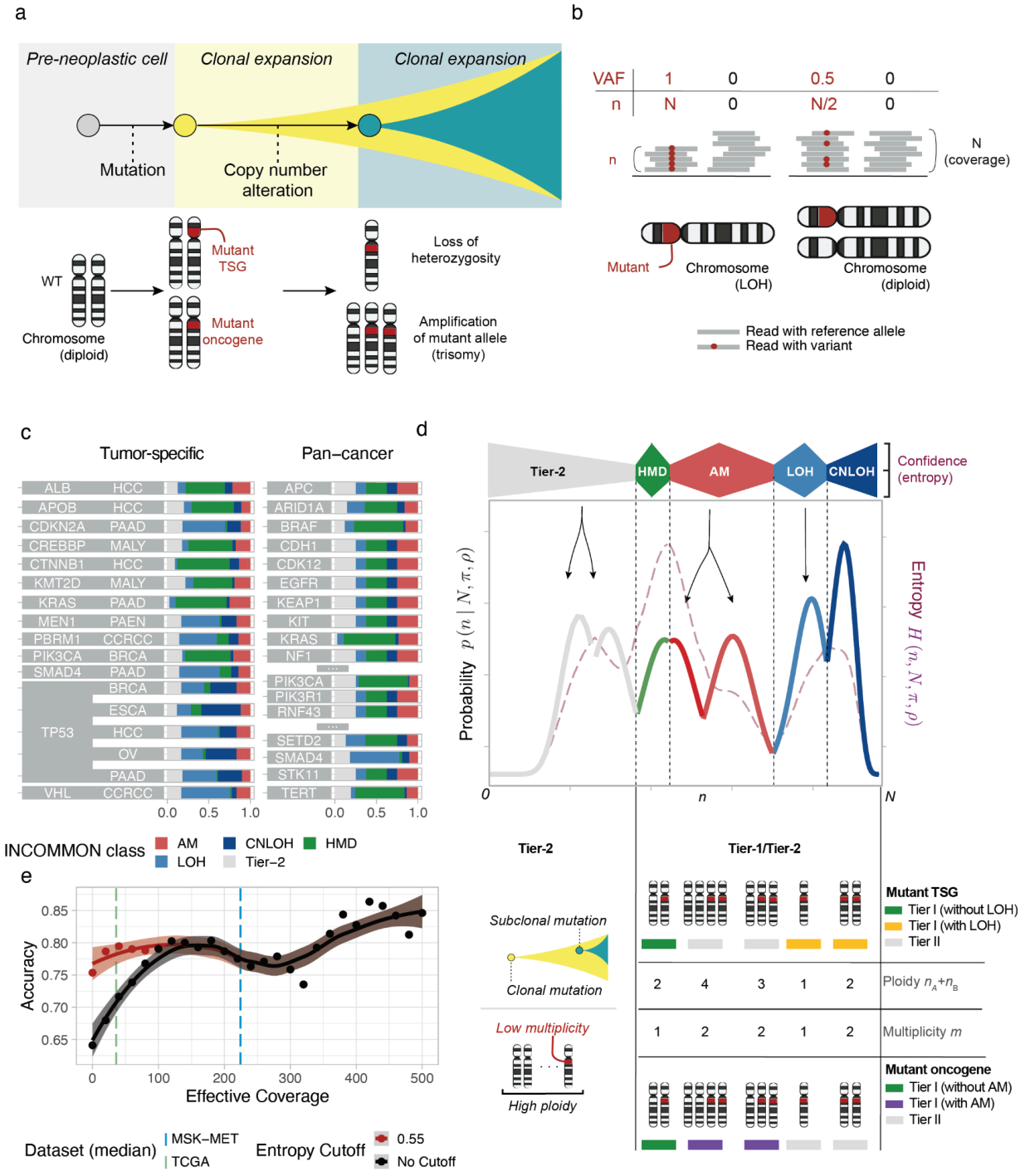

Figure 1. Copy-number and mutation multiplicity inference with INCOMMON. a. Clonal expansions from a normal cell that acquires first a mutation in a tumour suppressor gene (TSG) or an oncogene, then a copy number alteration (CNA) on the mutant locus, pictured as a loss of the wild-type (WT) allele or an amplification of the mutant one. **b.** INCOMMON processes read counts that define the variant allele frequency (VAF) of a mutation (the total number of reads $N \gg 1$, and the number $n \leq N$ of mutant

