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

Abstract

Classical models of cancer posit that only a small number of variants strongly drive tumor progression, and that the remaining variants (termed "nominal passengers") are inconsequential for tumorigenesis. In this study, we leverage the comprehensive variant data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project to evaluate the impact of each variant including nominal passengers to decipher their overall functional impact on different genomic elements. The overall functional impact distribution of PCAWG SNVs shows that, in addition to high impact drivers and low-impact passengers, there is a group of medium-impact passenger variants predicted to influence gene expression or activity. Furthermore, functional burdening relates to the underlying mutational signature and thus different signatures contribute to the functional burdening to different extent. Moreover, burdening of variants is non-random in terms of effecting different functional subsystems and for different categories of genes. We find that functional burdening varies based on subclonal architecture (i.e. early vs late mutations) and also further can be related to survivability of patients. Finally, we speculate on how the differential burdening might be related to both weak positive and negative selection during tumor evolution.

Introduction

Previous studies have extensively focused on characterizing variants occupying coding regions of various cancer genomes \cite{391996}. However, the extensive Pan-cancer Analysis of Whole Genomes (PCAWG) variant dataset, which comprises variant calls from ~2500 uniformly processed whole cancer genomes, offers an unparalleled opportunity to investigate the overall functional impact of variants influencing different non-coding genomic elements. Given that the majority of cancer variants lie in non-coding regions \cite{ 26781813}, 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 single-nucleotide variants (SNVs) and INDELs.

To a first approximation, all clinically significant consequences of genomic variants in cancer are mediated through changes in gene expression or gene activity; i.e. "functional impact." The relevance of functional impact is firmly established for driver mutations - positively-selected variants promoting tumor growth \cite{ 26304545, 23770567}. Nonetheless, of the thousands of variants in a typical tumor, very few of these ($< 5/\text{tumor}^1$) \cite{26559569} can be identified to drive tumor growth. The remaining variants, which we call nominal passenger variants, 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), some may weakly affect tumor cell fitness by promoting or inhibiting tumor growth, which in the literature have been reported as "mini-drivers" and "deleterious passengers," respectively \cite{ 23388632}. We reason that if any nominal passenger variants do indeed impact tumor cell fitness, this effect should be mediated by their functional impact; therefore, predicted functional impact serves as a good starting point to assess the potential tumor fitness effects of nominal passenger variants.

In this work, we explore the functional landscape of passenger variants in various cancer cohorts by leveraging extensive pan-cancer variant calls from PCAWG. More specifically, we build upon existing tools \cite{ 25273974} 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. Furthermore, integration of annotation and impact score allows for quantification of overall functional impact of variants occupying different genomic elements. We observe that disruption of regulatory elements in the noncoding genome correlates with altered gene expression. Moreover, regulatory elements are differentially affected by mutation process, as elucidated by our signature analysis.

Furthermore, we found that overall functional impact correlates with age at cancer diagnosis, patient survival, and tumor clonality. Finally, the ensemble of our work provides suggestive evidence that subsets of functionally impactful passenger variants confer weak selection preference to tumor cells.

Classifying variants based on impact and underlying selection process

In a continuum model, variants can be classified into three categories based on their fitness effects on tumor cells: positively-selected driver variants, neutrally-selected neutral passenger variants, and negatively-selected deleterious passenger variants. This broad classification can be further refined by considering ascertainment-bias and functional impact of different variants (**Fig 1**). Previous power analyses suggest that \cite{ 24390350}, in practice, existing cohort sizes support the identification of the mostly-strong positively-selected driver variants, but that many weaker drivers, and even some moderately strong driver variants would be missed. However, these moderately strong and weak driver variants also provide fitness advantage to tumor cells albeit at lower extent. As for the functional-impact-based-classification: The philosophy of molecular reductionism holds that any positively or negatively selected variants have some functional impact (i.e. effect on gene expression or activity). Some high functional impact variants may alter tumor gene expression or activity in ways that are not ultimately relevant for tumor fitness; hence, will be under neutral selection. However, all low impact non-functional variants will be neutrally selected. Similarly, rapid accumulation of moderately deleterious passengers, which undergo negative selection, could adversely affect the fitness of tumor cell and even drive tumor cell population to extinction in certain conditions \cite{23388632}.

