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 "*putative passengers*") are inconsequential for tumorigenesis. In this study, we leveraged the comprehensive variant data from PCAWG to ascertain the molecular functional impact of each variant, including *putative passengers*. This allowed us to decipher their overall impact uniformly over different genomic elements. The functional impact distribution of PCAWG mutations shows that, in addition to high and low impact mutations, there is a group of medium-impact *putative passengers* predicted to influence gene expression or activity. Moreover, we found that functional impact relates to the underlying mutational signature: different signatures confer contrasting impact, differentially affecting distinct regulatory subsystems and 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. Furthermore, we adapted an additive effects model derived from complex trait studies to show that aggregating nominal passenger variants provide significant predictability for cancer phenotypes beyond the characterized driver mutations. We further used the additive effects model to provide a conservative estimate on the number of mutations with weak positive and negative fitness effects in different cancer cohorts.

Deleted: nominal

Formatted: Font:Italic

Deleted: nominal

Formatted: Font:Italic

Deleted: , both coding and non-coding.

Deleted: variants

Deleted: nominal

Formatted: Font:Italic

Deleted: weak drivers and deleterious passengers in different cancer cohorts. Finally, we delineate multiple lines of evidence that correlate the overall burdening of cancer

Deleted: the existence of both

Deleted: selection during tumor evolution.

Formatted: Level 1

Introduction

Previous studies have focused on characterizing variants occupying coding regions of cancer genomes[1]. 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[2], 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 somatic copy number alterations (SCNAs) 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/tumor[3]) 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 (referred as *putative passengers* in this work), with poorly understood molecular consequences and fitness effects. Recent studies have proposed that, among *putative passengers*, some may weakly affect tumor cell fitness by promoting or inhibiting tumor growth. In prior studies, these variants have been described as "mini-drivers"[4] and "deleterious passengers"[5], respectively.

It is interesting to note that in a cancer genome, the presence of few drivers (with high positive fitness effects) and large numbers of *putative passengers* (with weak or neutral fitness effects) is analogous to prior observations in genome-wide association studies (GWAS) that implicated a handful of variants influencing 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[6,7]. However, it has been shown that aggregating remaining variants with weak effects can explain a significant part of any "missing heritability"[6] and is predictive of disease risk[8]. 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[9]. 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 *putative passengers* to

Deleted: variants (CNVs

Deleted: ,

Deleted: variants that have not been found to be driver variants (i.e. nominal passenger variants),

Deleted: key variants

Formatted: Font:Italic

Deleted: that significantly influence

Deleted: nominal passenger mutations

understand their potential role in cancer. Furthermore, we can adapt this model to estimate frequencies of rare drivers, which might be misannotated as passengers using recurrence-based approaches due to limitation of current sample sizes.

Overall functional impact

If these *putative passenger* mutations 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 predicted 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 strong drivers and low impact neutral passengers, respectively. In contrast, the middle peak in the intermediate molecular functional impact regime corresponds to variants which may or may not have non-neutral effects. These medium impact variants potentially include undiscovered drivers (strong & weak positive effects) and deleterious passengers (strong & weak negative effects) (**Fig 1c**).

Subsequently, we investigated whether the frequency of medium-high impact *putative passengers* in a cancer cohort is proportionate to its total mutational burden. For a uniform mutation distribution, we expect that the fraction of these *putative passengers* would remain constant as cancer samples accumulate more mutations. In contrast, we observed that as a cancer acquires more SNVs, the fraction of medium and high impact *putative passengers* often 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 1d).**

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 (**Fig 1e**). Many of these correlations have previously been observed[12]. 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

Deleted: progression

Moved down [1]: Overall effects of nominal passengers and additive variance

Moved down [2]: , we first adapted an additive effects model[6,10], originally used in complex trait analysis, to quantify the relative size of these aggregated effects in relation to known drivers. With a number of caveats regarding interpretation arising due to differences between germline 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 (**Fig 1a**). 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[11]; 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 phen ... [2]

Deleted: To estimate the overall effects of nominal ... [1]

Deleted: passengers, for instance due to epistatic effects.

Moved down [3]: Furthermore, we observed that acr ... [3]

Deleted: nominal

Formatted: Font:Italic

Deleted: variants

Deleted: putative

Deleted: putative

Deleted: what we term *impactful nominal passengers.* ... [4]

Deleted: category

Deleted: includes

Deleted:) as well as potentially

Deleted:

Formatted: Indent: First line: 0.5"

Deleted: these *impactful nominal*

Deleted: would

Deleted: *impactful nominal*

Deleted: will

Deleted: one accumulates large number of

Deleted: in a given cancer sample.

