Passenger mutations in ~2500 cancer genomes: Overall functional impact & its consequences

To a first approximation, all clinically significant consequences of genomic variants in cancer are mediated through their functional impact, such as changes in gene expression or gene activity. Certain key alterations in tumor genome, often identified through the detection of strong signals of positive selection on individual variants, have been shown to play pivotal role in tumor progression. Although a typical tumor has thousands of genomic variants, very few of these (~3/tumor[1]) are thought to drive tumor growth. The remaining variants, often termed passengers, represent the overwhelming majority of the variants in cancer genomes, and their functional consequences are poorly understood. Furthermore, the bulk of these passengers fall within noncoding regions of the genome, making these the main product of whole-genome sequencing of tumors. Recent studies have proposed that, among variants that have not been found to be driver variants (i.e. *nominal* passenger variants), may impact tumor cell biology along a range of dimensions and weakly affect tumor cell fitness by promoting or inhibiting tumor growth (*latent driver variants* [2,3]) (*deleterious passengers*[4]) (**Fig 1**).

Previous studies have extensively focused on characterizing variants occupying coding regions of various cancer genomes. However, the exhaustive pan-cancer analysis of whole genome (PCAWG) variant dataset, which comprises pan-cancer variant calls from ~2500 uniformly processed whole cancer genomes, gives us an unparalleled opportunity to investigate the overall functional burdening of different non-coding genomic elements. Given that the majority of cancer variants lie in non-coding regions, this variant dataset serves as a substantially more informative resource than the many existing datasets focused on exomes. In addition, it also contains a full spectrum of variants, including copy number variants (CNVs) and large structural variants (SVs) in addition to SNVs and INDELs.

In this work, we explore the functional landscape of passenger variants in various cancer cohorts by leveraging extensive pan-cancer variant calls PCAWG. More specifically, we build on and apply existing tools to annotate and score the predicted functional impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset. This systematic annotation effort generates a comprehensive annotation compendium of PCAWG variants, which can serve as a useful resource for the wider cancer genomics community. Furthermore, we integrate annotation and impact score of each variant to quantify the overall burdening of various elements in cancer genomes. We observed that disruption of genetic regulatory elements in the noncoding genome correlates with altered gene expression. Moreover, various mutation processes have different impacts on the regulatory elements, as elucidated by our signature analysis. Furthermore, we also show how overall functional burdening of various genomic elements correlate with age at cancer diagnosis, patient survival time, and tumor clonality. Finally, we observe

suggestive evidences, which are consistent with the notion that aggregated subsets of functionally impactful passenger variants confer weak fitness effects.

In order to substantiate the presence of impactful passenger SNVs and their role in cancer progression, we surveyed the functional impact distribution of somatic variants in different cancer genomes. The functional impact distribution varies among different cancer types and different genomic elements. For instance, impact score distributions of non-coding variants in different cancer genomes indicate three distinct peaks. The upper and the lower extremes of this distribution correspond to traditional definitions of high-impact putative driver variants and low impact neutral passengers, respectively. In contrast, the middle peak in the intermediate functional impact regime corresponds to what we term *impactful passengers* (**Fig 2a**).

According to a simple random expectation, one would assume that the overall functional burdening in a cancer genome will be uniformly distributed across different functional elements and among different gene categories. In contrast, we observe that the functional burdening in certain cancers is concentrated in particular gene categories. In particular, we show that impactful variants tend to occur in essential genes more often compared to low impact variants (**Fig 2b**). Conversely, low impact passengers constitute larger fractions of variants influencing non-essential genes. Similarly, we examined essential genes affected by loss of function SNVs in cancer patients. We found that cancer related somatic loss of function SNVs disproportionately affect essential genes, while germline loss of function variants rarely affects essential genes (**Fig 2b**). This observation matches an expectation of positive selection for cancer related mutation of essential genes, and negative selection for germline loss of function mutation of essential genes.

Furthermore, in the random model, we would expect that the fraction of impactful variant will remain constant as one accumulate large amount mutation in certain cancer sample. In contrast, we observe that as we acquire more SNVs in cancer, the fraction of impactful mutations decreases suggesting that the earlier variants tend to be impactful and driving the cancer whereas the later are more likely to be random, i.e. collateral damage. This trend is particularly strong and in CNS medulloblastoma ($p < 4e-8$, Bonferroni's correction), lung adenocarcinoma ($p<3e-4$, Bonferroni's correction), and a few other cancers (**Fig 2c**).

