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 leveraged the exhaustive variant data from PCAWG to ascertain the molecular functional impact of each variant, including nominal passengers. This allowed us to decipher their overall impact uniformly over different genomic elements, both coding and non-coding. The functional impact distribution of PCAWG mutations shows that, in addition to high and low impact variants, there is a group of medium-impact nominal passenger variants predicted to influence gene expression or activity. Furthermore, we found that functional impact relates to the underlying mutational signature: different signatures confer contrasting impact, differentially affecting distinct regulatory subsystems and different categories of genes. Also, we find that functional impact varies based on subclonal architecture (i.e. early vs. late mutations) and can be related to patient survival. Subsequently, we adapted an additive effects model derived from complex trait studies to show that aggregating nominal passenger variants provide significant predictive ability for cancer phenotypes beyond the characterized driver mutations. We further used the additive effects model to provide a conservative estimate on the number of weak drivers and deleterious passengers in different cancer cohorts. Finally, we delineate multiple lines of evidence 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, i.e. positively selected variants that favor tumor growth, by recurrence based driver detection methods. The remaining ~99% of SNVs are termed 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.

It is interesting to note that in a cancer genome, the presence of few key variants (with high positive fitness effects) and large numbers of passengers (with weak or neutral fitness effects) is analogous to prior observations in genome-wide association studies (GWAS) that implicate a handful of variants that significantly influence complex traits. These modest numbers of variants explain only a small proportion of the genetic variance, thus contributing to the “missing heritability” problem in GWAS \cite{20562875,19571811}. However, it has been shown that aggregating remaining variants with weak effects can explain a significant part of any “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 influenced by thousands of variants with individually small effects*** \cite{28622516}. Although these models for complex disease are intriguing, they are also controversial, and further studies are required to test them. Nonetheless, these models highlight the importance of investigating the cumulative

effect of nominal passenger mutations in cancer to understand their potential role in cancer progression.

Overall effects of nominal passengers and additive variance

To estimate the overall effects of nominal passengers on tumorigenesis, we first adapted an additive effects model, originally used in complex trait analysis, to quantify the relative size of these aggregated effects in relation to known drivers \cite{20562875,21167468}. With a number of caveats regarding interpretation arising due to differences between germ-line and cancer evolutionary processes (see supplemental note X.b), we tested the ability of this model to predict cancerous from null samples as a binary phenotypic trait. Briefly, we created a balanced dataset of observed tumor and matched neutral (null) model samples, where the latter preserve the mutational signatures and local mutation rates of the observed samples (see Supplemental note Xa). Subsequently, we apply different thresholds on predicted molecular functional impact levels (using Funseq impact scores \cite{24092746}; See Supplemental Note) to identify different sets of variants. Using a linear model, for each SNV the additive effects model associates a positive or negative effect (coefficient), considering them to be sampled from a normal distribution. The model has the form $y_j = \mu + \sum_i z_{ij} u_i + e_j$, where y_j is the phenotype (0/1) of sample j, z_{ij} is the normalized SNV dosage of SNV i in sample j (z-score), e_j is the residual effect for sample j, and μ is the mean phenotype. The u_i's are normally distributed with variance σ_A^2/m, where σ_A^2 is the additive variance and m is the number of SNVs, and the e_j's are normally distributed with variance σ_E^2. The variance of y is denoted σ_P^2 (the 'phenotypic' variance), where $\sigma_P^2 = \sigma_A^2 + \sigma_E^2$. The hyper-parameters σ_A^2 and σ_E^2 are optimized using restricted maximum-likelihood (REML) \cite{21167468}, and the predictive power of the model can be summarized by σ_A^2/σ_P^2.

We applied this model in 8 cancer cohorts having sample size greater than 100. Across cancers, we found that the nominal passengers predicted a large fraction of the variance (64.5% median), a significant fraction of which remained even when coding variants were excluded (57.9%) (see Fig/Supp. Fig X). We compared this with a model including all known drivers, which predict ~52.5% of the variance. The ability of the nominal passengers to achieve higher predictive accuracy in many tumor types implies that these variants must contain additional information to the known drivers. However, there may be mutual information shared between the known drivers and passengers, for instance due to epistatic effects. Furthermore, we

observed that across tumor types, the predicted variance per nominal passenger increases with impact score for both coding and non-coding variants, with the increase being stronger for coding variants (Fig. X). However, the fact that the largest amount of variance is explained at the lowest impact threshold suggests that weak drivers and deleterious passengers at all impact levels might have functional consequence (Table X). Moreover, their effect sizes may become detectable individually with the increased power of larger datasets.

Overall functional impact

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 functional impact distribution of somatic variants in different cancer genomes. The predicted 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 are presumably enriched with high-impact putative drivers 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 potentially includes undiscovered drivers (strong & weak) as well as potentially deleterious passengers (**Fig 1a**).