Deleted: we acquire

Deleted: in cancer

Deleted: impactful

Formatted: Font:Italic

Deleted: <

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 *putative 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 six categories of cancer-related functional annotation (**Fig 2a**). As expected, driver LoF variants showed significant enrichment in four categories (cell cycle, cancer pathway, apoptosis & DNA repair) of cancer-related genes compared to a random (shuffled-variant) control ($p < 0.001$). Conversely, non-driver LoFs displayed depletion compared to random expectation, in each of these categories ($p < 0.001$). However, non-driver LoFs in metabolic and essential genes were slightly enriched compared to the random expectation.

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 SNVs, 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, molecular impact of transcription factor binding site (TFBS) variants 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.

Deleted: nominal

Formatted: Font:Italic

Deleted: In contrast

Deleted: Moreover, driver, non-driver, and random LoFs were all enriched in comparison to germline LoFs ($p < 0.001$).

Deleted: variants

Deleted: are among the noncoding variants where molecular impact

Furthermore, for a particular TF family, one can identify their target genes affected due to the bias towards creation or disruption of specific motifs in their regulatory elements (promoters and enhancers). For instance, the TERT gene shows the largest alteration bias for ETS motif creation across a variety of cancer types (**Fig 2c**). 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 promoters that significantly perturb the overall downstream gene expression (**Fig 2d**). 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 (**Fig 2e**). 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[13]. 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 random but biased 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. For instance, one would expect that mutational processes creating stop codons will highly correlate with the number of LoF variants observed in a cancer sample. Indeed, we were able to identify a high correlation between the mutation spectrum and the number of 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. For example, Lung–SCC and Esophageal adenocarcinoma cohorts exhibit a high correlation between their mutation pattern and the number of LoFs per tumor

Formatted: Not Highlight

Formatted: Not Highlight

Deleted:

Deleted: provided

sample (r=0.55 and 0.46 respectively) (see supplement table X). Other cancer cohorts such as colorectal adenocarcinoma and non-Hodgkin lymphomas were able to withhold the majority of their LoFs with the ratio of observed vs expected close to 1(**Fig 3a**).

Similarly, the disproportionate functional load on certain TFs in cancers can be related to an underlying mutational spectrum influencing their binding sites. This can be partially explained by the different nucleotide context among TF biding sites (TFBS). For instance, the mutational spectrum of motif breaking events observed in SP1 TFBS suggests major contribution from C>T and C>A mutation (**Fig 3b**). In contrast, motif-breaking events at the TFBS of HDAC2 and EWSR1 have relatively uniform mutation spectrum profiles. Based on the mutational context, we can further decompose all observed mutations into a linear combination of mutational signatures, which presumably represent the mutational processes (\cite{}). Every signature has varying influence on different cancer types and in a given cancer type, different signatures disproportionally burden the genome. Comparing the signature composition of low and high impact *putative passengers* in certain cancer-cohorts can help us to distinguish between mutational processes that generate distinct variant impact classes. For instance, in the Kidney-RCC cohort, although the majority of passenger variants can be explained by signature 39, high impact and low impact passengers have different proportion of signature 5 and signature 1(**Fig 3c**). We further generalized this analysis across multiple cohorts in PCAWG. Similar to Kideny-RCC cohort, we observed distinct signature distributions for the low and high impact non-coding *putative passengers* in Liver-HCC, Prost-AdenoCA, Eso-AdenoCA and Ovary-AdenoCA cohorts (**Fig 3c**). Collectively, these findings suggest that various mutational processes shape and disproportionally burden cancer genomes.

Subclonal architecture and cancer progression

Cancer is an evolutionary process, often characterized by the presence of different sub-clones. These can be further categorized as early and late subclones based on the overall subclonal architecture of a cancer sample. Thus, we explored the relative population of high and low impact *putative passengers* in different sub-clones of a tumor sample to decipher their progression during tumor evolution. Intuitively, 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. As expected, we observe this to be true among driver variants.

Formatted: Highlight

Deleted: Finally, even though Skin-melanoma cancers contain the highest ratio of stop-codon triplets - T(x>A)G, T(x>A)A, C(x>T)A, T(x>T)A and T(x>C)A - they showed a negative correlation between mutation pattern and the number of LoFs while presenting the lowest number of observed vs expected LoF mutations (**Fig 3a**).

