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. Furthermore, we adapt an additive effects model from complex trait studies to show that the differential presence/absence of low-medium impact variants across the genome is significantly predictive of observed cancer genotypes against a neutral (null) model. We further use the additive effects model to provide a conservative estimate of the number of weak drivers and deleterious passengers in different cancer cohorts. Finally, we delineate multiple suggestive evidences that correlate the overall burdening of cancer mutations with 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 coding and 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 variant data set, a few thousand ($< 5/\text{tumor}^1$) \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. 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{2338863}, respectively.

The identification of a few cancer driver mutations associated with a given cancer is analogous to genome-wide association studies (GWAS) that implicate handful of moderate-effect variants in a complex disease. Moreover, these modest number of variants explain only a small proportion of the predicted genetic variance, thus contributing to the "missing heritability" problem in GWAS \cite{20562875,19571811}. However, aggregating remaining variants with weak-effect explains a significant part of "missing heritability" and is predictive of disease risk. A recently proposed "omnigenic model" takes this logic a step further, arguing that the majority of complex traits are driven by a large number of regulatory variants with small effects. Although these models for complex disease are intriguing, they are also controversial, and further studies are required to confirm or refute them. Nonetheless, in the context of cancer, such models make a case to investigate the cumulative effect of nominal passengers and understand their role in cancer progression. Thus, we leverage the comprehensive PCAWG variant dataset to address these key questions. We built 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. 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. We also found that overall functional impact correlates with patient survival, subclonal architecture, and tumor heterogeneity. Finally, we observed suggestive evidences, which are consistent with the notion that aggregated subsets of functionally impactful passenger variants may confer weak selection avenues to tumor cells.

Aggregated nominal passengers and additive variance

We first adapted an additive effects model originally used in complex trait analysis to test for the existence of aggregated effects of non-driver mutations on tumorigenesis, and the relative size of these aggregated effects in relation to known drivers \cite{20562875,21167468}. We created a balanced sample of observed tumor genotypes and match neutral (null) model samples, where the latter preserve the mutational signatures and local mutation rates of the observed samples. We tested the ability of this model to predict cancerous from null samples using variants appearing in at least two samples. We threshold variants at a sequence of functional impact levels (using Funseq impact scores \cite{24092746}; See Supplemental Note). The additive effects model associated an effect with each SNV in a linear model, and considers they are sampled from a normal distribution. The variance of this distribution is optimized as a hyper-parameter along with the variance of residuals using restricted maximum-likelihood (REML) \cite{21167468}, while the predictive power of the model can be measured by the additive variance, which is equivalent to σ_A^2. We test 8 tumors using this model, and observed that the low & medium impact variants have significant predictive power in 7/8 of the tumors at an FDR<0.001 when non-driver variants are included from both coding and non-coding regions. In contrast, at this threshold, we observed significant predictive power in only 3 tumors, while considering variants in only the non-coding regions. In a further 2 tumors, non-coding low-medium impact variants show significance in the non-coding only regions at FDR<0.1 (Supp Fig. X).

We observed that across tumors, the predictive power of non-drivers is optimized at a higher impact threshold in coding regions than in non-coding regions, suggesting an enrichment of medium impact SNVs with non-neutral effects in coding regions. Further, we compared a variation of the additive effects model in which variants are pooled at the gene level, which demonstrated increased predictive power (Fig. 1A), and showed that, as a whole, the non-drivers mutations increase the predictive power by ~X% relative to known drivers.

Impactful passengers and their prevalence

If these nominal passenger variants do indeed exert a combined effect on tumor cell fitness, one would expect that this effect is mediated through their molecular functional impact. Therefore, we surveyed the

putative molecular functional impact distribution of somatic variants in different cancer genomes. The molecular functional impact distribution varies among different cancer types and for different genomic elements. A closer inspection of the pan-cancer impact score distributions for non-coding variants demonstrated 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 is most likely to include undiscovered drivers (strong & weak) as well as potentially deleterious passengers (**Fig 1b**).

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 1c).**

Overall variant impact

Furthermore, one might assume that the overall burden of nominal passengers in a cancer genome will be uniformly distributed across different functional elements and among different gene categories. In contrast, we observe that the predicted 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 putative 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 2a**). As expected, 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 ($p < 0.001$).

Additionally, we sought to examine whether the observed 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) (**Fig2b**). 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, large structural variations (SVs) also play important role in cancer progression. Thus, we analyzed the overall burden of structural variants (SVs) in various genomic elements and 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, which is contrary to what one would expect from a purely randomized model.

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.

TF binding landscape

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, predicted molecular 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 3a**) 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 3b**) 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. 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 3c**). For example, ETS family transcription factor at the regulatory region of TERT and PIM1 gene displayed a strong motif creation bias and a significant change in gene expression (with p-value TERT=0.001 and p-value PIM1=0.019).