Impactful passenger and their prevalence

In order to substantiate the presence of the above-mentioned continuum among nominal 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 nominal passengers,* which could include undiscovered drivers as well as deleterious passengers (**Fig 2a**). The interpretation of functional impact distributions depends on which null distribution serves as the comparison. A uniform null distribution is useful for making descriptive statements about the functional properties of the human genome and the functional impact of

mutational processes in cancer. A more sophisticated null distribution formed by variant shuffling has the potential to show evidence of selection and is described in more detail in Supplemental Method X.X.

According to a simple expectation, we might assume that the overall burden of variants in a cancer genome will be uniformly distributed across different functional elements and among different gene categories. In contrast, we observe that the functional burden in certain cancers is concentrated in particular gene categories. For instance, as a measure of the functional impact of both driver and non-driver loss of function (LoF) SNVs, we examined the fraction of deleterious LoFs affecting genes across four categories of cancer-related functional annotation (**Fig 2d**). Driver LOFs, which are well understood high impact variants, showed significantly high enrichment in each category of cancer-related functional annotation compared to random (shuffled-variant) control ($p < 0.001$). Conversely, non-driver LoF SNVs displayed depletion in each of these categories ($p < 0.001$). Driver, non-driver, and random loss of function mutations were all enriched in comparison to germline LoF mutations ($p < 0.001$). Given the high selective pressure presumed to act against germline deleterious loss of function mutations *in vitro*, our observations suggest that both driver and non-driver LoF mutations exert functional impact. Similarly, compared with the uniform null distribution, we observe that *impactful variants* (nonsynonymous & promoter SNVs) 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. This observation is consistent with underlying functional properties of the human genome.

Furthermore, based on uniform expectation, we would assume that the fraction of *impactful variant* will remain constant as one accumulate large amount of 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 drive 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).**

TF binding landscape and overall impact of variants

One might further expect that nominal passenger variants will contribute uniform functional burden 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, compared with uniform expectation. 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 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, thyroid adenocarcinoma and oligoastrocytoma. Furthermore, enrichment of SNVs in selective TF motifs leading to gain and break events in promoter significantly perturb the overall downstream gene expression (**Fig 3g**). For example, a close inspection of overall expression level of target genes for different TFs undergoing motif breaking events in lung adenocarcinoma cohort, indicate significantly lower expression values compared to instances when there was no loss in those TF motifs. Moreover, in lung adenocarcinoma, we found three TFBSs gain events (ZBTB14, E2F and HNF4) that 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).

Signature Analysis

The disproportionate functional load on 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.

Annotating structural variants

In addition to SNVs, large structural variations (SVs) are considered to play an 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 uniformly across the genome regardless of their extent of overlap with functional elements of the genome. However, our annotation analysis of somatic and germline SVs in PCAWG portrays a different picture. As expected, we observed higher enrichment of somatic SVs compared to germline SVs engulfing or partially overlapping with either coding region or transcription factor peaks, which suggest positive and negative selection of large SVs in somatic and germline cancer genomes, respectively. Similarly, we also observe significant enrichment of large engulfing somatic deletions as well as duplications among pseudogenes, non-coding RNAs, UTRs and ultra-conserved regions of the genome. 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**.**

Survivability and age of onset

Additionally, we sought to examine whether impactful passengers might be associated with tumor initiation and progression. Therefore, 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) (**Fig5d**). These observations remained after redefining somatic impact burden in relation to the burdening of corresponding variant-shuffled 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 number of impactful passengers is clinically meaningful. More specifically, these results suggest that undiscovered drivers are clinically more important than deleterious passengers in CLL, but that the situation is reversed in RCC. In addition, we observed similar correlation between age at cancer diagnosis with their impactful germline

mutation burden. More specifically, we found that patients harboring a larger number of high-impact rare germline alleles were diagnosed with cancer at earlier ages in three cancer subtypes.