Formatted: Highlight

Deleted: (signature)

Deleted: mutations

Formatted: Font color: Auto

Deleted: In addition, comparing the signature composition of low and high impact SNVs

Formatted: Highlight

Deleted: 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 (**Fig 3c**).

Formatted: Highlight

Deleted: 3d).

Formatted: Highlight

Formatted: Pattern: Clear, Highlight

Formatted: Indent: First line: 0.5"

Deleted: in

Deleted: nominal

Formatted: Font:Italic

However, interestingly, we observe that high impact *putative passengers* in coding regions have greater prevalence among parental subclones (**Fig 4a**) – an effect driven by high impact *putative passenger* SNVs in tumor suppressor and apoptotic genes (**Fig 4a**). In contrast, high impact *putative passenger* SNVs in oncogenes appear slightly depleted. Similarly, high impact *putative passengers* in DNA repair genes and cell cycle genes are depleted in early subclones (**Fig 4a**). We obtained similar results when we simply categorized mutations based on variant allele frequency (VAF) (supplement Fig X). Note that these subclones and VAF based analyses are not reliant on any particular randomized model and so will be robust to potential inaccuracies in the null model.

In non-rearranged genomic intervals, the VAF of a mutation is expected to be proportional to the fraction of tumor cells bearing that mutation. Previous studies have measured the divergence in VAFs to indirectly quantify heterogeneity in mutational burden among different sub-clones in a cancer. Here, we quantified this heterogeneity among low, medium and high impact *putative passengers* for different cancer cohorts. Overall, we observe lower mutational heterogeneity among high impact *putative passenger* SNVs. This observation is consistent for both coding and non-coding *putative passenger* variants (**Fig 4b**).

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 (with high GERP score[14]) would be more likely to interfere with cellular processes and reduce cellular fitness. An exception is in cancer driver genes, where disruption of conserved nucleotides could be oncogenic, increasing cellular proliferative potential (**Fig 4c**). We find that within driver genes and their regulators, variants that disrupt more conserved positons tend to have higher VAFs. This trend remains true even after excluding SNVs that have been individually called as driver variants, suggesting the existence of weak driver variants within driver genes. We also find that outside of driver genes, variants that disrupt more conserved positions tend to have lower VAFs.

Similar to the clonal status of a tumor, clinical outcomes such as survivability provides an alternative measure for tumor evolution. Therefore, we performed survival analysis to see if somatic molecular impact burden – here measured as the mean GERP of somatic nominal

Deleted: there is evidence to corroborate this hypothesis even among impactful passenger variants. We observed

Deleted: passenger variants

Deleted: nominal

Formatted: Font:Italic

Formatted: Font:Italic

Deleted: impactful nominal

Formatted: Font:Italic

Deleted: nominal

Formatted: Font:Italic

Deleted: As expected

Deleted: nominal

Formatted: Font:Italic

Deleted: nominal

Formatted: Font:Italic

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) (**Fig 4d**). The use of *average* impact rather than summed impact ensures that these results do not simply reflect more advanced progression (i.e. more mutations) of the cancer at the time of sequencing.

Deleted: 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.

Formatted: Indent: First line: 0"

Moved (insertion) [1]

Overall effects of nominal passengers and additive variance

Moved (insertion) [2]

In addition to comprehensive characterization of *putative passengers* in PCAWG, we also estimated the overall effects of *putative passengers* on tumorigenesis. To address this, we first adapted an additive effects model[6,10], originally used in complex trait analysis, to quantify the relative size of these aggregated effects in relation to known drivers. With a number of caveats regarding interpretation arising due to differences between germline 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 (**Fig 1a**). 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[11]; 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)[10], 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 **1b**; FDR<0.1 for all tests using a gene-level variant of the additive model except non-coding variants in CNS, Ovary and Prostate cancers, and FDR<0.001 for all tests using the basic (SNV-level) model, see supplemental table 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 nominal 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. **1c**). 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 (supplement table X). Moreover, their effect sizes may become detectable individually with the increased power of larger datasets.

Moved (insertion) [3]

Categorizing nominal passenger variants

Through our analysis of the molecular functional impact of nominal passenger variants, we observed multiple manifestations that are suggestive of *putative 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, nominal 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 5a**). Previous power analyses[15,16] 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).

Deleted: nominal

Formatted: Font:Italic

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 cells[5]. 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 the 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).