One might further expect that passenger variants will be contributing functional burden uniformly across the genome. Consequently, we comprehensively analyzed the overall mutational burdening of various genomic elements, including TF (transcription factor) binding motifs in various cancer genomes. The presence of a variant within a TF binding site(TFBS) can lead to either the creation or destruction of binding motifs (gain or loss of function). In both cases, we observe significant differential burdening of TFBS among different cancer cohorts. For instance, we observe significant enrichment of high impact

variants creating new motifs in various TFs such as GATA, PRRX2 and SOX10 (**Fig 3b**) across major cancer types analyzed in this study. Similarly, high impact variants influencing gene expression by breaking TF motifs, were highly enriched in TFs such as IRF, POU2F2, NR3C1and STAT (**Fig 3f**) in a majority of cohorts. This selective enrichment or depletion suggests distinct alteration profiles associated with different components of regulatory networks in various cancers. For instance, a strong motif creation bias event among ETS family TFs was detected in the TERT promoter region in various cancer cohorts including glioblastoma, medulloblastoma, bladder transitional cell carcinoma, thyroid adenocarcinoma and oligoastrocytoma. Furthermore, enrichment of SNVs in selective TF motifs leading to gain and break events in promoter significantly perturb the downstream gene expression (**Fig 3g**). For instance, in lung adenocarcinoma, we found three TFBSs gain events (ZBTB14, E2F and HNF4) significantly increase downstream expression level ($p<5e-7$, 3e-6 and 2e-4 respectively) (**Fig 3c**). Similarly, ETS family transcription factor at the regulatory region of IRF and PSIP1 gene display a strong motif creation bias and a significant change in their expression (with p-value IRF=0.001 and p-value PSIP1=0.019).

The disproportionate burdening of certain TFs in different cancers can be further related to the underlying mutational spectrum (ie signature) of variants influencing their binding sites. For instance, mutation spectrum of motif breaking events observed in SP1 TF binding sites (TFBS) suggest major contribution from C>T and C>A mutation (**Fig 4b**). In contrast, motif breaking events at TFBS of HDAC2 and EWSR1 have relatively uniform mutation spectrum profiles. Similarly, comparing signature composition of low and high impact SNVs in certain cancer-cohort can help us to distinguish between mutational processes that generate distinct impact classes of variants. For instance, we observed distinct signature distributions for the low and high impact non-coding passengers in the kidney-RCC cohort. While the majority of passengers can be explained by signature 5, high impact passengers have a higher fraction of SNVs explained by signature 4 (**Fig4a**). Moreover, we observed cancers showing microsatellite instability (MSI) due to failure of DNA mismatch repair, have higher percentage of high impact non-coding passengers (**Fig4c**). Our findings suggest various mutational processes shape and disproportionally burden cancer genomes.

In addition to SNVs, large structural variations (SVs) are considered to play important role in cancer progression. Thus we annotated and evaluated the impact of large SVs in the entire PCAWG cohort. Simplistically, we would expect majority of SVs to be distributed across the genome regardless of their extent of overlap with functional elements of the genome. However, our annotation analysis of somatic SVs in PCAWG portrays a different picture. We observe significant enrichment of large engulfing somatic deletions as well as duplications among pseudogenes, coding regions, UTRs and TF peak regions. Moreover, engulfing SVs tend to have higher enrichment value compared to partially overlapping SVs. The observed enrichment bias of SVs toward certain regions of the genome as well as

the extent of their overlap suggest that selection processes play a key in role in emergence of somatic SVs. We quantified the effect of these selection processes by evaluating functional impact of these large deletions and duplications across various cancer-types. The functional impact score distribution of SVs for different cancer-types indicate that meta tumor cohorts such as CNS, glioma and sarcoma tend to harbor higher impact large deletions and duplications compared to others. In addition, gene-centric analysis on the pan-cancer level reveals that CDKN2A and TEKT2 genes have the largest observed enrichment of high impact deletions and duplications, respectively.