Subclonality and impact score

Furthermore, we also 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 weak driver activity and that passenger variants in oncogenes impair oncogenic activity as 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**).

Functional impact and variant allele frequency

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.

Discussion

There are good *a priori* reasons to think that nominal passenger variants could have fitness effects on tumor cells. 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 on these essential genes would be deleterious to tumor cells \cite{23388632}. For instance, radiation therapy and some chemotherapies are believed to kill tumor cells by causing DNA damage \cite{}. Similarly, increased mutation counts in coding genes or regions relevant for splicing increase the antigenic cross-section of tumor cells, making them potentially vulnerable to immune surveillance \cite{}. Conversely, any variants that reduces the energy a cell spends on its organism-supporting functions to optimize cell-division could be expected to have a small but not easily detected positive effect on tumor fitness. Moreover, certain variants through their complex genetic regulatory interactions might moderately increase the expression levels of canonical oncogenes. These weak undiscovered driver variants have been proposed to undergo small positive selection to benefit tumor growth.

In this work, we came across multiple observations that support the notion that some nominal passenger variants affect tumor fitness. First, the finding that variants at more conserved positions have lower VAFs suggests that impactful passenger variants can encumber the tumor cells they inhabit. Second, in some cancer subtypes, the most mutated tumors have a lower fraction of impactful variants than do less-mutated tumors, suggesting either that the aggregate impact of impactful passenger variants becomes more deleterious at higher mutation loads, or alternatively but equally interestingly, that some fixed number of undiscovered drivers is diluted at higher mutation counts. Third, our LoF related analysis indicate that driver LoF mutations exert a positive selective effect, whereas non-driver LoF mutations apparently exert a net negative selective pressure. This observation is consistent with prior evidence of net negative selective effect among nominal passenger missense mutations. This putative fitness impact of nominal passenger variants may help explain why patient survival times are correlated with functional impact load in select 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).