Deleted: this

Further, having estimated σ_A^2, we find the maximum a-posteriori estimate for the effect of each individual SNV, and use the effect signs from this estimate to then predict the number of weak drivers and deleterious passengers per tumor across the smallest subset (i.e. using positive effect for weak and negative effect for deleterious passengers). Next, A conservative estimate of the number of deleterious passengers *removed* can be made by comparing this prediction to the mean number estimated in the neutral samples. In general, we observe that the number of deleterious passengers removed is predicted to exceed the number of weak drivers across most tumors. The pan-cancer average of weak drivers per tumor falls in the range of 11.6(lower bound) to 16.1 (upper bound). Similarly, pan-cancer average of *remove*d deleterious passenger per patient is in the limits of 23.6 (lower bound) to 57.9 (upper bound) (**Fig 5b**). 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 is consistent with earlier prediction of the removed deleterious passenger frequency based on the additive effects model. Additionally, the observed depletion of high-impact mutations was most pronounced at the promoters of essential genes in genomic regions impacted by loss-of-heterozygosity (32%). Orthogonally, we used VAF deficits to estimate on average 8.6 *retained* deleterious passenger mutations per tumor. These are again conservative estimate, as we assume that latent drivers and deleterious passengers exert a VAF effect equal in magnitude to discovered drivers, when in fact, their true effect is likely smaller.

Discussion

Previous studies[6] related to the missing heritability problem in GWAS, indicate that the cumulative effect of SNPs 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 nominal passenger variants influencing these essential genes would be deleterious to tumor cells[5]. Similarly, 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. In this work, we comprehensively characterized *putative passengers* in the PCAWG dataset.

First, we evaluated the molecular functional impact of each variant in the PCAWG including *putative passengers*. We observed 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 *putative passengers* with an increase in total mutation burden. Additionally, we also observed strong correlation between differential functional burden and patient survival in certain cancer cohorts. These trends can be explained by mutational signature and their differences in *putative passengers* with varying impact level.

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

Deleted: In this work, we came across several orthogonal lines of evidence that suggest presence of weak positive and negative selective effect in different cancer genomes

Formatted: Font color: Text 1, Pattern: Clear (White)

Deleted: First, we observe that the

Deleted: nominal passengers with an increase in total mutation burden. This trend can be explained either by selection on impactful nominal passenger variants, or by changes in 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.

Formatted: Font color: R,G,B (33,33,33), Pattern: Clear (White)

Deleted: These correlations can be also inferred as presence of weak selection. For instance, prolonged survival of Lymph-CLL might be attributed to the presence of deleterious passenger variants. Lymph-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

Formatted: Font color: R,G,B (33,33,33), Pattern: Clear (White)

operation of various signatures, which in itself is interesting. However, in certain contexts this can be potentially related to presence of weak negative selection. For instance, depletion of nominal passenger LoFs in key gene categories including dna repair and cell cycle compared to a random expectation can be interpreted as presence of negative selection pressure. Interestingly, we do not observe such signal of weak negative selection among non-essential genes. This is consistent with prior studies suggesting role of negative selection in different cancers[5].

Third, we also detect a differential functional 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. A speculative interpretation of this finding can be that a subset of *putative 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 effects 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 *putative passengers* with weak positive and negative effect in various cancer cohorts.

We note that discussion of these selective effects is meaningful only in the context of a proper background (null) model. For instance, one can identify a role of positive or negative selection based on differences between an observed attribute and the corresponding 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 discovery. However, our understanding of the underlying mutational processes and genome structure of a tumor sample is limited, which can be a hindrance in achieving the accurate null model. Nonetheless, we our additive variance analysis suggest potential role of weak positive and negative selection among *putative passengers*. These observations further motivate follow up experiments and additional whole genome analyses to explore the role of weak *putative passengers* with weak (positive and negative) fitness effects in cancer. In conclusion, our work highlights that an important subset of somatic variants originally identified as *putative passengers* nonetheless show biologically and clinically relevant functional roles across a range of cancers.

Deleted: An

Deleted: is

Deleted: nominal

Formatted: Font:Italic

Deleted: .

Deleted: weak drivers and deleterious

Formatted: Font:Italic

Formatted: Pattern: Clear

Deleted: have delineated multiple set of intriguing observations suggesting

Deleted: nominal

Formatted: Font:Italic

Deleted: drivers and deleterious

Formatted: Font:Italic

Formatted: Font:Italic

References

1. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
2. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17,** 93–108 (2016).
3. Vogelstein, B. & Kinzler, K. W. The Path to Cancer — Three Strikes and You're Out. *N. Engl. J. Med.* **373,** 1895–1898 (2015).
4. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15,** 680–685 (2015).
5. 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).
6. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–9 (2010).
7. International Schizophrenia Consortium, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–52 (2009).
8. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29,** 150–159 (2013).
9. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
10. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
11. Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15,** 480 (2014).
12. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–33 (2013).
13. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
14. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–13 (2005).
15. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
16. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* (2017). doi:10.1038/nature22992