Additionally, we sought to examine whether impactful passengers might be associated with tumor initiation and progression. Therefore, we correlated patient impactful somatic mutation burden with patient survival. We performed survival analysis to see if somatic impact burden –the ranked sum of the impact scores of coding and noncoding variants – predicted patient survival within individual cancer subtypes. These correlations varied substantially in different cancer types. For instance, we observed that somatic mutation burden predicted substantially earlier death in chronic lymphocytic leukemia (CLL) and substantially prolonged survival in renal cell carcinoma (RCC), respectively (**Fig5d**). These observations remained after redefining somatic impact burden in relation to the burdening of corresponding randomized sets. Furthermore, these patterns remained after adjusting for patient age at diagnosis, low-impact mutation load, and –in the case of CLL, including a covariate for IgVH mutation status. These results lend support to the hypothesis that the aggregate amount of impactful passengers is clinically meaningful. More specifically, these results suggest that latent drivers are more important than deleterious passengers in CLL, but that the situation is reversed in RCC. We also related age at diagnosis with their impactful germline mutation burden. We observed that patients harboring a larger number of high-impact rare germline alleles were diagnosed with cancer at earlier ages in three cancer subtypes.

Additionally, we explored the role of impactful variants in cancer evolution by integrating their sub-clonality information. Intuitively, one might hypothesize that high impact mutations should either achieve higher prevalence in tumor cells if they are advantageous to the tumor, or a lower prevalence if deleterious. Interestingly, one finds suggestive evidences corroborating this hypothesis. We observe that high functional impact passenger variants in coding regions have higher pervasiveness among parental subclones (**Fig 5a**). High impact nominal passenger SNVs in tumor suppressor and apoptotic gene regions show enrichment in early subclones (**Fig 5a**). In contrast, high impact passenger SNVs in oncogenes appear slightly depleted. One interpretation of these findings is that passenger variants in tumor suppressor genes may have residual driver activity and that passenger variants in oncogenes impair oncogene activity at a detriment to tumor fitness. Similarly, impactful SNVs in DNA repair and cell cycle genes are depleted in early subclones (**Fig 5a**), suggesting that a high impact variant might eventually provide a critical burden for the survival of tumor cell. This observation is consistent with prior studies

highlighting role of deleterious passengers inhibiting cancer progression. Furthermore, we also observe lower heterogeneity among higher impact variants suggesting that pervasiveness of high impact variants within a tumor is more uniform compared to lower impact variants. This observation is consistent for both coding and non-coding variants (**Fig 5c**).

We employed a similar analysis using variant allele frequency (VAF) to explore whether passenger variants with high functional impact also conferred a fitness impact to tumor cells. We would expect for variants that enhance tumor cell fitness to achieve an overall higher than average mean VAF, while variants that reduce tumor cell fitness to occur at an overall lower mean VAF. Indeed, driver SNVs occur at higher mean VAF, non-silent coding SNVs and noncoding variants in sensitive regions occur at lower mean VAF, and synonymous variants along with variants in inter-genomic regions occur at intermediate mean VAF (**Fig 5b**). This suggest that in aggregate, non-silent passenger variants and noncoding variants in sensitive regions impair cancer cell fitness. Additionally, we generalize our observations among functional classes by correlating their respective variant frequency with the degree of conservation. Highly conserved positions (i.e. those with high GERP) are expected to be important for organismal fitness, as polymorphisms at those positions could hurt cellular function and in other cases because polymorphisms at those positions could promote undue cellular fitness (i.e. cancer) at the cost of organismal fitness. As expected, we observe that in PCAWG driver genes, VAF and GERP have a small but statistically significant positive correlation (with coefficient 0.0040 and p-value 0.0046). Interestingly, VAF and GERP have a correlation of similar magnitude but in opposite direction among variants not in driver genes, with very high significance (coefficient -0.0034, p-value < 2.2e-16). The observed trend for passenger variants at more conserved positions to occur at lower VAF is consistent with the deleterious passenger hypothesis.

Intuitively, tumor cells must require some minimal set of essential genes in working order to maintain homeostasis. One might imagine then that aggregate effect of functionally impactful passenger variants would be deleterious to tumor cells. Similarly, it has been proposed that presence of mildly beneficial passengers (latent/mini drivers) could provide fitness advantage to the tumor cell and maintain cellular homeostasis. In this work, we came across three different observations that support the notion that some nominal passenger variants affect tumor fitness. First, the VAF-related findings, such as the finding that nonsynonymous passenger mutations have lower VAF than do synonymous variants, suggest that impactful passenger variants hurt tumor cells enough to lower their VAF even if not enough to complete remove the variants from the tumor. Second, in some cancer subtypes, the most mutated tumors have a lower fraction of impactful variants than do less-mutated tumors, suggesting that impactful passenger variants become more deleterious when present in high numbers. Third, functional impact burden predicts patient survival time in select tumor subtypes. In conclusion, our work highlights that an important

subset of somatic variants originally identified as passengers nonetheless show biologically and clinically relevant functional roles across a range of cancers.

