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 uniformly decipher their overall impact 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 *putative passenger* variants provides significant predictability for cancer phenotypes beyond the characterized driver mutations.

Style Definition: Normal (Web): Space Before: Auto, After: Auto

Style Definition: Balloon Text

Style Definition: p1

Formatted: Pattern: Clear, Highlight

Deleted: nominal

Formatted: Font:Italic

Deleted: provide

Deleted: We further used the additive effects model to provide a conservative estimate of the number of mutations with weak positive and negative fitness effects in different cancer cohorts.

Formatted: Pattern: Clear, Highlight

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 both coding and 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, several 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.

In this work, we explore the landscape of *putative passengers* in various cancer cohorts by leveraging extensive pan-cancer variant calls in PCAWG. More specifically, we build on and apply existing tools to annotate and score the predicted molecular functional impact of variants in the pan-cancer dataset. Furthermore, we integrated the annotation and impact score of each variant to quantify the overall burdening of various genomic elements in different cancer cohorts. We observed that disruption of genetic regulatory elements in the non-coding genome correlates with altered gene expression. Moreover, various mutational processes have different impact on regulatory elements, as elucidated by our signature analysis. Furthermore, we also show that the molecular functional impact burden of various genomic elements correlates with patient survival time and tumor clonality. Finally, we show that aggregating *putative passengers* provides significant predictive power beyond common driver mutations to distinguish cancer phenotypes from non-cancerous ones.

Formatted: Pattern: Clear, Highlight

Deleted: different

Deleted: noncoding

Deleted: mutation

Deleted: impacts

Deleted: the

Deleted: overall

Deleted: burdening

Formatted: Not Highlight

Deleted: come across observations which are consistent with the notion

Formatted: Not Highlight

Deleted: aggregated subsets of

Formatted: Not Highlight

Deleted: might confer weak fitness effects

Formatted: Not Highlight

Overall functional impact

In order to characterize the landscape of *putative passenger* mutations in PCAWG, we first surveyed the predicted molecular functional impact (quantified by funseq score \cite {}) 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 regions. 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 corresponds to *putative passenger*s with intermediate molecular functional impact (**Fig 1a**).

Subsequently, we investigated whether the frequency of medium- and high-impact *putative passengers* (see supp.X for classification threshold) 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 tumor acquires more SNVs, the fraction of medium-and high-impact *putative passengers* often decreases. This trend is particularly strong in CNS medulloblastoma ($p < 4e-8$), lung adenocarcinoma ($p < 3e-4$), 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 (**Fig 1c**). 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 predominance of high-impact large deletions compared to impactful SNVs in the bone leiomyoma cohort.

Burdening of different genomic elements

Furthermore, we investigated the overall mutational burden observed among different genomic elements in various cancer cohorts. *A priori*, one might assume that the overall burden of

Deleted: . ... [1]

Deleted: peaks

Deleted: in the

Deleted: regime corresponds to variants with potentially weak fitness effects

Formatted: Font color: Auto

Deleted: 1d

Deleted: 1e

Formatted: Font color: Auto

Formatted: Font:Not Bold, Font color: Auto

Formatted: Font:Bold, Font color: Text 1

Deleted: . ... [2]

Formatted: Font:Italic

putative passengers in a cancer genome would be uniformly distributed across different functional elements and among different gene categories. In contrast, we observed 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 variants (LoFs), where the putative molecular impact is most intuitive. We thus 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 of cancer-related genes (cell cycle, cancer pathway, apoptosis and DNA repair) relative to a random (shuffled-variant) control ($p < 0.001$). Conversely, non-driver LoFs displayed depletion relative 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. [[SK2MG: will update this highlighted section with other LoF analysis later.]]

Deleted: observe

Formatted: Highlight

Deleted: &

Formatted: Highlight

Formatted: Font color: Auto

As with 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 would most likely have little functional consequence. In contrast, the 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.

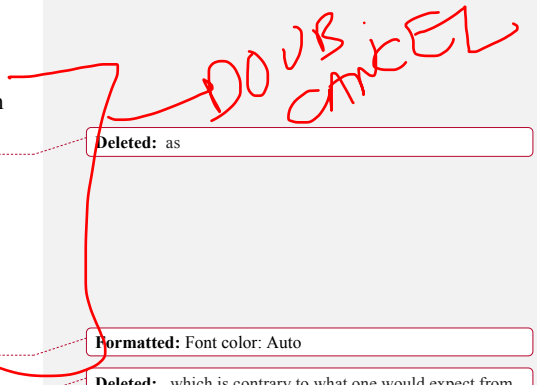
Furthermore, for a particular TF family, one can identify the associated target genes affected due to the bias towards creation or disruption of specific motifs in their regulatory elements (promoters and enhancers). For instance, the TERT gene shows the largest alteration bias for ETS motif creation across a variety of cancer types (**Fig 2c**). Other genes (such as

BCL6) showed a similar bias, albeit in fewer cancers. Moreover, the 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 the germline (**Fig 2e**). As expected, we observed that 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.