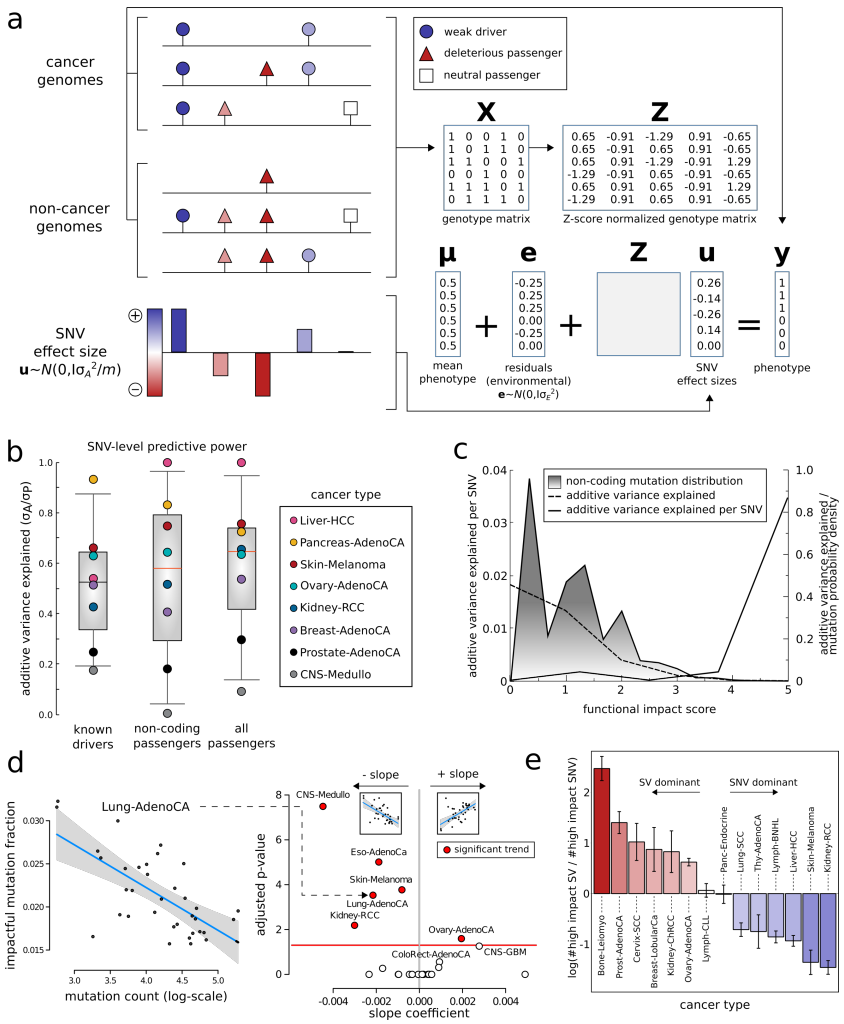

Figure 1: Additive effect and overall functional impact of PCAWG variants: *Additive effects model for nominal passengers*: The combined effects of many nominal passengers are modeled using a linear model, which predicts whether a genotype arises from an observed cancer sample or from a null (neutral) model (notation defined in text). The model is fitted by optimizing the hyper-parameter σ_A^2, and a test for significant combined effects of the nominal passengers is made by performing a log-likelihood ratio test against a restricted model which includes only **μ** and **e**. **b)** *Predictive power of known drivers and nominal passengers using the additive effects model*: Figure compares the maximum possible variance which can be explained using known drivers with the performance of the model from using either non-coding passengers or all nominal passengers. **c)** Functional impact distribution in noncoding region: three peaks correspond to low, medium and high impact variants. **d)** Correlation between number of impactful and total SNV frequencies for different cohorts. **e)** log ratio of high impact structural variants(SVs) and SNVs in different cancer cohorts.