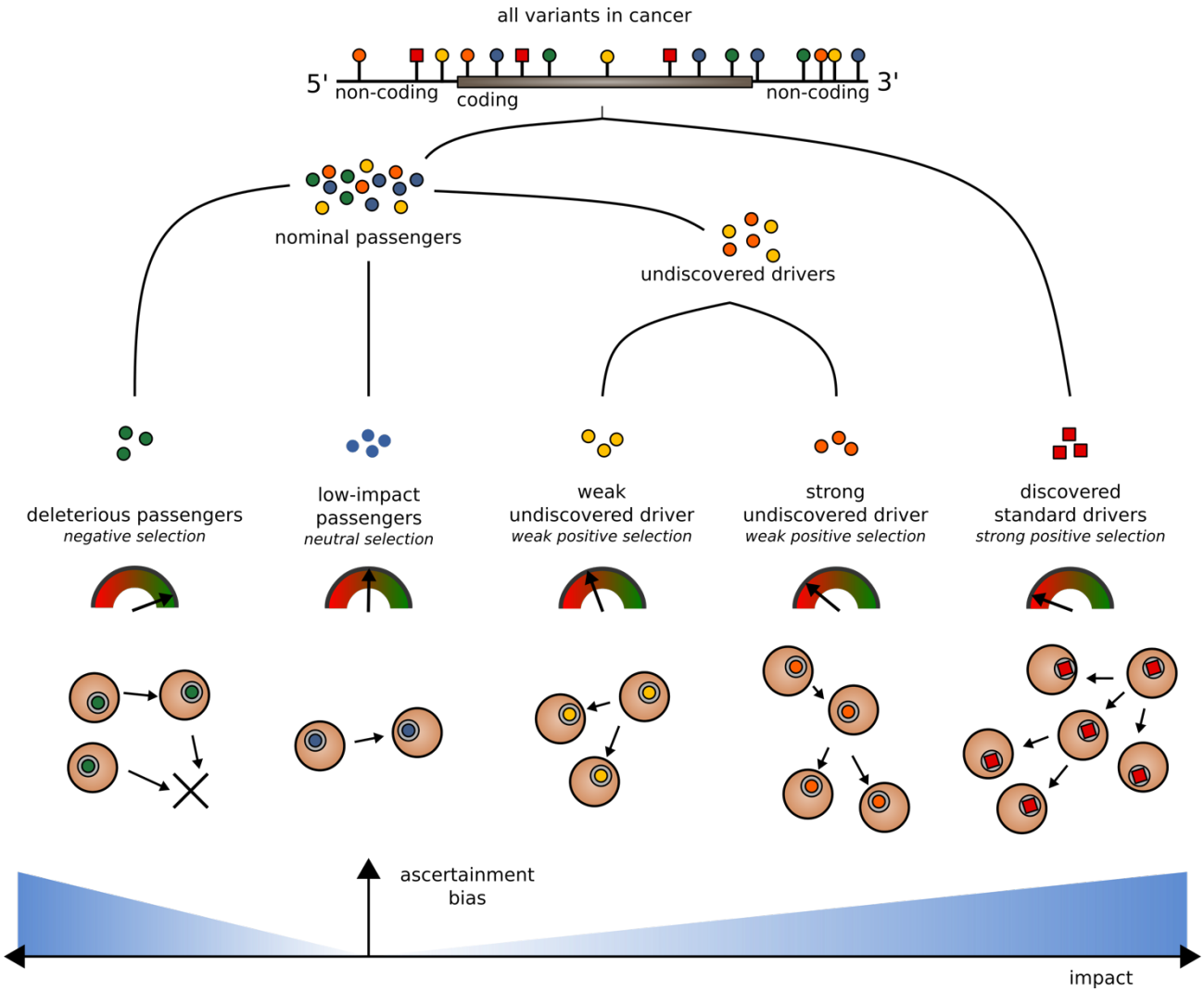

Figure 1. Classification of somatic variants into different categories based on their functional impact and selection characteristics: Both coding and non-coding variants can be classified as drivers and passengers based on their impact and signal of positive selection. Among nominated passengers, true 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.

