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. 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. More specifically, we annotate and evaluate the impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset. Subsequently, we integrate their annotations and impact scores to quantify the overall burdening of various elements in cancer genomes. Furthermore, we also show how overall functional burdening correlates with age at cancer diagnosis, patient survival time, and tumor clonality.

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. 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. More specifically, we found that these impactful passengers are highly enriched among patients in Myeloid-MPN, colorectal and uterus adenocarcinoma cohorts. This observation is consistent with prior studies, which suggest role of non-neutral passengers in cancer progression in these cancer-subtypes. In addition, gene-centric analysis indicate enrichment of these non-neutral passenger variants in key genes such as metabolic, immune-responsive and essential genes in general. Moreover, our analysis clearly suggest that impactful passenger SNVs show shift in mutational signatures compared to the neutral ones.

One might further expect that 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) 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 majority of cohorts. This selective enrichment or depletion suggest distinct alteration profiles associated with different components of regulatory networks in various cancers. Furthermore, signature analysis of these variants influencing TF binding sites suggest 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 region, UTRs and TF peak regions. Moreover, engulfing SVs tend to have higher enrichment value compared to partially overlapping SVs. Furthermore, we also observed high impact large deletions and duplications in meta tumor cohorts such as CNS, Glioma and Sarcoma.

Additionally, we explored the role of impactful variations in cancer evolution by integrating them with sub-clonal 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. 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. Nine subtypes are represented among the nine gene-gene-subtype triples, especially squamous cell cancers – both of the lung and of the head.

Finally, we sought to examine whether impactful passengers might exert a clinically meaningful effect on cancer initiation and progression. To study role of impactful passenger variants on tumor progression, we performed survival analysis to see if impact burden in non-driver genes predicted patient survival within individual cancer subtypes. Interestingly, we discerned a statistically significant correlation between functional burden and patient survival for few cancer cohorts. However, 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 results lend support to the hypothesis that the aggregate amount of impactful passengers is clinically meaningful. More specifically, the 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.

Figure1

Panel A:

- Thousands of genomic variants mostly in noncoding region. very few of these (<5/tumor) are thought to drive tumor growth.

- Functional consequence of passenger variants poorly understood.

- Passengers can be subdivided into neutral and impactful based on functional impact score.
    - Low-impact passengers are thought to be inconsequential for tumor progression.
    - impactful passengers can alter cancer progression in two distinct mode
        - deleterious passengers decrease the cancer cell fitness and inhibit cancer growth
        - impactful passengers can drive cancer progression by interacting epistatically

Panel B:

- Functional impact distribution for non-coding variants indicate presence of non-neutral variants
- The upper and the lower extremes of the distribution corresponds to high-impact putative driver variants and neutral passengers.
- Intermediate functional impact regime comprises of *impactful passengers* (~3,500 noncoding variants per patient)

Panel C:

- Cohort level analysis indicate prevalence of impactful passenger SNVs in certain cancer-subtype compared to others.
- Myeloid-MPN, colorectal & uterus adenocarcinoma and melanoma cohorts are highly enriched in impactful passenger variants compared to others.
- Observation consistent with prior studies indicating role of impactful passenger variants in Myeloid-MPN and colorectal cancers.

Panel D:

- Previous study suggests that non-neutral passenger variants influence cancer progression by burdening key genes including housekeeping, metabolic and immune-response genes.
- We observe higher fraction of impactful variants in these key genes (essential, metabolic & immune-response genes) compared to neutral passenger variants.
- Moreover, neutral passengers constitute larger fractions of variants influencing the non-essential genes.
- In addition, somatic LOF variants (both SNVs and INDELs) are highly enriched compared to germline LOFs in the entire pan-cancer dataset. This is intuitively consistent.

Panel E:

- Signature composition of neutral and non-neutral passengers in a given cancer-cohort can help decipher underlying mutational processes
- A close inspection of neutral and impactful non-coding passengers in Kidney-RCC indicate distinct signature profile
    - Majority of neutral and non-neutral passengers can be explained by signature 5.

- However, non-neutral passengers have higher fraction of SNVs explained by signature 4. In contrast, larger fraction of neutral passengers are constituted by signature 8.