reads). For example, a clonal mutation with LOH has $n = N$ and VAF $= 1$, whereas VAF $= 0.5$ if the mutation harbours a diploid heterozygous site. **c.** Pan-cancer or tumour-specific Bayesian classification prior obtained from validated copy-number calls in PCAWG[16]. **d.** INCOMMON identifies Tier-1 clonal heterozygous mutant diploid (HMD) states, deletion of the WT allele in monosomy (LOH) or copy-neutral (CNLOH) states, amplifications (AM) of the mutant allele in trisomy and tetrasomy states. It also detects Tier-2 mutations compatible either with subclonal frequencies or with high-ploidy low-multiplicity states. Genotypes are interpreted in terms of mutant TSGs with/without LOH and oncogenes with/without amplification. The uncertainty of each classification is estimated via the entropy. **e.** We measured the performance of INCOMMON using ground truth data from TCGA[20], determining the accuracy of copy number and multiplicity estimates, which can be controlled through a cutoff on the entropy not needed for high effective coverage (sequencing depth times sample purity).

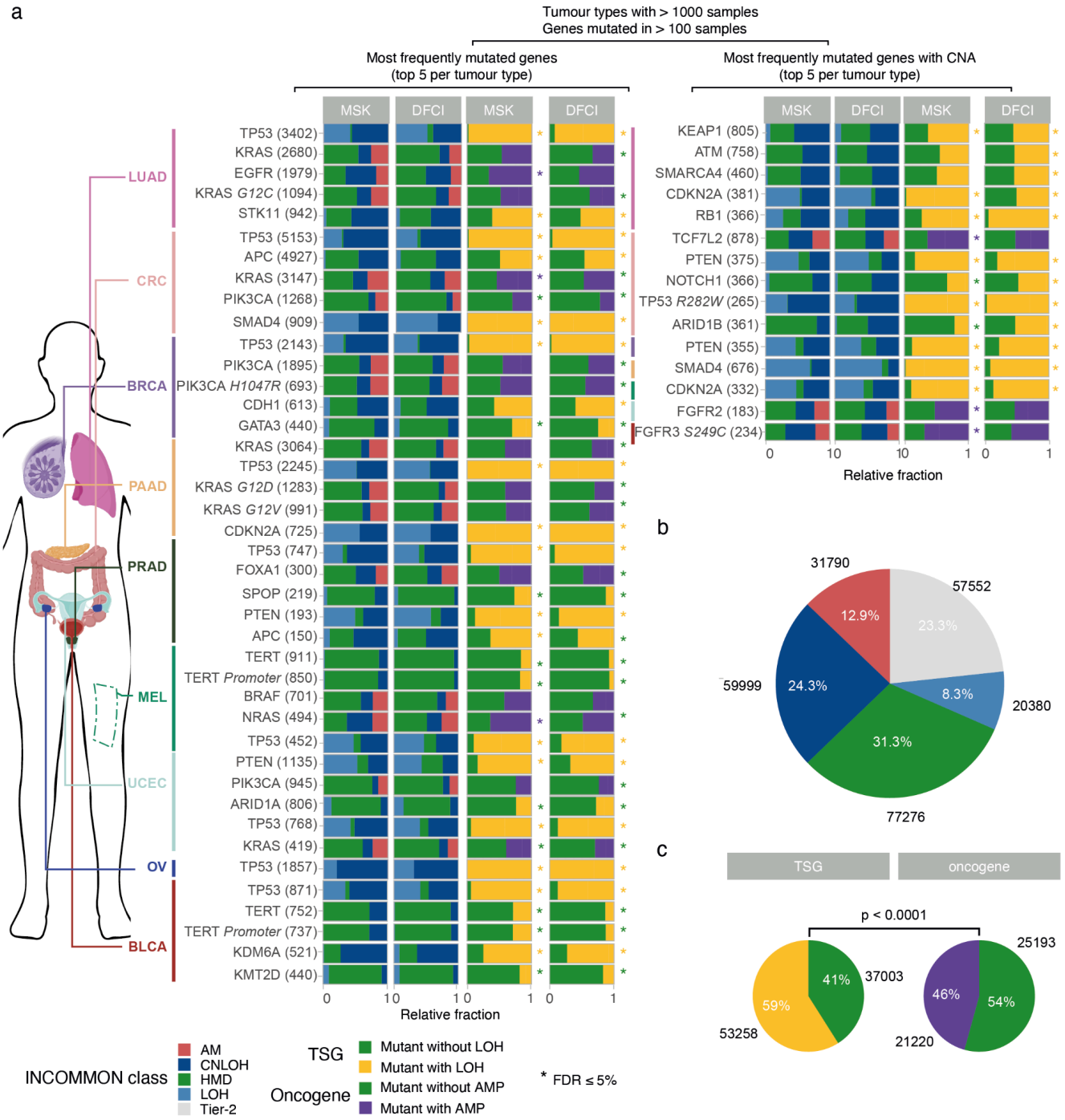

Figure 2. INCOMMON classification of mutations in 61,690 clinical samples. **a.** Classification of gene-level or hotspot mutations for the MSK-MET (MSK) and GENIE-DFCI (DFCI) cohorts. The map reports results for the most common tumour type (on the left the most frequently mutated entries, on the right the ones with the strongest statistical association with copy numbers). For each entry, we report a Chi-squared test for the enrichment of mutants without CNAs, against mutants with CNAs (Mutant+CNA). Genes are also colour-coded to denote suppressors (TSG) and oncogenes (ONC). **b.** Gene-level and hotspot mutations were included by filtering tumour types with at least 1000 samples and considering

gene mutants in at least 100 samples in both cohorts. **c,d.** Pan-cancer distribution of INCOMMON classes and genotypes across both cohorts.