References

1. Vogelstein, B. & Kinzler, K. W. The Path to Cancer --Three Strikes and You're Out. *N. Engl. J. Med.* **373,** 1895–8 (2015).
2. Nussinov, R. & Tsai, C. J. 'Latent drivers' expand the cancer mutational landscape. *Current Opinion in Structural Biology* **32,** 25–32 (2015).
3. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15,** 680–685 (2015).
4. McFarland, C. D., Korolev, K. S., Kryukov, G. V, Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2910–5 (2013).

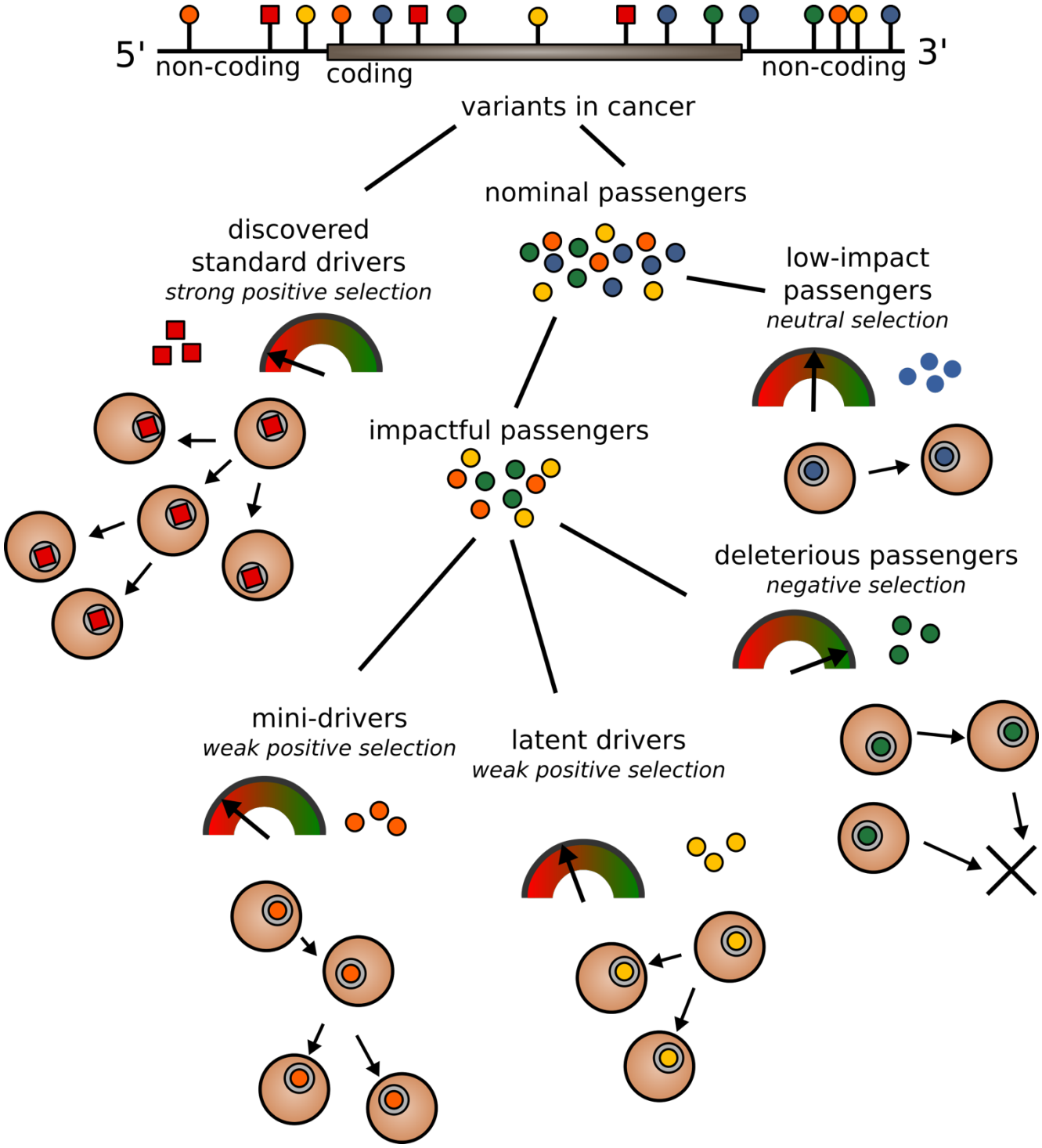

Figure 1. Classification of somatic variants into different categories based on their functional impact and selection characteristics: Both coding and non-coding variants can be classified as drivers and passengers based on their impact and signal of positive selection. Among nominated passengers, true passengers undergo neutral selection and tend to have low functional impact. Deleterious passengers, latent drivers and mini-drivers represent various categories of higher impact nominal passenger variants, which undergo weak negative or positive sections.