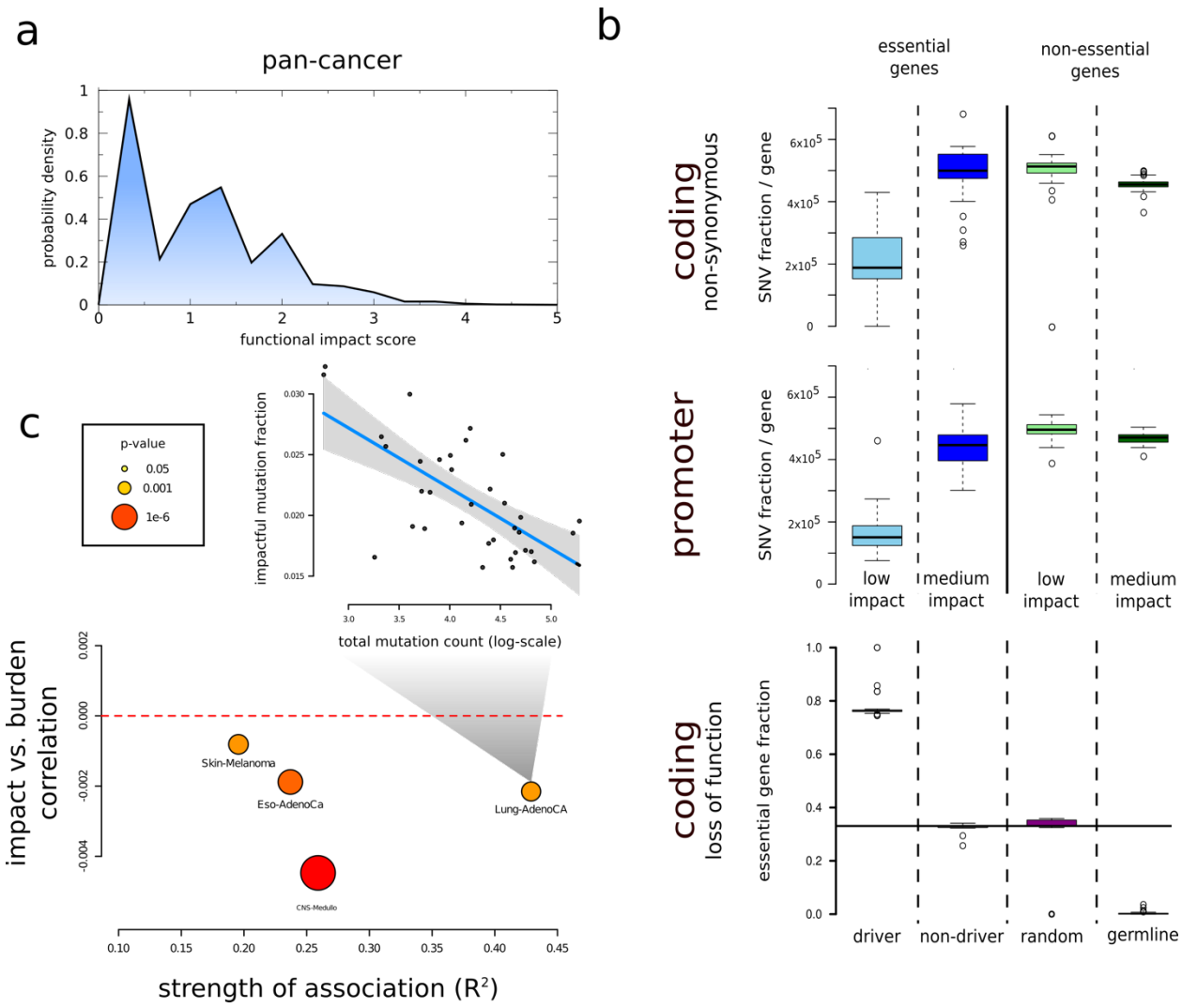

Figure 2: Functional impact scores for PCAWG SNVs: a) Functional impact distribution in noncoding region: three peaks correspond to low, medium and high impact variants. b) Fraction of impactful variants per gene in essential and non-essential gene sets: non-synonymous(top), promoter(middle) and loss-of-function(bottom).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

