Passenger mutations in 2500 cancer genomes: Overall molecular functional impact and consequences

Abstract

The Pan-cancer Analysis of Whole Genomes (PCAWG) project provides an unprecedented opportunity to comprehensively characterize a vast set of uniformly annotated coding and non-coding mutations present in thousands of cancer genomes. Classical models of cancer progression posit that only a small number of these mutations strongly drive tumor progression and that the remaining ones (termed "nominal passengers") are considered inconsequential for tumorigenesis. In this study, we leverage the comprehensive variant data from PCAWG to ascertain the molecular functional impact of each variant, including nominal passengers. This allows us to decipher their overall impact uniformly over different genomic elements, both coding and non-coding. The molecular impact distribution of PCAWG mutations 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, we find that functional impact relates to the underlying mutational signature: different signatures confer contrasting molecular impact, differentially affecting distinct regulatory subsystems and different categories of genes. Also, we find that molecular functional impact varies based on subclonal architecture (i.e. early vs. late mutations) and can be related to patient survival. We further show that nominal passenger mutations explain a significant portion of the total variance (25-40% additive variance) when observed cancer variants are compared to random expectation from a null model. Finally, we speculate on how the overall burdening of cancer mutations might be related to the existence of both weak positive and negative selection during tumor evolution.

Introduction

Previous studies have focused on characterizing variants occupying coding regions of cancer genomes \cite{24071849}. However, the extensive Pan-cancer Analysis of Whole Genomes (PCAWG) dataset, which includes variant calls from ~2500 uniformly processed whole cancer genomes, offers an unparalleled opportunity to investigate the overall molecular 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. Moreover, 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 small insertion & deletions (INDELS).

Of the 30 million SNVs in the PCAWG data set, a few thousand (< 5/tumor[1]) \cite{26559569} can be identified as driver variants – positively selected variants that favor tumor growth. The remaining ~99% of SNVs are termed nominal passenger variants, with poorly understood molecular consequences and fitness effects. Furthermore, the bulk of these nominal 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. In prior studies, these variants have been described as “mini-drivers” \cite{26456849} and “deleterious passengers” \cite{23388632}.

Conceptually, variants can be classified into three categories based on their impact on tumor cell fitness: positively-selected driver variants, neutrally-selected passenger variants, and negatively-selected deleterious passenger variants. This broad classification can be further refined by considering ascertainment-bias and the putative molecular impact of different variants (**Fig 1**). Previous power analyses \cite{24390350} suggest that existing cohort sizes support the identification of 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 can also provide potential fitness advantage to tumor cells. 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). The relevance of molecular functional impact is firmly established for driver mutations - positively-selected variants promoting tumor growth. However, rapid accumulation of weak and strong deleterious passengers, which undergo negative selection, could adversely affect the fitness of tumor cell \cite{23388632. Moreover, majority of low impact and some

high functional impact variants may alter tumor gene expression or activity in ways that are not ultimately relevant for tumor fitness; hence, these variants will be under neutral selection.

Impactful passengers and their prevalence

In this work, we leveraged the PCAWG variant dataset to assess the molecular consequence of nominal passenger variants in 37 cancer histological subtypes. We build on existing tools \cite{25273974} to annotate and score the predicted molecular impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset. The integration of annotation and impact score allows for the quantification of overall molecular functional impact of variants occupying different genomic elements.

A closer inspection of the pan-cancer impact score distributions for non-coding variants demonstrate 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 molecular functional impact regime corresponds to what we term *impactful nominal passengers.* This intermediate functional impact category could include undiscovered drivers (strong & weak) as well as potentially deleterious passengers (**Fig 2a**).