Figure2

- We performed gene level analysis (pan-cancer & cancer-specific) to decipher the overall functional burdening observed by various functional element of the cancer genome.
- We compared these gene-centric burden in context of highly disrupting (LOFs) and mildly deleterious coding and non-coding variants.

Panel A:

- As expected, gene centric analysis of disruptive LOF variants on pan-cancer level indicate higher overall burdening in known cancer genes such as TP53, CDKN2A, KMT2D,APC and SMAD4.
- Cancer-specific analysis indicates higher burdening of TP53 across cancer types, which is consistent with prior studies. Similarly, APC gene in colorectal cancer is known to carry higher burden of LOF variants.
- More interestingly, we observe high enrichment of LOF variants in multiple genes in melanoma, which is not observed in other cancer types.

Panel B:

- Nonsynonymous passenger variants are highly enriched among essential genes, which are involved in important biological functions such as gene regulation (INTS1), metabolism(AMPD1), signal transduction (SHC1), enzymatic activity(KMT2B) and cell maintenance (LMNA).
- We also identify essential genes, which are highly enriched in impactful nonsynonymous passenger variants in different cancer types.
    - For instance, impactful nonsynonymous variants are highly enriched in KAT8 genes, which is involved in chromatin organization and P53 pathway
    - Similarly, KMT2B gene has large burden of nonsynonymous non-neutral passenger variants in Melanoma. Prior studies suggest role of this gene in various solid cancer.
    - ACTB and TMSB4X gene, which are involved in cell proliferation, differentiation and integrity are highly burdened with impactful nonsynonymous variants in mature B-cell lymphoma cohort.

Panel C:

- Our pan-cancer analyses of noncoding variants indicate large burdening of non-neutral passenger variants in promoter region of various known cancer genes such as TRAF7, RB1, GNA11 and NF2.
- We also identified promoter region of known cancer genes with relatively high enrichment of impactful variants in different cancer type.
    - For instance, promoter regions of CEBPA, ASXL1 and NF2 genes are highly enriched in impactful non-coding variants in the breast-adenocarcinoma cohort.
    - Similarly, promoter regions of RB1 and BRCA1 have higher burden of impactful promoter variants compared to by chance.

- In addition, PIK3R1 and NF2 gene promoters are also highly enriched in impactful noncoding variants in melanoma.

Figure3:

- We evaluate impact of somatic variants on the transcriptional landscape of various cancer genomes.
- More specifically, we calculate the overall functional burdening of various transcription factor in the context of SNVs generating (gain-of-motif) and disrupting (loss-of-motif) their binding sites.
- Clustering of TFs suggest that majority of TFs undergoing break/gain of events belong to the same family.

Panel A:

- Pan-cancer analysis of all SNVs breaking TF motifs indicate that
  - Motif breaking SNs are highly enriched among TFs such as SP1, ETS, ELF, EGR1, ZNF143, YY1, BCL, RAD21 and CTCF.
  - We observe high enrichment in TF breaking events in cancer subtypes such as Lymph-NOS, AML and Bone-Epith.

Panel B:

- Pan-cancer analysis of all SNVs leading to gain of TF motifs indicate that
  - Gain of motif events is widely observed across various cancer types and influence wide-range of TFs.
  - Gain of motif event is highly enriched in TFs such as GATA, PRRX2 and SOX10 across all cancer types.
  - Interestingly, we don't observe enrichment of SNVs leading to gain of TF motif events in Myeloid-MDS, Breast-DCIS and Bone-Cart.

Panel C:

- Signature profile of motifs breaking different TFs varies a lot
  - Mutation spectrum of motif breaking events in SP1 suggest major contribution from C>T and C>A mutation. In contrast, T>A, T>C and T>G mutations are lowly contributing in the overall mutational profile.
  - In contrast, both HDAC2 and EWSR1 has relatively uniform mutation spectrum profile.
  - Interestingly, T>A, T>C and T>G mutation contribute higher in HDAC2 breaking compared to breaking events in EWSR1 and SP1 TF motifs.