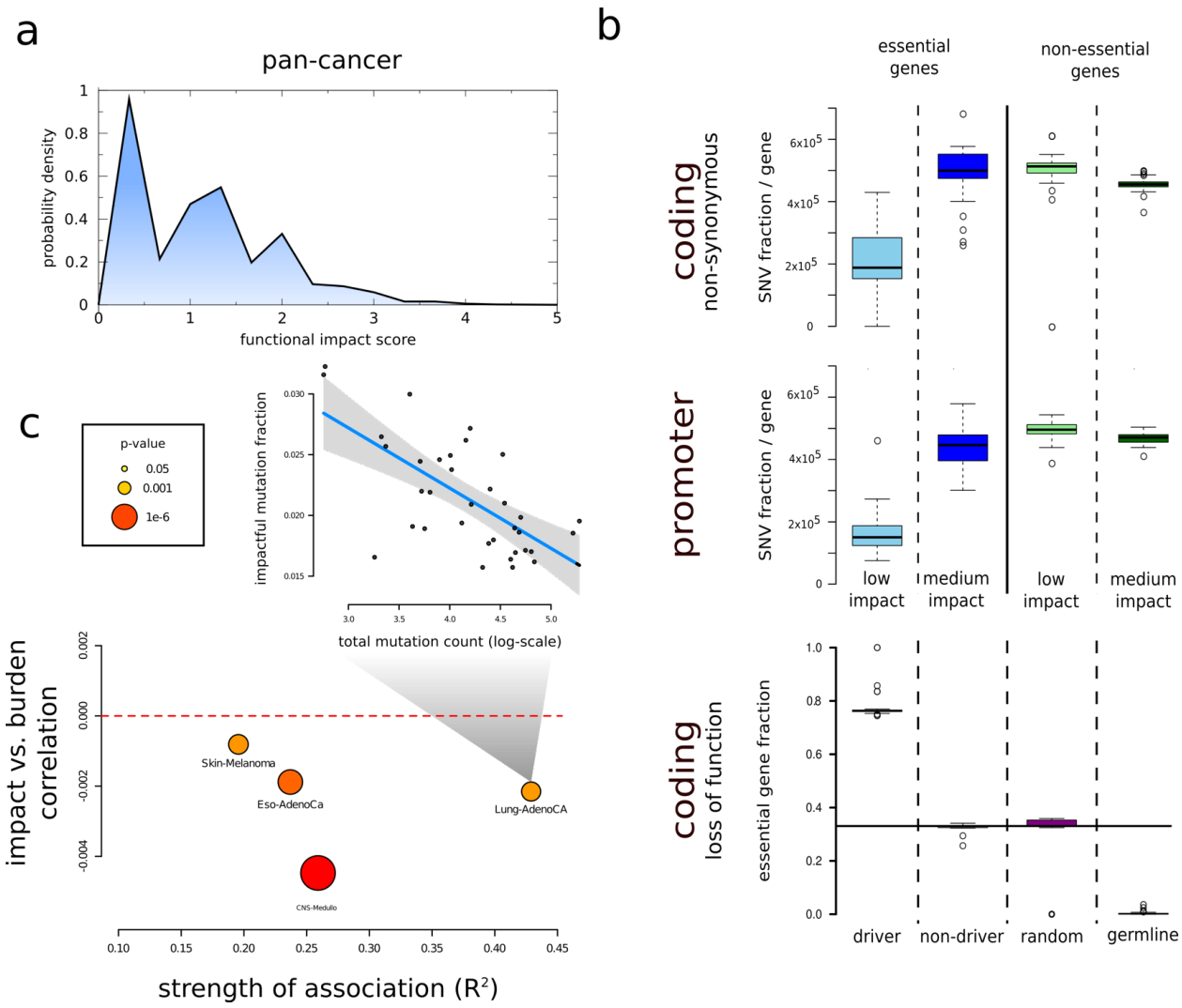

Figure 2: Functional impact scores for PCAWG SNVs: a) Functional impact distribution in noncoding region: three peaks correspond to low, medium and high impact variants. b) Fraction of impactful variants per gene in essential and non-essential gene sets: non-synonymous(top), promoter(middle) and loss-of-function(bottom). c) Correlation between number of impactful and total SNV frequencies for different cohorts.