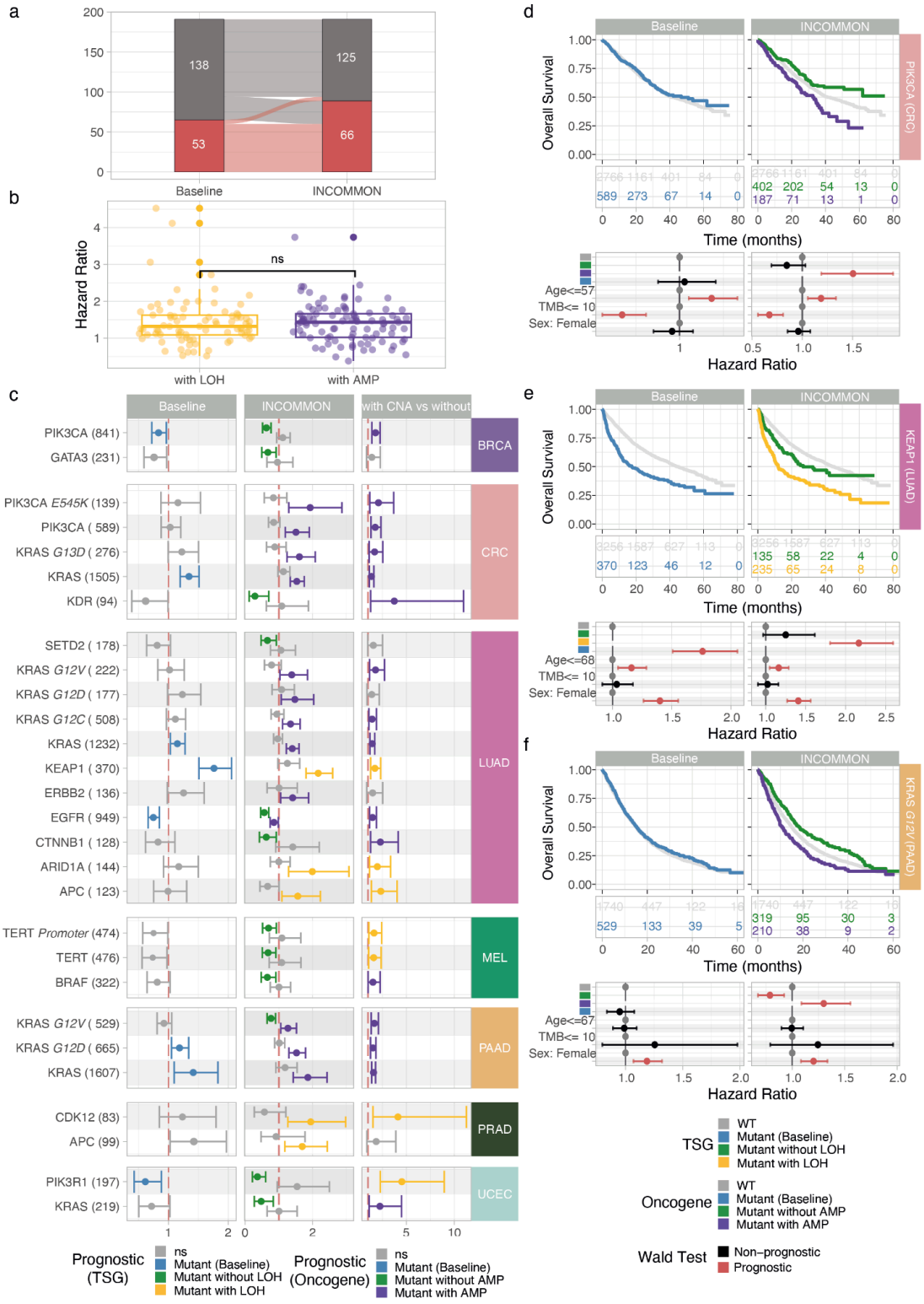

Figure 3. The INCOMMON classification increases the number of prognostic markers. a. Increase in the number of prognostic versus non-prognostic groups between the baseline survival analysis (mutants irrespective of CNAs vs WT) and the one enabled by INCOMMON (mutant with or without CNAs vs WT). Prognostic power is estimated by computing hazard ratios (HR) from multivariate Cox regression,

accounting for sex, age and tumour mutational burden (TMB), and using the Wald test for statistical significance. **b.** The overall trend of HR is similar for TSGs and oncogenes with CNAs (worse prognosis) vs without. **c.** Results per gene or hotspot mutation, split by tumour type. HR values with 95% confidence intervals are reported for every group. **d-f.** Kaplan-Meier curves annotated with numbers at risk and HR for *PIK3CA* in colorectal cancers, *KEAP1* in lung adenocarcinoma and *KRAS G12V* in pancreatic adenocarcinomas, evidence the better resolution provided by INCOMMON classification in stratifying patients. Forest plots show the contribution of each covariate in the model.