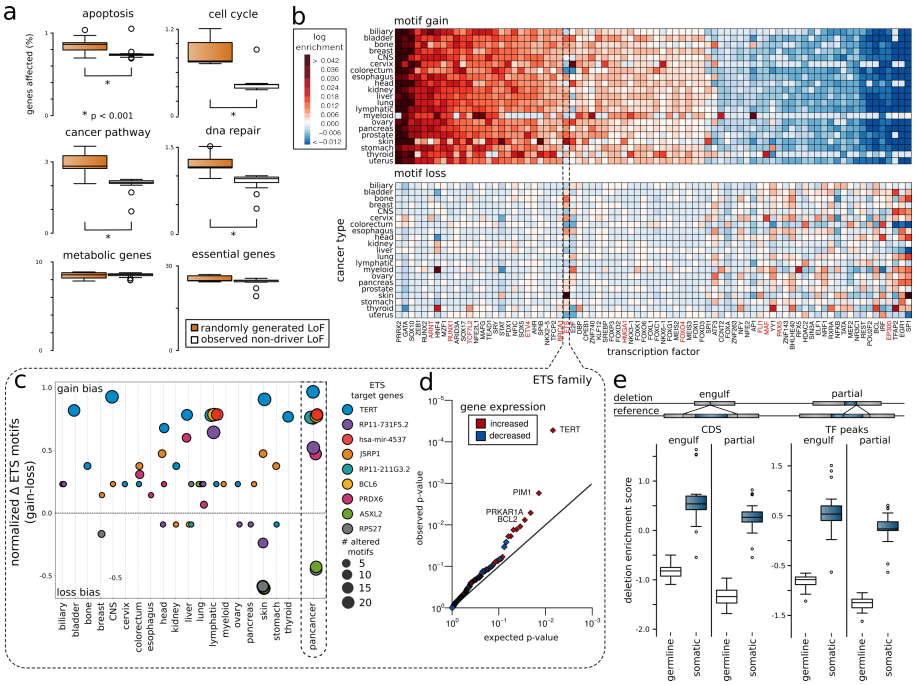

Figure 2: Overall functional burdening of different genomic elements: a) Percentage of genes in different gene categories (apoptosis, cell cycle, cancer pathway, dna repair, metabolic and essential genes) affected by non-driver LoFs in observed and random model, **b)** *Pan-cancer overview of TFs burdening:* Heat map presenting differential burdening of various TFs due to SNVs inducing motif breaking and motif gain events in different cohorts compared to the genomic background. **c)***target genes affected due to motif gain and loss in ETS transcription factor family*: genes such as TERT, RP17-731F5.2 and JSRP1 are affected due to gain of motif event, whereas ASXL2 and RPS27 are affected due to loss of motif event. **d)** q-q plot showing genes such as TERT, PIM1 and BCL2, which are differentially expressed due to gain of motif event in ETS TFs. **e)** enrichment of germline and somatic large deletions in coding region and transcription factor binding peaks. Large deletions can engulf or partially delete various genomic elements.