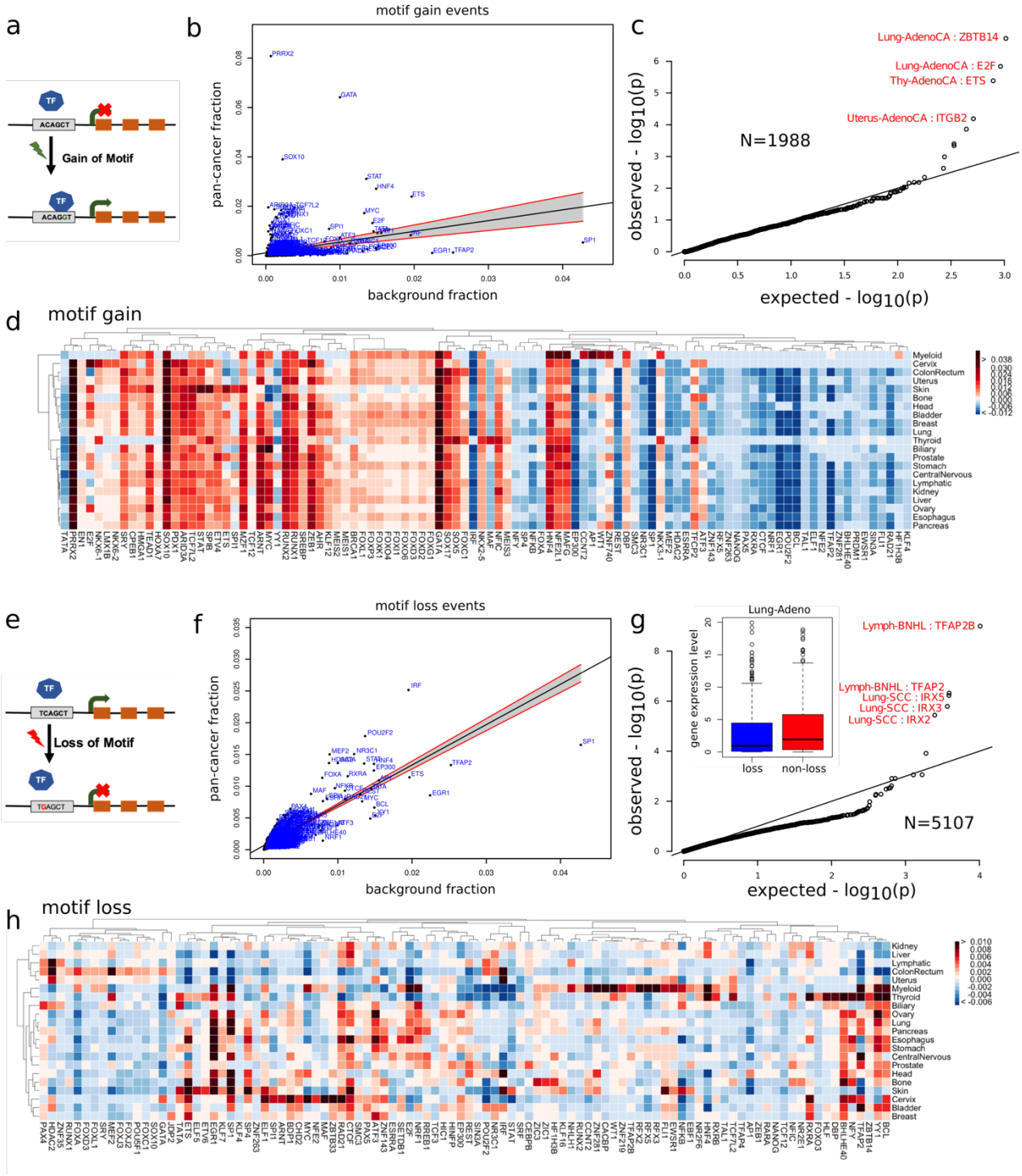

Figure 3: **Overall functional burdening of TF motifs:** ***Pan-cancer overview of TFs burdening****:* scatter plots for b) motif loss and f) motif gain events, ***Heat map presenting differential burdening of various TFs****:* SNVs leading to d) motif breaking and H) motif gain events in different cohorts 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.