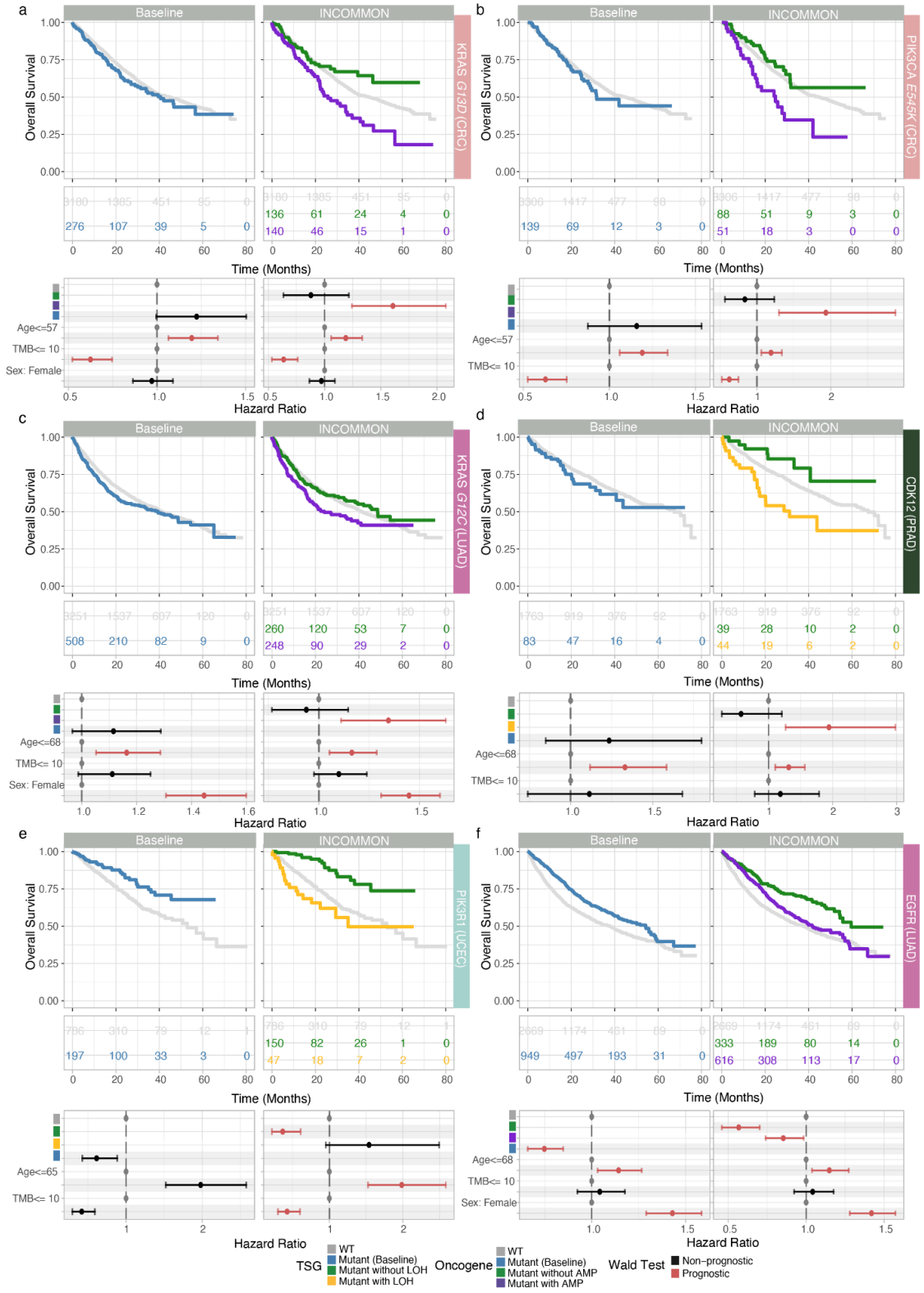

Figure 4. Survival analysis for INCOMMON groups. Comparison (as in Figure 3c-e) of the baseline analysis (mutant with/without CNAs vs WT) and the one enabled by INCOMMON (mutant with or without

CNAs vs WT). **a,b.** Results for *KRAS G13D* and *PIK3CA E545K* in colorectal cancers, **c.** Results for *KRAS G12C* in lung adenocarcinoma. **d.** Results for *CDK12* in prostate adenocarcinoma. **e.** Results for *PIK3R1* in lung adenocarcinoma. **f.** Results for *EGFR* in lung adenocarcinoma. Other survival curves and results of Cox regression are in Supplementary Figure S9.