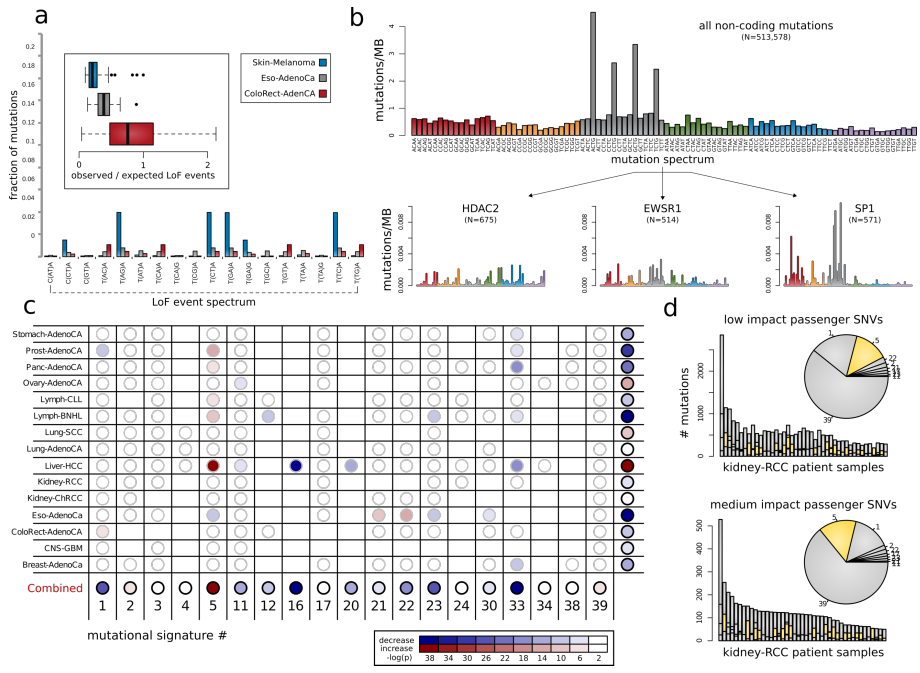

Figure 3. Mutational signatures associated with different categories of impactful variants: a) Differences in mutation spectrum leading to stop-coding triplets as a fraction of the total number of mutations per sample between three cancer cohorts: Colorectal Adenocarcinoma, Esophageal Adenocarcinoma and Skin Melanoma. In addition, we also present the ratio between observed/expected LoFs mutations per sample for these cohorts. **b)** Mutation spectra associated with motif breaking events observed in HDAC2, EWSR1 and SP1 in the kidney-RCC cohort. **c)** Differences in underlying signatures between high and low impact nominal passengers in different cancer cohorts. **d)** Distribution of canonical signatures in the kidney-RCC cohort for impactful (bottom) and low-impact SNVs (top).

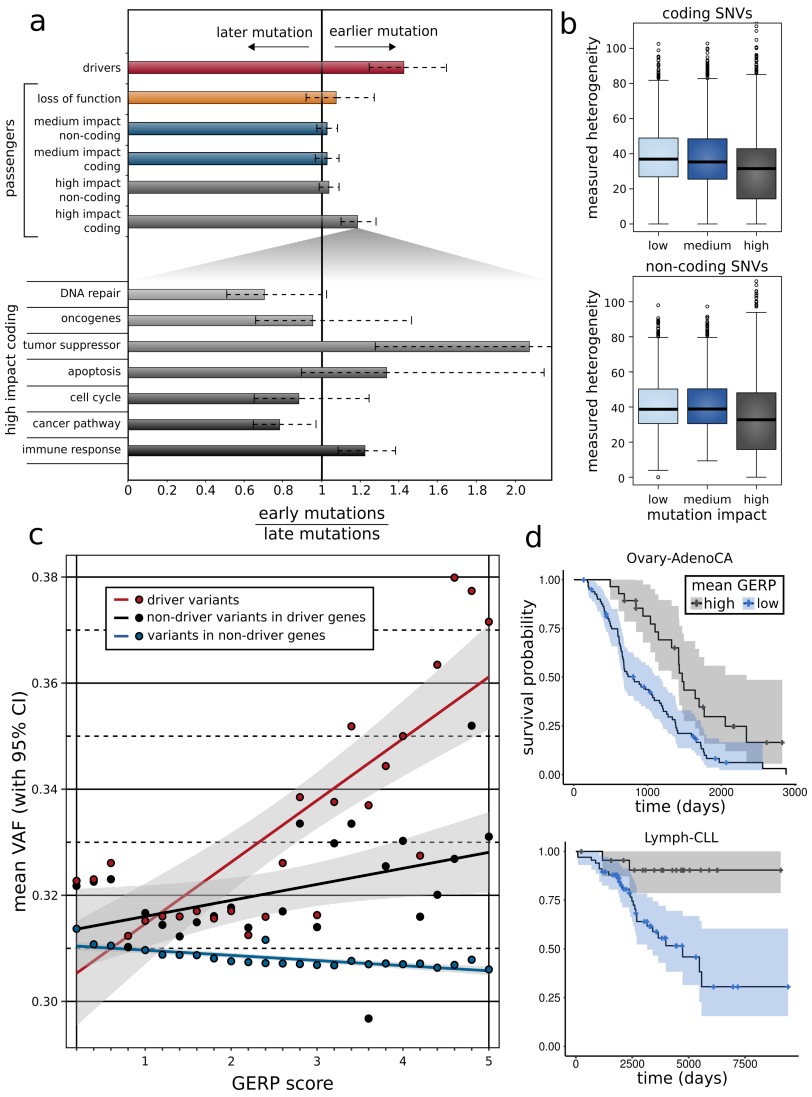

Figure 4: Correlating functional burdening with subclonal information and patient survival: **a)** Subclonal ratio (early/late) for different categories of SNVs (coding/non-coding) based on their impact score. Subclonal ratio for high impact SNVs occupying distinct gene sets. **b)** Mutant tumor allele heterogeneity difference comparison between high, medium and low impact SNVs for coding(left) and non-coding regions(right). **c)** correlation between mean VAF and GERP score of different categories of variants (driver SNVs, non-driver SNVs in known cancer genes & passenger variants in non-driver genes) on a pan-cancer level. **d)** Survival curves in CLL (*left panel*) and RCC (*right panel*) with 95% confidence intervals, stratified by mean GERP score.