According to a null 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 molecular impact burden in certain cancers is concentrated in particular gene categories. This is easiest to understand in terms of coding loss-of-function (LoF) variants, where the molecular impact is most intuitive. For instance, as a measure of the molecular impact of both driver and non-driver loss of function SNVs, we examined the fraction of deleterious LoFs affecting genes across four categories of cancer-related functional annotation (**Fig 2d**). Driver LoF variants showed significant enrichment in each category of cancer-related genes compared to a random (shuffled-variant) control ($p < 0.001$). Conversely, non-driver LoF SNVs displayed depletion in each of these categories ($p < 0.001$). In addition, 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 LoFs *in utero*, our observations suggest that both driver and non-driver LoF mutations exert molecular functional impact.

Similar to LoF variants, we can also quantify the overall burden of the noncoding SNVs in a cancer genome. However, for majority of noncoding variant, functional impact is less easy to gauge. For instance, noncoding and coding variants occupying the terminal region of the gene or undergoing alternatively splicing, will have little functional consequence. In contrast, transcription factor binding site (TFBS) variants are among the noncoding variants, where molecular impact is clearly manifested through the creation or destruction of TF binding motifs (gain or loss of motif). In both cases (gain or loss), we

observed significant differential burdening of TFBS among different cancer cohorts. For instance, we detect significant enrichment of high impact variants creating new motifs in various TFs including GATA, PRRX2 and SOX10 (**Fig 3b**) across major cancer types, compared with uniform expectation. Similarly, high impact variants breaking motifs, were highly enriched in TFs such as IRF, POU2F2, NR3C1and STAT (**Fig 3f**) in the majority of cohorts. This selective enrichment or depletion suggests distinct alteration profiles associated with different components of regulatory networks in various cancers.

Furthermore, for a particular TF family, one can identify their target genes affected due to bias towards creation or disruption of specific motifs in their regulatory elements (promoters and enhancers). For instance, TERT shows the largest alteration bias for ETS motif creation across a variety of cancer types (Fig 3d), with other genes (such as BCL6) showing a similar bias, albeit in fewer cancers. Interestingly, ETS motifs appear to show a systematic bias towards motif creation and loss, whereas MYC-family motif alterations show alteration biases for motif creation only (Fig. 3d). Furthermore, enrichment of SNVs in selective TF motifs leads to gain and break events in promoter that significantly perturb the overall downstream gene expression (**Fig 3g**). For example, ETS family transcription factor at the regulatory region of IRF4 and PSIP1 gene displayed a strong motif creation bias and a significant change in gene expression (with p-value IRF4=0.001 and p-value PSIP1=0.019).

Signature Analysis

The disproportionate functional load on certain TFs in cancers can be related to an underlying mutational spectrum (signature) influencing their binding sites. For instance, the mutational spectrum of motif breaking events observed in SP1 TF binding sites (TFBS) suggests 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 mutational profiles. Similarly, comparing the signature composition of low and high impact SNVs in certain cancer-cohorts can help us to distinguish between mutational processes that generate distinct variant impact classes. We observed distinct signature distributions for the low and high impact non-coding passengers for multiple cancer cohorts including Liver-HCC, Prost-AdenoCA and Kidney-RCC. For instance, in the Kidney-RCC cohort, although the majority of passenger variants 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 a higher percentage of high impact non-coding passengers (**Fig4c**). Our findings suggest that various mutational processes shape and disproportionally burden cancer genomes.

Overall variant impact

For a uniform mutation distribution, we would expect that the fraction of *impactful variants* will remain constant as one accumulates large number of mutations in a given cancer sample. In contrast, we observed that as we acquire more SNVs in cancer, the fraction of impactful mutations decreases. This suggests that earlier variants tend to be impactful, and drive the cancer progression. Conversely, later variants are more likely to be random, i.e. collateral damage. This trend is particularly strong in CNS medulloblastoma ($p < 4e-8$, Bonferroni's correction), lung adenocarcinoma ($p<3e-4$, Bonferroni's correction), and a few other cancers (**Fig 2c).**

