Passenger mutations in >2500 cancer genomes: Overall burdening & selective effects

A typical tumor has thousands of genomic variants, yet very few of these (<5/tumor[1]) are thought to drive tumor growth. The remaining variants, 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. Formally, passengers can be subdivided into neutral and impactful based on their predicted functional impact on the genome. Low-impact passengers are thought to be inconsequential for tumor progression. However, impactful passengers can alter gene expression or activity, and while some of these changes may be irrelevant, others may promote or inhibit tumor cell growth and survival, as has been suggested for *latent driver variants* [2,3] ("mini-drivers") and *deleterious passengers*[4], respectively.

Here, we explore the landscape of passenger impact in various cancer cohorts by leveraging extensive pan-cancer variant calls from ~2700 uniformly processed whole cancer genomes. This variant dataset serves as an ideal resource to explore the role of impactful passenger variants, considering that the majority of passenger variants occupy non-coding regions of the genome and that some of these variants such as large SVs are difficult to identify from exomes alone. Our purposes are two-fold: First, we build on and apply existing tools to score the predicted impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset, relying on features such as the degree of evolutionary conservation and overlapping functional elements of the involved positions. Subsequently, we search for signals of positive and negative selection of passenger variants, particularly those we predict to be impactful.

We find seven kinds of evidence of selection upon passenger variants. We find that the distribution of impact scores among passenger variants, their mutational signatures, co-mutation frequencies, variant allele frequencies, evolutionary timing, and ability to predict age of cancer onset (in the case of germline variants) and patient survival all clash to varying degrees with the null hypothesis that all nominal passenger variants are neutral.

In order to substantiate the presence of various categories of passenger variants, we surveyed the functional impact distribution of somatic variants in the pan-cancer dataset. While some variants are impactful and play prominent role in cancer progression, others are less effective and are generally ignored during cancer studies. Based on canonical classification of somatic variants as passenger and drivers, one might expect their functional impact score distribution to be unimodal and centered around 0 (as a result of a large number of neutral passengers), along with a tail in the high-impact score regime, corresponding to putative drivers. However, inspection of impact scores for somatic variants across

cancer cohorts reveals a very different picture: passengers can be broadly classified into three distinct subgroups. The upper and the lower extremes, which comprise ~23 and ~13,500 noncoding variants per patient, fall under traditional definitions of high-impact putative driver variants and neutral passengers. In contrast, the intermediate functional impact regime comprises of *impactful passengers* (~3,500 noncoding variants per patient), which can further influence cancer progression by acting as latent drivers or through aggregate burdening of functional elements.

We observe a heterogeneous enrichment profile of these impactful passengers in different cancer-subtypes and different categories of genes. Cohort level analysis indicates enrichment of impactful passenger SNVs in various cancer types. Myeloid-MPN, colorectal & uterus adenocarcinoma cohorts stood out with higher enrichment level of impactful passengers compared to others. In addition, a gene-centric analysis indicates a higher fraction of impactful variants in key genes (essential, metabolic & immune-response genes) compared to neutral passenger variants. Conversely, neutral passengers constitute larger fractions of variants influencing non-essential genes. This observation is consistent with previous studies suggesting role of non-neutral passenger variants in cancer progression by burdening key genes including housekeeping, metabolic and immune-response genes. Similarly, somatic LOF variants (both SNVs and INDELs) are highly enriched among essential genes in the entire pan-cancer dataset.

Furthermore, we closely inspected signature composition of neutral and impactful passengers in each cancer-cohort to distinguish between mutational processes that generate these distinct classes of passenger variants. For instance, we observed distinct signature distributions for the impactful and neutral non-coding passengers in the Kidney-RCC cohort. While the majority of neutral and non-neutral passengers can be explained by signature 5, impactful passengers have a higher fraction of SNVs explained by signature 4. This suggests that impactful passenger SNVs show shifts in mutational signatures compared to the neutral ones.

One might further expect that the presence of impactful passengers varies among different genomic elements as well as different cancer cohorts. Consequently, we comprehensively analyzed the overall burdening of various genomic elements, including TF (transcription factor) binding motifs in the pan-cancer somatic variant dataset. The presence of a variant within a TF binding site 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 *impactful variants* 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 across major cancer types analyzed in this study. Similarly, high impact variants influencing gene expression by breaking TF motifs, were highly enriched in YY1, BCL, RAD21 and CTCF in a majority of cohorts. This selective enrichment or depletion suggests distinct alteration profiles associated with different components of regulatory networks in various cancers. Furthermore, signature

analysis of these variants influencing TF binding sites suggests that distinct signatures burden motifs disproportionately.

Similarly, structural variants are considered to play a pivotal role in driving cancer progression, thus we annotated and evaluated the impact of large SVs in the entire PCAWG cohort. Our annotation analysis suggests 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 also explored the role of impactful variations in cancer evolution by integrating them with subclonal information and allele frequencies. Intuitively, one might hypothesize that high impact mutations should either achieve higher frequency if they are advantageous to the tumor, or a lower frequency if deleterious. Interestingly, one finds suggestive observations that this is the case. In particular, we observe that high functional impact non-coding variants (along with high impacting coding LOF variants) have a higher allelic frequency and a higher prevalence in parental subclones, signifying a potential important role in the early phases of cancer progression or providing a higher fitness advantage.

Furthermore, it has been proposed that two or more low impact variants might confer a selective advantage to tumor cells when mutated together – the so-called, epistatically interacting passengers. We find statistical evidence for the existence of epistatic drivers among the PCAWG variants in the form of gene-pairs that are co-mutated more frequently than expected under additive-effects assumptions. In our co-mutation analysis, we observed XXX significantly co-mutated and XXX significantly under-co-mutated gene-pairs containing at least one passenger gene, with an FDR of 10%, as well as XXX and XXX among germline passengers. One interesting observation from this analysis is that anti-correlated gene-pairs are substantially more likely to participate in the same pathway than are randomly chosen gene-pairs, which is consistent with plausible mechanisms of synergy and redundancy.

Finally, we sought to examine whether impactful passengers might exert a clinically meaningful effect on tumor initiation and progression. Therefore, we correlated patient impactful germline mutation burden with patient age at diagnosis and patient impactful somatic burden with patient survival. 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. We then performed survival analysis to see if somatic impact burden 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. These observations remained after defining 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. This can be explained by the large share of missing drivers in CLL, which suggests a greater role for latent drivers in CLL.

In conclusion, our work highlights an important subset of somatic variants originally identified as passengers nonetheless show biologically and clinically relevant functional roles across a range of cancers.