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 would 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 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**).



Deleted: as

Formatted: Font color: Auto

Deleted: , which is contrary to what one would expect from a purely random background model

Formatted: Font color: Auto

Formatted: Indent: First line: 0"

Formatted: Not Highlight

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 depending on the cancer type 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-chRCC 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 Kidney-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.

Formatted: Font:Italic

Deleted: different

Deleted: types

Deleted: RCC

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. 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

Formatted: Indent: First line: 0"

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 on the basis of variant allele frequency (VAF) (supplement Fig X).

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. We generally 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**).

Furthermore, we correlated the functional impact (measured through GERP score here) of each variant with their corresponding cellular prevalence measure through VAF. We find that, within driver genes and their regulators, variants that disrupt more conserved positions (high GERP score) tend to have higher VAF values. This trend remains true even after excluding SNVs that have been individually called as driver variants. We also find that outside of driver genes, variants that disrupt more conserved positions tend to have lower VAF values.

As with the clonal status of a tumor, clinical outcomes (such as patient survival) provide 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 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 somatic molecular impact burden and patient survival 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 $2.3x10^{-4}$) and ovary adenocarcinoma (Ovary-AdenoCA, p-value $2x10^{-3}$) (**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.

Categorizing putative passenger variants

Deleted: 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 lower VAF values. 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 VAF values. 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 VAF values.

Deleted: survivability

Deleted: overall

Deleted: survivability

Deleted: 0.00023

Deleted: 0.0020

Deleted: nominal

Our comprehensive characterizations of *putative passenger* mutations highlight some of their key attributes. These can be further explained through the underlying mutational processes or might be indicative of weak selective effects among subset of these mutations. For instance, the multi-modal functional impact distribution suggests that a subset of mutations among *putative passengers* might confer potentially weak fitness effect to tumors. Similarly, strong correlation between differential functional impact burden and patient survival, can be also inferred as presence of weak selection in certain cancer cohorts. Furthermore, differential burdening of distinct genomic elements in cancer can be associated with the operation of various signatures, which in itself is interesting. However, in certain contexts this can be potentially related to presence of weak fitness effects. For instance, depletion of *putative passenger* LoFs in key gene categories including DNA repair and cell cycle can be potentially interpreted as presence of weak negative selection in different cancers[5].

Additionally, overall enrichment and depletion of high impact *putative passengers* among TSGs and oncogenes can be indicative of weak selective effects as well. Similarly, positive and negative correlation between conservation score of *putative passengers* and their corresponding VAF, potentially suggests the presence of weak positive and negative fitness effect among subset of these mutations. However, we note that differences in signatures between and early and late subclones can also contribute to these observed differences.

If indeed a subset of putative passengers possess weak fitness effect, then we can extend the canonical model of driver and passengers into a continuum model. Conceptually, in such extended model, somatic variants can be classified into multiple categories while considering their impact on tumor cell fitness: drivers with strong positive selective effects, *putative passengers* with neutral, weak positive and weak negative selective effects. This broad classification scheme 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 a potential fitness advantage to tumor cells. With respect to 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

Formatted: Indent: First line: 0"

Moved (insertion) [1]

Formatted: Font color: Auto, Pattern: Clear

Formatted: Font color: Auto, Pattern: Clear

Deleted: Classical view of cancer progression posits that most of somatic variants confer no selective advantage to the cancer cell and are considered to occur neutrally. In contrast, a handful of driver variants are thought to give a positive selection advantage to the cancer cell. This canonical dichotomy is often useful, but it is also imprecise. Our comprehensive characterization of the passenger landscape in PCAWG present a more nuanced view compared to the canonical dichotomy of cancer variants as drivers and passengers. Conceptually, somatic 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.

Formatted: Indent: First line: 0.5"

Deleted: ,

driver mutations, defined as positively-selected variants promoting tumor growth. However, rapid accumulation of *putative passengers*, which undergo weak/strong negative selection, could adversely affect the fitness of tumor cells[5]. Moreover, a majority of low-impact and some high-functional impact *putative passengers* may alter tumor gene expression or activity in ways that are not ultimately relevant for tumor fitness; hence, these variants will undergo neutral evolution. A general approach for identifying the presence of variants with effects on tumor fitness is to compare observed mutation distributions with ones generated by simulating neutral processes. Such an approach is potentially powerful since it allows the use of complex background mutational models, although the possibility of detecting artefacts due to the inadequacy of current models of neutral mutational processes remains a possible caveat. We explore this approach below using a variety of recent neutral models in the context of an additive model of fitness effects.

Overall effects of putative passengers and additive variance

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 the remaining variants with weak effects can explain a significant part of the "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]. Despite their limitations, these models highlight the importance of investigating the cumulative effect of *putative passengers* on cancer progression.