Additionally, we sought to examine whether cumulative molecular impact of variants can be associated with tumor initiation and progression. Therefore, we performed survival analysis to see if somatic molecular impact burden – the mean GERP of somatic passenger variants per patient – predicted patient survival within individual cancer subtypes. Furthermore, patient age at diagnosis and total number of mutations were used as covariates in the survival analysis. We obtained significant correlation between overall molecular impact burden and survivability in two cancer subtypes after multiple test correction. For instance, we observed that somatic mutation burden predicted substantially better patient survival in lymphocytic leukemia (Lymph-CLL, p-value 0.00023) and ovary adenocarcinoma (Ovary-AdenoCA, p-value 0.0020) (**Fig5d**). The prolonged survival of high mean GERP patients in these subtypes is consistent with the possibility that an important subset of mutations at conserved positions are deleterious to tumor cells and benefit the patient.

In addition to SNVs, we also analyzed the annotation and overall impact of structural variants (SVs) in the PCAWG dataset. We compared the pattern of somatic SV enrichment in cancer genomes with those from germline. First, we observed, that as expected, somatic SVs were more enriched among functional regions compared to germline SVs, because the latter ones will be under negative selection for disrupting functional regions. Furthermore, we observed a distinct pattern of enrichment for SVs that split a functional element versus those that engulf it. As has been previously noted, there is greater enrichment of germline SVs that engulf an entire functional element rather than for those break a functional element partially. Furthermore, we observed the same pattern for somatic SVs. This is contrary to what one would expect from a purely randomized model, and suggests some form of selection. Finally, we also quantified the functional impact of somatic SVs across various cancer-types. A close inspection of SV and SNV impact scores suggest that certain cancer subtypes tend to harbor large number of high impact SVs, while others were more burdened with high impact SNVs. Many of these correlations have previously been observed \cite{24071851}. For example, it is known that ovarian cancer tends to be associated with driver SVs, whereas clear cell kidney cancer is often driven by SNVs. However, we also find new associations, such as the predominance of high impact SVs compared to SNVs in the bone leiomyoma cohort.

Subclonality and impact score

Furthermore, we also explored the role of impactful variants in cancer evolution by analyzing variants in the context of their associated tumor sub-clone. One might hypothesize that high impact mutations achieve greater prevalence in tumor cells if they are advantageous to the tumor, and a lower prevalence if deleterious. Interestingly, there is evidence to corroborate this hypothesis. We observed that high impact passenger variants in coding regions have greater prevalence among parental subclones (**Fig 5a**) – an effect driven by high impact nominal passenger SNVs in tumor suppressor and apoptotic genes (**Fig 5a**). In contrast, high impact passenger SNVs in oncogenes appear slightly depleted. Similarly, impactful SNVs in DNA repair genes and cell cycle genes are depleted in early subclones (**Fig 5a**). Furthermore, we also observe low heterogeneity in prevalence among higher impact variants compared to lower impact variants. This observation is consistent for both coding and non-coding variants (**Fig 5c**).

Estimating number of weak drivers and deleterious passenger variants

Finally, we estimated the frequency of weak drivers and deleterious passengers in various cancer cohorts. These variants tend to have small effect sizes and current datasets are underpowered to detect them individually. In addition to increasing the sample size, power to detect these variants can be increased by testing for their combined effect. In the context of germline variants, such approach has revealed that, while variants found significant in current GWAS studies only explain a small proportion of total heritability of quantitative traits \cite{20562875,19571811}, the combined analysis of all common variants can account for most of the missing heritability through additive effects. Thus, we investigate the combined effect of somatic nominal passengers by estimating their additive variance explained as a proportion of the total variance observed in a cancer cohort and the corresponding null model which preserves mutational signatures.