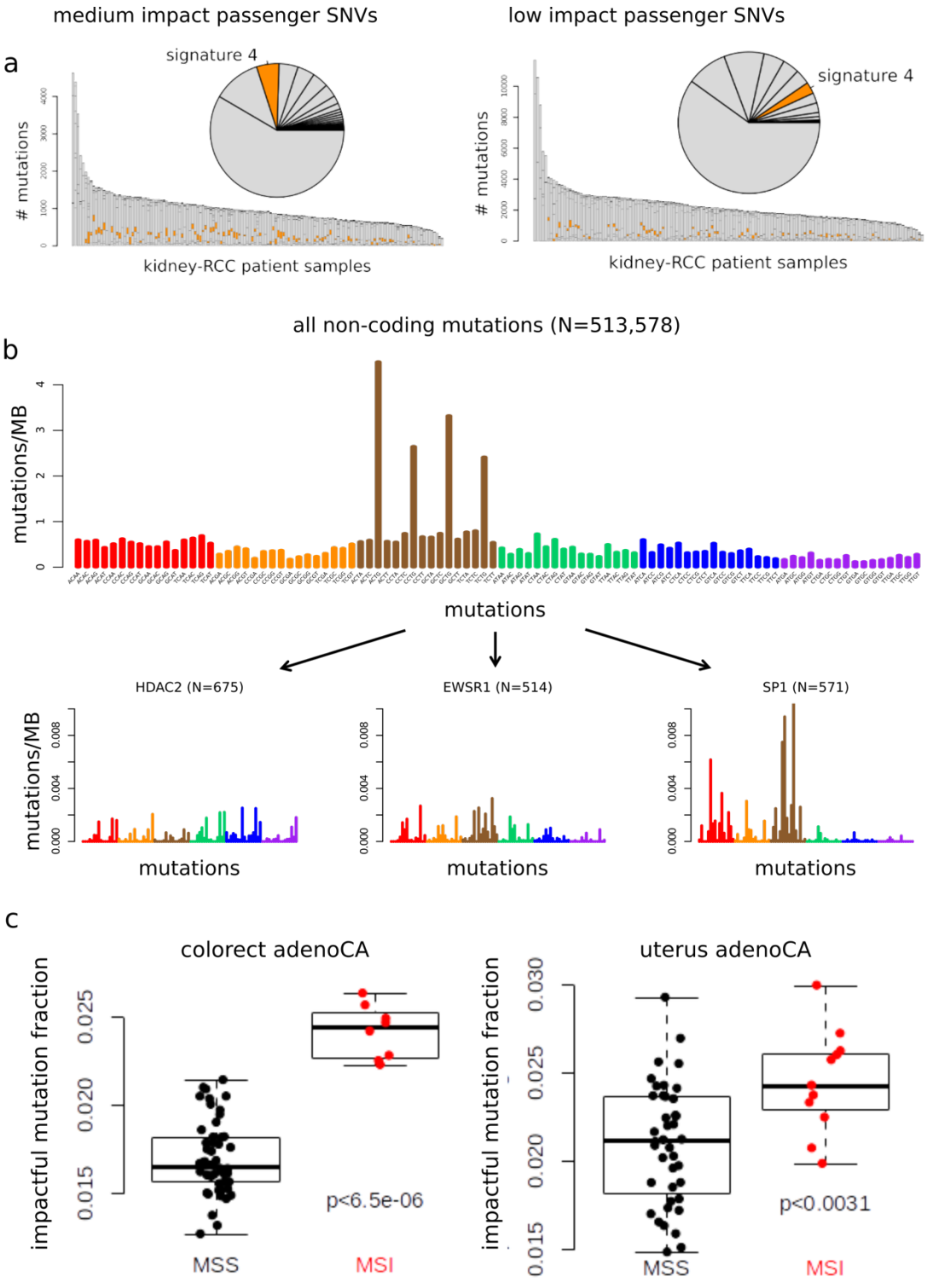

Figure 4: Mutational signatures associated with different categories of impactful variants: a) Distribution of canonical signatures in the kidney-RCC cohort for impactful (left) and low-impact SNVs (right). b) Mutation spectra associated with motif breaking events observed in HDAC2, EWSR1 and SP1 in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