Subsequently, we investigated whether frequency of these *impactful nominal passengers* in a cancer cohort is proportionate to its total mutational burden. For a uniform mutation distribution, we would expect that the fraction of *impactful nominal passengers* 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 passengers decreases. 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 cancer cohorts(**Fig 1b).**

In addition to SNVs, large structural variations (SVs) also play important role in cancer progression. Thus, we quantified the putative functional impact of SVs (deletions and duplications). A close inspection of both 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 large deletions play role of drivers in ovarian cancer, whereas clear cell kidney cancer is often driven by SNVs. However, we also find new associations, such as the predominance of high impact large deletions compared to impactful SNVs in the bone leiomyoma cohort.

Burdening of different genomic elements

Simplistically, 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 regulatory regions and gene categories. This is easiest to understand in terms of coding loss-of-function (LoF) variants, where the putative molecular impact is most intuitive. Consequently, 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 ($p < 0.001$). As expected, driver, non-driver, and random LoFs were all enriched in comparison to germline LoFs ($p < 0.001$).

Similar to LoF variants, we can also quantify the overall burden of the noncoding SNVs in a cancer genome. However, for the majority of noncoding variants, predicted molecular functional impact is less easy to gauge. For instance, coding and noncoding variants occupying the terminal region of the gene or intronic regions will most likely 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 transcription factor (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 detected significant enrichment of high impact variants creating new motifs in various TFs including GATA, PRRX2 and SOX10 (**Fig 2b**) 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 2b**) 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). Other genes (such as BCL6) showed a similar bias, albeit in fewer cancers. Moreover, enrichment of SNVs in selective TF motifs leads to gain and break events in promoter that significantly perturb the overall downstream gene expression (**Fig 2c**). 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) (supplement X).

Finally, we also 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. As expected, we observed that as 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 \cite{26432246}. Moreover, we observed the same pattern for somatic SVs, which is contrary to what one would expect from a purely random background model.

Signature Analysis

The differential burdening of various genomic elements can be attributed to either the underlying mutational processes or selection on variants occupying these elements. Thus, we closely inspected the underlying mutational signatures generating SNVs in coding and non-coding regions of cancer genomes. One would expect that mutational processes influencing stop codons will highly correlate with the number of LoFs observed in a cancer sample. Indeed, we were able to identify a high correlation between the mutation spectrum and the number of expected LoFs within some cancer types. However, these correlations are highly heterogeneous among different cancer cohorts and the number of LoF mutations might be often driven by other factors, such as

tumor size and total number of mutations. For example, even though Skin-melanoma cancers contain the highest ratio of stop-codon triplets (TAG, TAA, CTA, TTA and TCA), they also have the smallest number of observed compared to expected LoFs, showing a negative correlation between mutation patterns and the number of LoFs. On the other hand, Lung-SCC and Esophageal adenocarcinoma cohorts provide a very high correlation between mutation patterns and the number of LoFs per tumor sample. However, far fewer LoFs are generated in melanoma than in colorectal cancer in relation to the mutational spectrum.

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 TFBS suggests major contribution from C>T and C>A mutations (**Fig 4b**). In contrast, motif-breaking events at TFBS of HDAC2 and EWSR1 have relatively uniform mutational profiles. In addition, 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**). Collectively, these findings suggest that various mutational processes shape and disproportionally burden cancer genomes.

Subclonal architecture and cancer progression

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 nominal passengers in DNA repair genes and cell cycle genes are depleted in early subclones (**Fig 5a**), potentially

showing evidence for providing a mutational burden. We obtained similar results when we simply categorized mutations based on variant allele frequency (VAF) (Fig. SXXX).

Moreover, we also measured divergence in variant allele frequency (VAF) to indirectly quantify tumor prevalence heterogeneity in the pan-cancer data. As expected, we observe lower heterogeneity among high impact nominal passenger SNVs. This observation is consistent for both coding and non-coding nominal passenger variants (**Fig 5b**). Moreover, these observations are not reliant on any particular randomized model and so will be robust to potential inaccuracies in the null model.

We next investigated how the VAF of variants are related to the degree of evolutionary constraint on the affected nucleotide position (using GERP score\cite{}). In non-rearranged genomic intervals, the VAF of a mutation is proportional to the fraction of tumor cells bearing that mutation. Conceptually, variants that increase tumor cell fitness should lead to greater proliferation of the tumor cells containing them and should therefore tend to be present at increased VAF, when averaged across many samples. Similarly, variants that decrease tumor cell fitness should tend to be present at decreased VAF. In general, we expect that disruption of more conserved nucleotides would be more likely to interfere with cellular processes and reduce cellular fitness. An exception is in cancer driver genes, where disruption of conserved nucleostides could be oncogenic, increasing cellular proliferative potential. We find that within driver genes and their regulators, variants that disrupt more conserved positons tend to have higher VAFs. This trend remains true after excluding SNVs that have been individually called as driver variants, suggesting the existence of latent driver variants within driver genes. We also find that outside of driver genes, variants that disrupt more conserved positions tend to have lower VAFs.