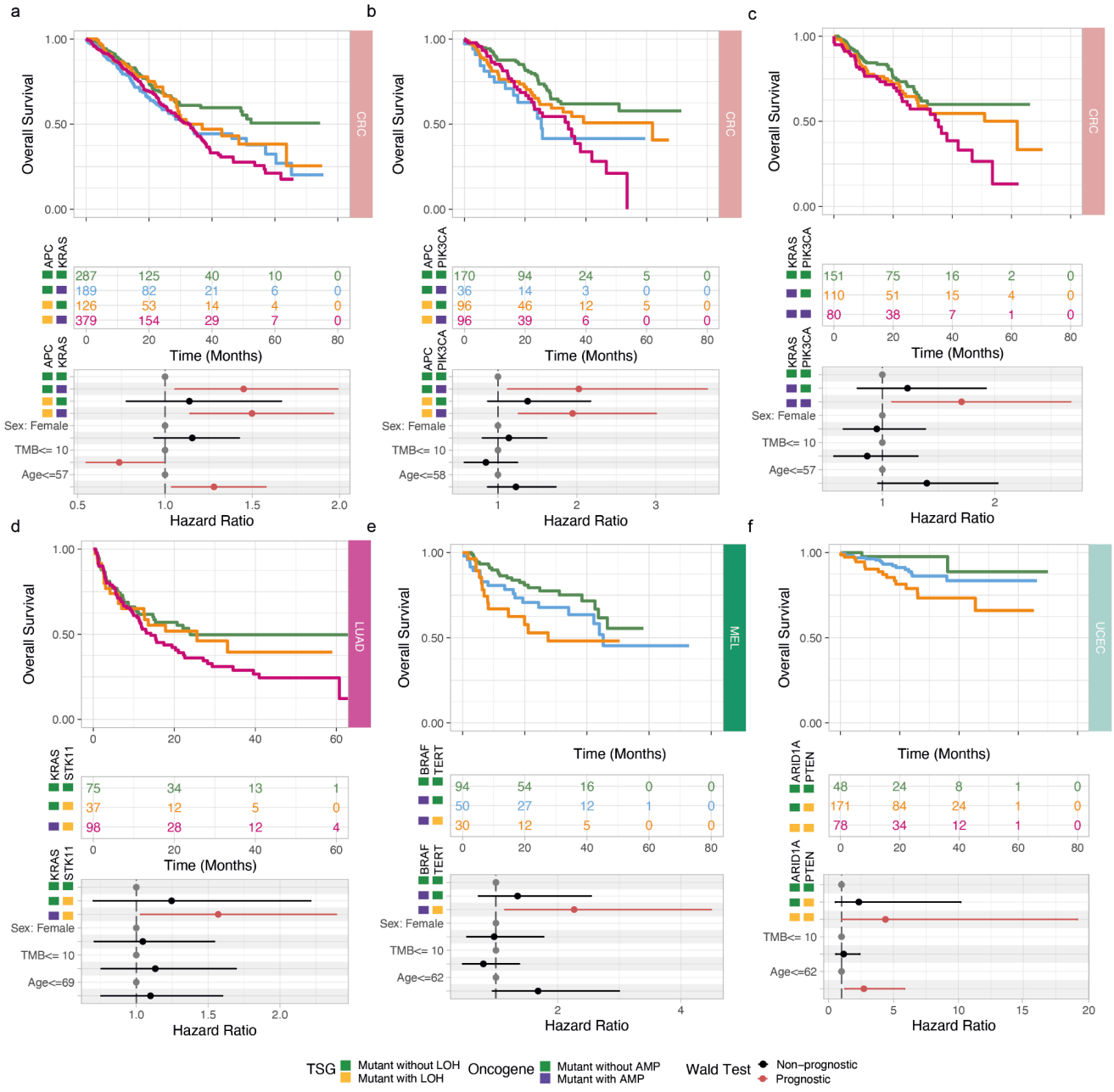

Figure 5. Survival outcome for paired genes classifications with INCOMMON. a,b,c. Paired survival analysis for *APC* and *KRAS*, *APC* and *PIK3CA* and *KRAS* and *PIK3CA* in colorectal cancers. **d.** Paired analysis of lung adenocarcinoma samples for *KRAS* and *STK11* mutations. The colours in the risk table reflect the group for the gene. In this analysis, the baseline group contains *TP53*-mutants without LOH and

KRAS-mutants without amplification **e.** Paired analysis of melanoma samples for *BRAF* and *TERT* mutations. **f.** Paired analysis of endometrial cancer samples for *ARID1A* and *PTEN* mutations.