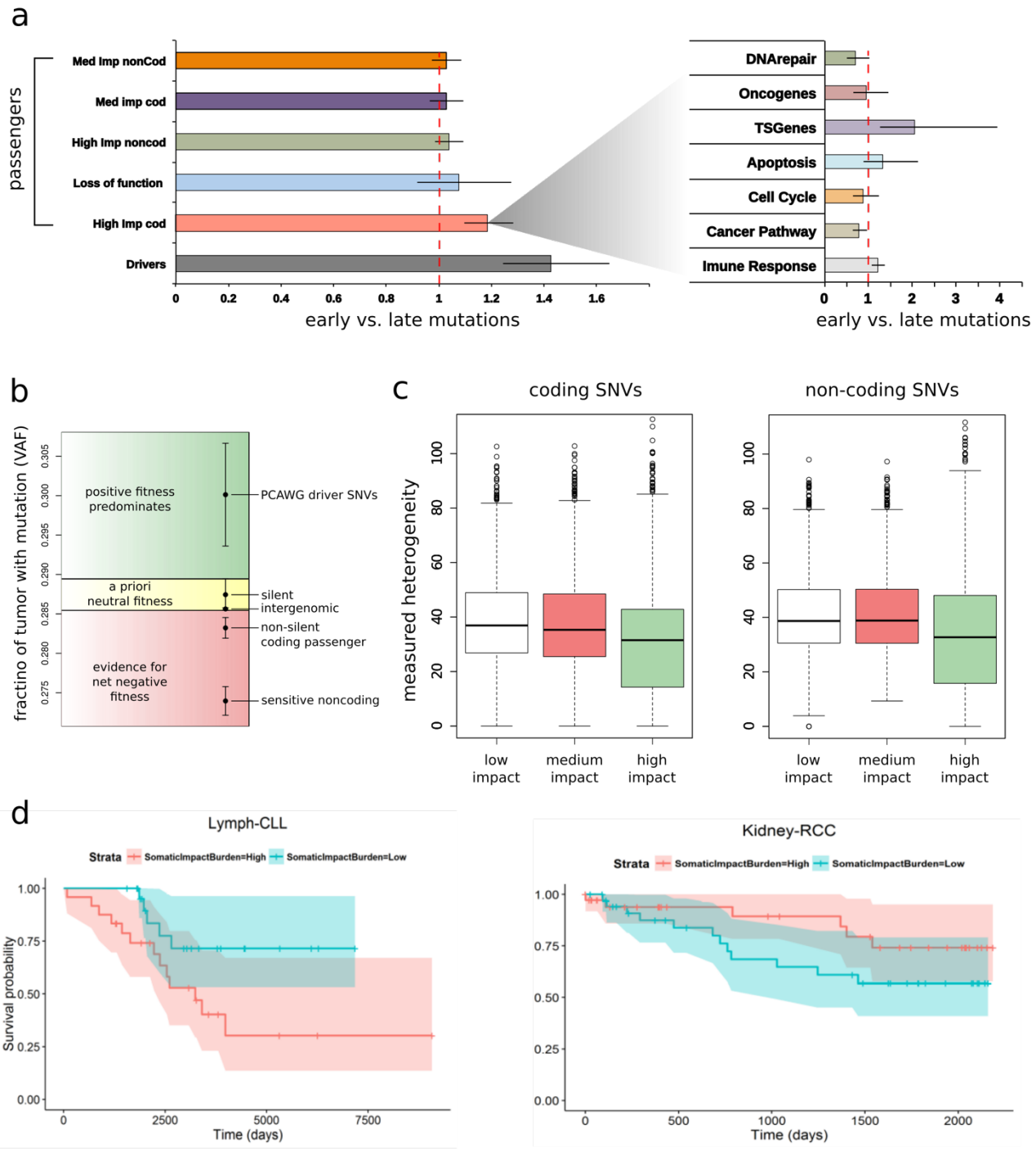

Figure 5: Correlating functional burdening with subclonal information and patient survival: a) Subclonal ratio (early/late) for different categories of SNVs (coding/non-coding) based on their impact score. Subclonal ratio for high impact SNVs occupying distinct gene sets. b) Stratifying SNVs in different 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.