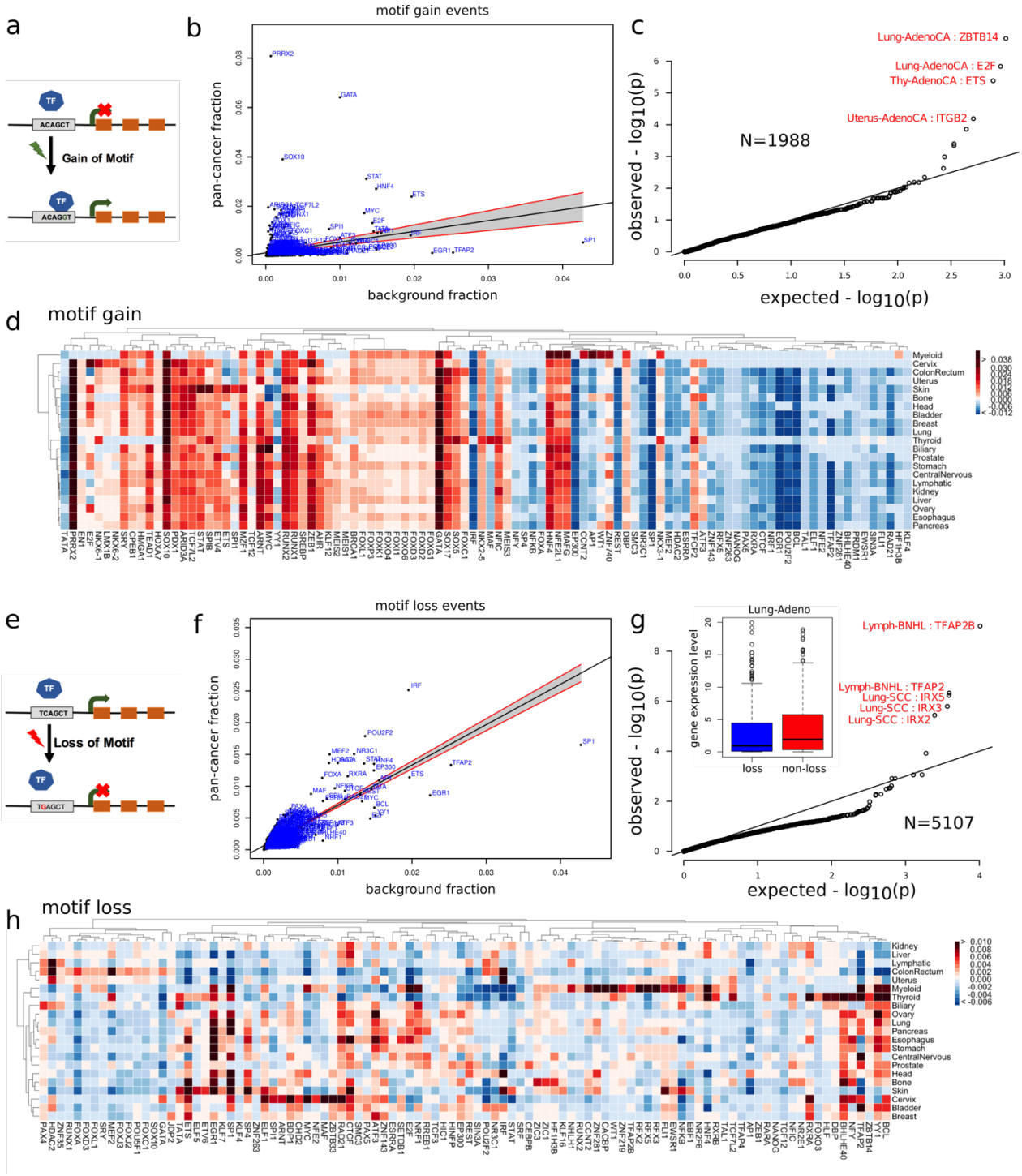

Figure 3: **Overall functional burdening of TF motifs:** ***Pan-cancer overview of TFs burdening****:* scatter plots for b) motif loss and f) motif gain events, ***Heat map presenting differential burdening of various TFs****:* SNVs leading to d) motif breaking and H) motif gain events in different cohorts compared to the genomic background. ***Gene expression changes due to motif alteration***: c) gene expression distribution for target genes for motif breaking and non-breaking scenario in Lung-Adenocarcinoma. g) Expression of target genes for TFs undergoing motif gain events.