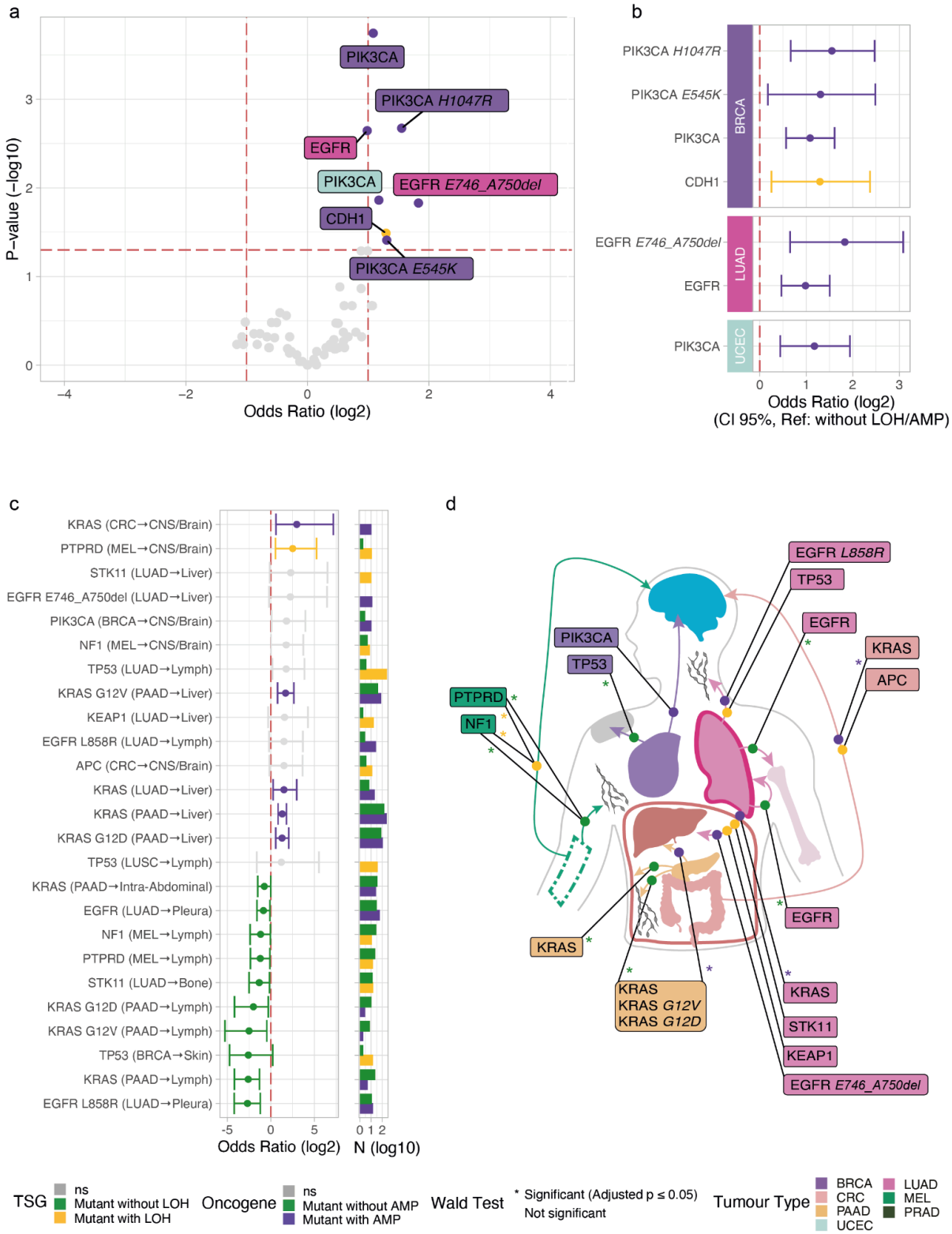

Figure 6. Metastatic propensity and organotropism analysis of 18,000 samples. a. Odds ratio (OR) of primary tumours with mutant TSGs with LOH vs without or mutant oncogenes with amplification vs without, of giving rise to metastasis. Reported is the adjusted p-value (log10 scale) against the OR (log2 scale). Genes and specific hotspots significant are labelled. **b.** Odds ratios with 95% confidence intervals for the significant cases. **c.** Combinations of mutations and CNAs that affect organotropism in metastatic samples with OR under the 10th and over the 90th percentile. The number of metastatic samples per group and metastatic site is also reported (right panel). Significant cases are highlighted **d.** Organotropic patterns associated with the combination of mutations and CNAs, represented by arrows starting from the primary site and pointing to the metastatic site. Thicker arrows represent patterns with at least one significant association, indicated by an asterisk.