We treat the observed (cancerous) versus null model(non-cancerous) labels as a discrete phenotypic trait, and use a liability scale to estimate the additive variance as in genetic disease models \cite{19571811}. The additive variance is expected to form a component of the heritability of a subclone's fitness, although non-additive effects are expected to play a significant role also in the context of asexual heritability in cancer evolution (supplemental note). We use a linear model with random effects, in which the effect of each SNV is sampled independently from a normal distribution with a common variance parameter, σ_A^2, which estimates the additive variance \cite{20562875,21167468}. Furthermore, we constrain the SNVs in each gene to have a common effect size (creating a 'pooled' model), and estimate σ_A^2 from the variance of the random gene-level effects. Using this approach, we identify low-medium impact somatic nominal passengers in non-coding regions to account for significant

proportions of the total variance observed in Liver Carcinoma (LC ~24.2%) and Skin Melanoma (SM ~42.9%). Furthermore, we estimate a lower bound on the number of these nominal passengers contributing causally to the cancer phenotype, by determining the smallest subset of SNVs necessary to account for an equivalent proportion of the total variance when they are incrementally added to the model according to their individual effect sizes. This approach leads to estimates of at least 225 and 140 weak drivers in LC and SM respectively (an average of 1.2 and 10.9 per tumor).

Similarly, we also employed a conservation based metric to estimate the number of deleterious passenger mutations, which are removed during tumor progression. We hypothesized that a subset of somatic variants will be deleterious to cancer cells, when they are highly conserved (high GERP score), recessively deleterious (unmasked by complementary deletions) and affect regulators of genes essential to cell survival. We observed 2% fewer of such non-coding variants compared to the expected fraction from the null model. One could interpret this as removal of 2% of conserved mutations (median of 48 mutations), corresponding to noncoding deleterious passenger mutations per tumor. Moreover, this shortfall was particularly pronounced in the noncoding regulators of essential genes in haploid regions (17%) and most intense at promoters (32%).

Discussion

Previous studies \cite{20562875} related to the missing heritability problem in GWAS, indicate the cumulative effect of SNVs can explain the majority of this missing associations. Similarly, here we investigate whether the cumulative molecular impact of many weak somatic SNVs can have a meaningful role in cancer progression. Intuitively, tumor cells must maintain function of some minimal set of essential genes in order to achieve homeostasis. It is plausible that the 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{PARP inhibitor}. Similarly, increased mutation counts in coding genes or regions relevant for splicing increase the antigenic cross-section of tumor cells, potentially making them vulnerable to immune surveillance \cite{}. Conversely, any variant that optimizes cell-division at the expense of organism-supporting functions is expected to have a small positive effect on tumor fitness that may be challenging to detect. Moreover, certain variants through their complex genetic regulatory interactions may moderately increase the expression levels of canonical oncogenes. It has been proposed that these weak undiscovered driver variants benefit tumor growth and have a small associated positive selection.

In this work, several observations support the notion that some nominal passenger variants undergo weak selection. First, we observed overall enrichment and depletion of nominal passenger

variants among TSGs and oncogenes, respectively. An interpretation of this finding is that passenger variants in tumor suppressor genes have weak driver activity, while passenger variants in oncogenes impair oncogenic activity to the detriment to tumor fitness. Similarly, our finding of depletion of nominal passenger variants among DNA repair and cell cycle genes may indicate that high impact variants affecting these genes decrease tumor cell survival in relation to greater mutational burden. Consistent with a possible deleterious effect of passenger variants on tumor growth, in some cancer subtypes, the most mutated tumors have a lower fraction of impactful variants. This may relate to the aggregate impact of passenger variants becoming more deleterious at higher mutation loads. Alternatively, a fixed number of undiscovered drivers may become diluted by neutral passengers at higher mutation counts. Our LoF mutation analysis indicates 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 \cite{}. The aggregate fitness impact of nominal passenger variants may help explain why patient survival times are correlated with functional impact load in select cancer subtypes. Finally, using the additive variance model, we provide a conservative estimation of the number of weak drivers and deleterious passengers in various cancer cohorts. 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).