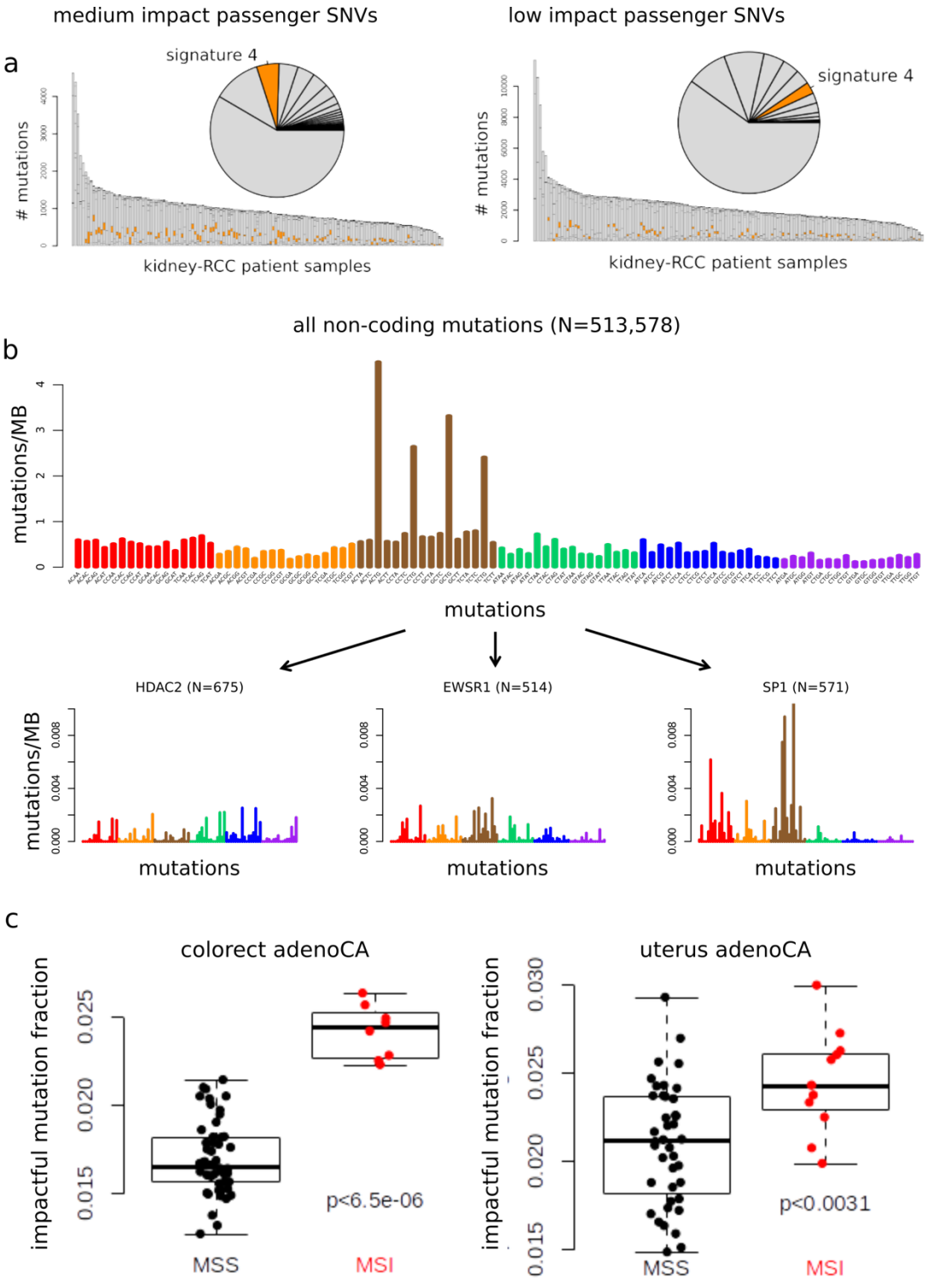

Figure 4: Mutational signatures associated with different categories of impactful variants: a) Distribution of canonical signatures in the kidney-RCC cohort for impactful (left) and low-impact SNVs (right). b) Mutation spectra associated with motif breaking events observed in HDAC2, EWSR1 and SP1 in the kidney-RCC cohort. c) fraction of impactful SNVs in MSI and MSS samples in Colorectal Adenocarcinoma(left) and Uterine Adenocarcinoma (right).