Figure4

Enrichment of functional Impact mutations in Early vs Late Subclones

- Mutations of positive fitness should appear with higher frequency (cell prevalence) or enriched in early/dominant subclones. Mutations of negative fitness should wash off the population
- Functional Impact mutations are enriched in Early vs Late Subclones and High Freq vs Low Frequency bins (Frequency graph not shown here)
- High Impact mutations in oncogenic regions are particularly enriched in early subclones showing higher fitness

Average number of LOF mutations in early vs late Subclones

- We expect: a) either higher fitness for each cell and therefore in higher prevalence or b) lower fitness and lower prevalence
- Tumor Suppressor gene LOF should -by definition- provide a higher fitness for tumor cells and therefore exist in higher than expected prevalence

Quantitative Changes in Signature proportion in Early vs Late Subclones

- Values of change in entropy closer to 1 signify higher Entropy in early signatures (Fewer of higher prevalence). Negative values signify higher entropy for late signatures. Closer to 0 Signify no change in signature proportions.
    - Female Reproductive, Lymph and CNS tumors show the most conservative nature, remaining mostly unchanged.
    - Sarcoma and Glioma show fewer and more dominant signatures in early/dominant subclones
    - Kidney, Lymph, Squamous and Digestive tumors show fewer signatures in later subclones

Figure5

Panel A:

- Correlations are observed between pairs of genes with loss of function.
- This may represent cooperativity between mutations, e.g. two-hits of the two-hit hypothesis, or the presence of such co-mutation may be correlated with cancer invasiveness (those cancers that are more likely to be detected and biopsied or resected).

Panel B:

- Notable anti-correlations are also observed between genes with loss of function.
- This may represent 'minimum' conditions for the development of malignancy -- a lowest common denominator.
- It may also be indicative among the heterogeneity of cancers among tissue of origin -- cancers do not develop by the same mechanism in different tumors/tissues.

- Finally, these mutations may be correlated with rate of cancer progression. Once one of these mutations is acquired (even if malignancy is already onset), the rate of cancer development is accelerated, leading to low likelihood of acquiring other mutations that accelerate cancer development.

Panel C:

- Despite being neutral on their own, two or more variants might confer a selective advantage to tumor cells when mutated.
- We investigated epistatic events among the PCAWG variants in the form of gene-pairs that are co-mutated more frequently than expected under additive-effects assumptions.
- We observe significant level of co-mutation between PIK3CA and TP53 as well as XIRP2.similarly, DDX3 gene is significantly co-mutated with PRKAR1A.

Panel D:

- survival analysis to see if impact burden predicted patient survival within individual cancer subtypes.
    - 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 adjusting for patient age at diagnosis and low-impact mutation load and after defining tumor impact burden in relation to the burdening of corresponding randomized sets.

    - These results lend support to the hypothesis that the aggregate amount of impactful passengers is clinically meaningful.
        - More specifically, the 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.

Figure6

Panel A & B:

- Enrichment of large engulfing somatic deletions is highest among Pseudogenes, coding region, UTRs and intronic region.

- Partially overlapping deletions are not highly enriched as engulfing ones.

- Partially overlapping deletions enriched among lincRNA and pseudogenes to similar extent compare to engulfing scenario.

- Cartilaginous neoplasm cohort has highest enrichment values for engulfing and partially overlapping deletions compared to other cancer types in the PCAWG. This is consistent across different genomic elements.

- Presence of high impact large deletions in CNS, Glioma , Sarcoma and Kidney cohorts.

- On average, deletions in Female reproductive meta tumor cohorts are have comparatively lower impact score.

Panel C & D:

- Enrichment pattern remains similar for large duplications as in large deletions.
    - Coding, intronic, UTR and pseudogenes overlap to a larger extent for both engulfing and partial overlap scenarios.

- Cartilaginous neoplasm, In-situ Breast Adenocarcinoma, Pancreatic Adenocarcinoma and Pilo-Astrocytoma cohorts have higher enrichment values for duplication overlap.

- Presence of high impact duplications in CNS, Glioma and Sarcoma meta tumor cohorts. Duplications in Squamous cohort tend to have lower impact score on average.