Signature Analysis

The differential burdening of various genomic subsystems can be further explained by closely inspecting the underlying mutational processes generating SNVs. These mutational processes are considered to play an important role during tumor growth through their frequent occurrence among specific tri-nucleotides, genes and regulatory elements. For instance, it is expected that mutational processes targeting stop codons will be highly correlated with the number of nonsense mutations identified across different samples. Indeed, we found that although we often observe a high correlation between the mutation pattern and the number of expected nonsense mutations within cancer types, these correlations are highly heterogeneous among different cancer cohorts. For instance, we observed a high degree of correlation between mutation spectrum profile and the number of nonsense mutations within melanoma and colorectal adenocarcinoma cohorts. However, these correlations are relatively lower among samples in melanoma compared to colorectal cancer cohort.

Similarly, 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 (**Fig4c**). Moreover, we also observed cancers showing microsatellite instability (MSI) due to failure of DNA mismatch repair, have a higher percentage of high impact non-coding passengers (**Fig4d**). Our findings suggest that various mutational processes shape and disproportionally burden cancer genomes.

Subclonal architecture and impact score

Furthermore, we also explored the role of impactful passenger 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 5b**).

We employed a similar analysis using variant allele frequency (VAF) to explore whether passenger variants with high functional impact correlates with their underlying conservation score. Highly conserved positions (i.e. those with high GERP) are expected to be important for organismal fitness, as polymorphisms at those positions could potentially hurt cellular function and in other cases because polymorphisms at those positions could potentially 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, among nominal passengers, we observed that increasing impact scores correlated with decreasing VAF.

Categorizing nominal passenger variants

Through our analysis of the molecular functional impact of nominal passenger variants, we observed effects that are suggestive of their impact on tumor cell fitness. Conceptually, variants can be classified into three categories based on their impact on tumor cell fitness: drivers with positive selective effects, passengers with neutral selective effects, and deleterious passengers with negative selective effects. 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: any positively or negatively selected variants will 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, a 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 undergo neutral evolution.

Estimating number of weak drivers and deleterious passenger variants

In the context of this conceptual categorization of variants in cancer, we used the additive effects model to estimate the frequency of weak drivers and deleterious passengers in various cancer cohorts through their combined ability to predict cancerous from matched neutral samples. As observed, these variants tend to have small effect sizes and current datasets are underpowered to detect them individually. However, we can estimate a lower bound on the number of the nominal passengers with non-neutral effects by determining the smallest subset of SNVs needed to reach the same predictive accuracy (measured using σ_A^2) determined to be significant when all nominal passengers collectively (See Supplemental Note). Further, having estimated σ_A^2, we find the maximum a-posteriori predictor for the effect of each individual SNV, and use the predicted effect signs to estimate the number of weak drivers and deleterious passengers per tumor across the smallest subset. A conservative estimate of the number of deleterious passengers removed can be made by comparing the mean estimate in the neutral samples with that in the observed tumors. In general, we observe that the number of deleterious passengers removed is predicted to exceed the number of weak drivers across most tumors, with a pan-cancer average of ~18 weak drivers per tumor, and ~60 deleterious passengers removed (Fig. 6B). We also observe differential proportions across tumors, where Skin-Melanoma is both highly mutated, and predicted to have a large proportion of removed deleterious passengers. We further analyze the enrichment of functional gene categories within the smallest subset of weak drivers and deleterious passengers, and show it to be enriched in essential genes, and specific gene categories for particular tumors (Supp. Fig. X).

Similarly, we also employed a conservation based metric to estimate the number of deleterious passenger mutations, which are removed during tumor progression. To estimate the number of removed noncoding deleterious passengers per tumor, we compared the observed number of high-impact noncoding mutations with the number expected under a neutral model. We observed a slight (2%)

depletion in high-impact mutations in the observed mutation set versus the null, corresponding to a median of 48 high-impact noncoding mutations removed per tumor. This depletion was most pronounced at the promoters of essential genes in genomic regions suffering loss-of-heterozygosity (32%). Details of this analysis are described in Supplemental Methods X.X.

Discussion

Previous studies \cite{20562875} related to the missing heritability problem in GWAS, indicate the cumulative effect of SNVs can explain the majority of 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 estimate of the number of weak drivers and deleterious passengers in various cancer cohorts.

Despite these suggestive evidences, we acknowledge that caveats associated with current background models and lack of our understanding related to different mutational processes might be confounding factor in some of observation. Thus, although intriguing, one would need to perform follow up experiments and further genomic analyses to confirm the presence of weak drivers and deleterious passengers. 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).