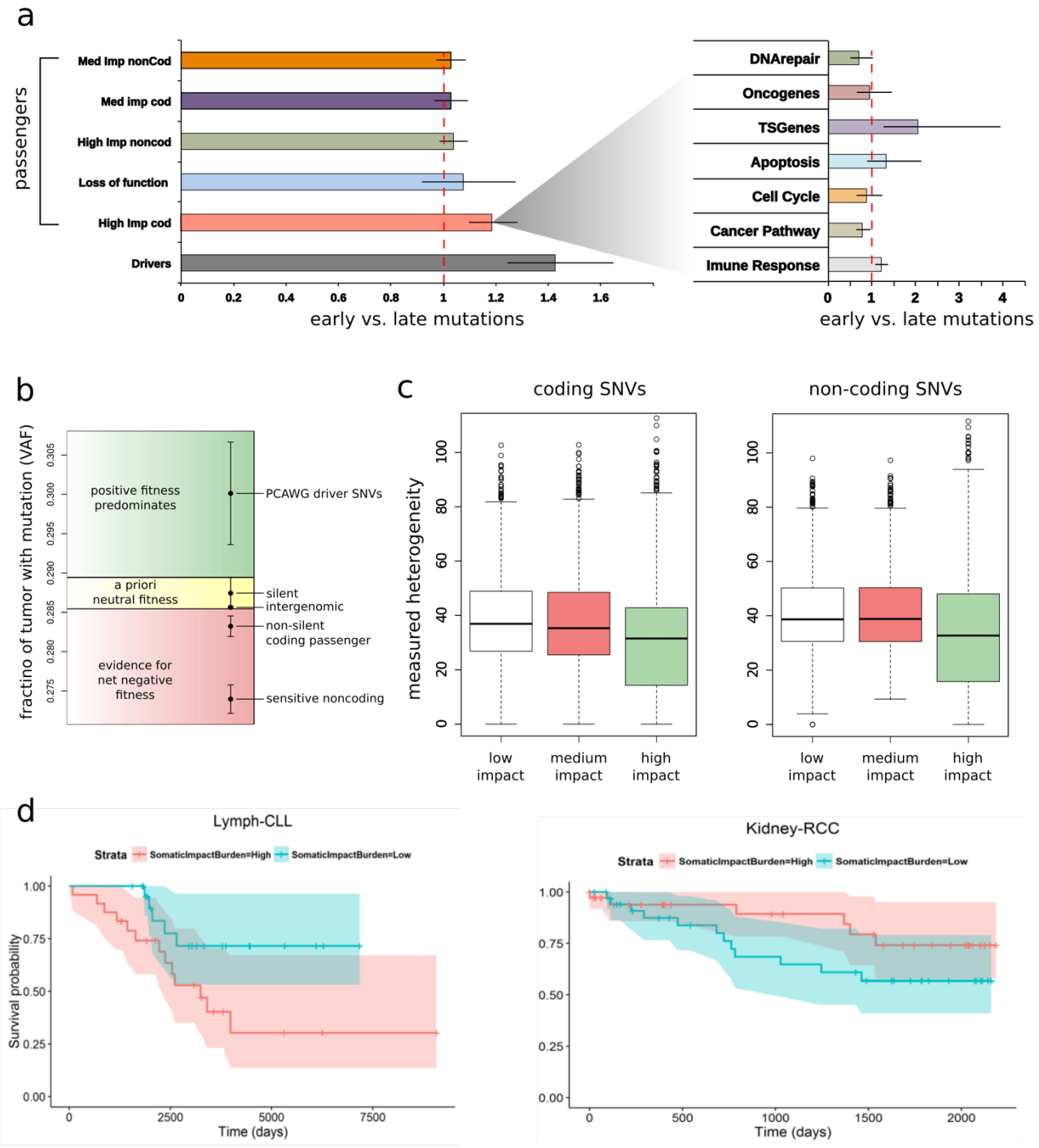

Figure 5: Correlating functional burdening with subclonal information and patient survival: a) Subclonal ratio (early/late) for different categories of SNVs (coding/non-coding) based on their impact score. Subclonal ratio for high impact SNVs occupying distinct gene sets. b) Stratifying SNVs in different selection classes based on their pervasiveness measured through mean VAF. c) Mutant tumor allele heterogeneity difference comparison between high, medium and low impact SNVs for coding(left) and non-coding regions(right). d) Survival curves in CLL (*left panel*) and RCC (*right panel*) with 95% confidence intervals, stratified by normalized impact burden.