Additionally, we sought to examine whether the observed molecular impact of variants can be associated with clinical outcomes. Therefore, we performed survival analysis to see if somatic molecular impact burden – here measured as the mean GERP of somatic passenger variants per patient – predicted patient survival within individual cancer subtypes. Patient age at diagnosis and total number of mutations were used as covariates in the survival analysis. We obtained significant correlations between overall molecular impact burden and survivability in two cancer subtypes after multiple test correction. Specifically, 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. The use of *average* impact rather than summed impact ensures that these results do not simply reflect more advanced progression of the patient at the time of sequencing.

Categorizing nominal passenger variants

Through our analysis of the molecular functional impact of nominal passenger variants, we observed multiple manifestations that are suggestive of nominal passenger's 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 6**). 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. This can be estimated to be the size of the smallest subset of SNVs needed to reach the same predictive accuracy (measured using σ_A^2) as when using all nominal passengers collectively (See Supplemental Note). Further, having estimated σ_A^2, we find the maximum a-posteriori estimate 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). These numbers are significantly higher than pan-cancer average of ~ 4.6 strong driver mutations.

We corroborate the quantification of deleterious passenger variants with two other methods: impact depletion-based and VAF deficit approaches. 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 impacted by loss-of-heterozygosity (32%). Orthogonally, we use VAF deficits to estimate the number of *retained* deleterious passenger mutations per tumor. Assuming conservatively that a deleterious passenger has as much negative impact on VAF as a driver has a positive impact on VAF, we calculate that there are on average 8.6 retained deleterious passenger variants per tumor. Since one deleterious passenger is unlikely to exert as large an effect on VAF as one discovered driver variant, the true number of retained deleterious passenger variants per tumor is likely larger.

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. In this work, we came across several orthogonal sources of evidence that suggest presence of weak positive and negative selective effect in a cancer genome.

First, we observe that functional impact distribution has a multi-modal characteristic with significant number of nominal passengers with intermediate functional impact. Furthermore, contrary to simple expectation, we observe lower amount of impactful nominal passengers with an increase in total mutation burden. This trend can be explained either by selection on impactful nominal passenger variants, or by changing mutational signatures. A selection-based explanation will be that negative selection on deleterious passengers becomes more pronounced at higher mutational loads, which tends to remove impactful nominal passengers. Additionally, we also observed strong correlation between differential functional burden and patient survival in certain cancer cohorts. These correlations can be also inferred as presence of weak selection. For instance, survival in specifically CLL might be prolonged by deleterious passenger variants because CLL is a particularly slow-growing tumor, such that the fitness cost of hitchhiking deleterious passengers may be at a magnitude more comparable to the overall tumor growth rate than in other tumors.

Second, we observe that various functional elements in a cancer genome are differentially burdened with distinct functional impact. To some extent, this can be associated with the operation of various signatures, which in itself is significant. However, in certain context this can be related to presence of weak negative selection. For instance, depletion of nominal passenger LoFs in key gene categories including essential and metabolic genes compared to a random expectation can be interpreted as presence of negative selection pressure.

Third, we also detect a differential functioning burdening between early and late subclones in a cancer. More specifically, we observed an overall enrichment and depletion of nominal passenger variants among TSGs and oncogenes, respectively. An interpretation of this finding is that nominal passengers in tumor suppressor genes may have potentially weak driver activity, while those in oncogenes impair oncogenic activity to the detriment to tumor fitness.

However, we note that difference in signatures between and early and late subclones can also contribute to these observed differences.

Finally, using an additive variance model, we show that aggregating nominal passengers in a cancer genome can provide significant predictive ability to distinguish cancer phenotype from non-cancerous ones. Moreover, this model can be also utilized to obtain a conservative estimate of the number of weak drivers and deleterious passengers in various cancer cohorts. We corroborate these estimates using a VAF-based method that does not rely on any particular randomized null.

We note that discussion of these selective effects is meaningful only in the realm of proper background(null) model. For instance, one can identify a role of positive or negative selection based on differences between observed and random expectation derived from a null model. However, this assumes that we apply an accurate randomized model to perform the comparison. In this work, we utilize a local background model that has been applied in other efforts in PCAWG including driver detection. However, our understanding of the underlying mutational processes and cancer genome structure is limited, thus, can hinder achieving an accurate null model. Nonetheless, these multiple set of observations are intriguing and further motivate follow up experiments and additional whole genome analyses to explore the role of weak drivers and deleterious passengers in cancer. 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.