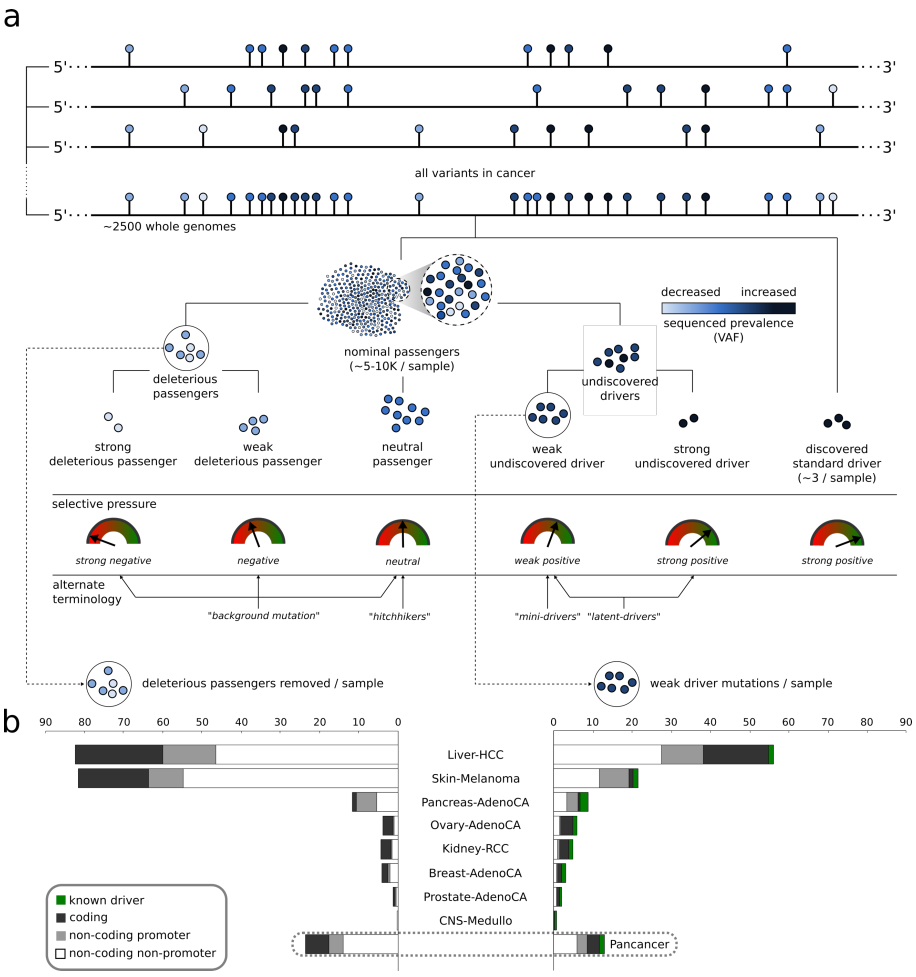

Figure 5. Conceptual classification of somatic variants into different categories based on their functional impact and selection characteristics: a) Both coding and non-coding variants can be classified as drivers and passengers based on their impact and signal of positive selection. Among nominated passengers, true passengers undergo neutral selection and tend to have low functional impact. Deleterious passengers (weak & strong) and mini-drivers (weak & strong) represent various categories of higher impact nominal passenger variants, which may undergo weak negative or positive sections. **b)** Conservative estimate (lower bound) of the number of *removed* deleterious passengers and weak drivers per sample in pan-cancer and individual cancer cohorts. Note that we only estimated these frequency for selected cohorts with sample size > 100.

Page 3: [1] Deleted	SK	10/15/17 8:09:00 AM

To estimate the overall effects of nominal passengers on tumorigenesis

Page 3: [2] Moved to page 4 (Move #2)	SK	10/15/17 8:09:00 AM

, we first adapted an additive effects model[6,10], originally used in complex trait analysis, to quantify the relative size of these aggregated effects in relation to known drivers. With a number of caveats regarding interpretation arising due to differences between germline 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 (**Fig 1a**). 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[11]; 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)[10], 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 **1b**; FDR<0.1 for all tests using a gene-level variant of the additive model except non-coding variants in CNS, Ovary and Prostate cancers, and FDR<0.001 for all tests using the basic (SNV-level) model, see supplemental table 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

Page 3: [3] Moved to page 4 (Move #3)	SK	10/15/17 8:09:00 AM

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. **1c**). 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 (supplement table X). Moreover, their effect sizes may become detectable individually with the increased power of larger datasets.

Page 3: [4] Deleted	SK	10/15/17 8:09:00 AM

what we term *impactful nominal passengers.* This intermediate functional