To address this, we 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 6a**). Briefly, we created a balanced dataset of observed tumor and matched neutral (null) model samples, using a

Deleted: -

Deleted: weak and strong deleterious

Formatted: Font:Italic

Deleted: variants

Deleted: nominal

Deleted: any

Deleted: first

Formatted: Font color: Auto

Deleted: 1a

recently proposed background model which preserves mutational signatures, local mutation rates, and coverage bias [ref Broad simulation]. Subsequently, using a linear model, for each SNV the additive effects model implicitly associates a positive or negative effect (coefficient), considering them to be sampled from a normal distribution (see Online Methods and Supplemental Note). Furthermore, in this model the individual effects of SNVs are not explicitly estimated; instead their variance is evaluated as a hyper-parameter using restricted maximum-likelihood (REML)[10], where separate variance terms can be associated with different groups of SNVs falling in distinct categories. In addition to the neutral model above, we utilized two further local background models including PCAWG-wide randomized dataset as well as our custom randomization correcting for various covariates.

We compared several versions of the additive variance model above in 8 cancer cohorts having sample size greater than 100. In the first model, we separated the mutations into two categories, corresponding to drivers (from the PCAWG analysis) and *putative passengers* (**Fig. 6bi**). *Putative passengers* were only included in the model if found in at least two samples from a cohort (which can be any combination of observed and simulated samples). Additionally, to maximize the predictive potential of the driver mutations, we used a binary variable which is 1 if any driver mutation is present in a sample as a predictor (details in Online Methods). In this model, we observed an increase in the variance explained from ~49.9% using drivers alone to ~59.4% with putative passengers when averaged across all cohorts, with the *putative passenger* contribution significant at FDR<0.1 in all cohorts except Kidney, suggesting that non-neutral effects are present among the putative passenger mutations (Supp Fig. X). We further tested a different version of the model in which we split mutations into coding, promoter and other non-coding categories, where the coding mutations are a superset of the PCAWG drivers (**Fig. 6bii**). Here, we observed that the coding mutations accounted for the largest overall proportion of the variance (~50.7% averaged across cohorts), while promoters and other non-coding also contributed significant amounts of extra variance (~1.9% and 6.9% respectively overall, with cohort-specific contributions from each category at FDR<0.1, Supp Fig. X). Although the total contribution of the promoters is lowest in this model, we calculated the additive variance per SNV by normalizing by the number of SNVs in each category (**Fig. 6biii**) and found that the normalized variance is substantially higher in promoters than other non-coding, although lower than coding. Further, we tested the sensitivity of our results to the choice of null model by

Deleted: neutral

Deleted: Using

Deleted: . The model has the form $y_j = \mu + \sum_{ik} z_{ijk} u_{ik} + e_j$, where y_j is the phenotype (0/1) of sample j, z_{ijk} is the normalized SNV dosage (z-scored) of SNV i in sample j belonging to category k, e_j is the residual effect for sample j, and μ is the mean phenotype. The u_{ik}'s are normally distributed with variance σ_k^2/m_k, where σ_k^2 is the additive variance and m_k the number of SNVs in category k (where the categories represent for instance coding, promoter and other non-coding mutations), 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 $\sigma_P^2 = \sum_k \sigma_k^2 + \sigma_E^2$. The individual effects u_i are not explicitly estimated; instead the hyper-parameters σ_k^2 and σ_E^2 are optimized

Deleted: and the estimator $\sigma_A^2 = \sum_k \sigma_k^2$ can be used to predict the proportion of the phenotypic variance explained by the SNVs as σ_A^2/σ_P^2.

Formatted: Font color: R,G,B (33,33,33)

Formatted: Font:Italic

Formatted: Font:Bold

Formatted: Font:Italic

Deleted: $z_{0j1} = \vee_{i>0} z_{ijk}$).

Deleted: 50

Deleted: 57

Formatted: Font:Italic

Deleted: all

Formatted: Font:Bold

Deleted: 21

Deleted: were

Deleted: associated with

Deleted: 20

Formatted: Font:Bold

repeating these analyses for two other randomization schemes, with quantitatively similar results (Supp Fig. X).

Discussion

To a first approximation, all clinically significant consequences of genomic variants in cancer are mediated through their molecular functional impact, such as changes in gene expression or gene activity. Certain key alterations in tumor genome, often identified through the detection of strong signals of positive selection on individual variants, have been shown to play pivotal role in tumor progression. Although a typical tumor has thousands of genomic variants, very few of these (~4/tumor[1]) are thought to drive tumor growth. The remaining variants, often termed passengers, represent the overwhelming majority of the variants in cancer genomes, and their functional consequences are poorly understood. In this work, we comprehensively characterized *putative passengers* in the PCAWG dataset. As described earlier, we came across multiple line of evidences, which suggested presence of putative passengers with weak fitness effects. Subsequently, we attempted to quantify the cumulative fitness effect of such putative passengers on tumor growth through our additive variance model. We note that the above approach relies on applying an accurate background model. However, current null models have inaccuracies due to our incomplete understanding of various mutational processes in cancer. Nonetheless, our additive variance analysis was robust for multiple background models and suggested a potential role of weak positive and negative selection among *putative passengers*. Also, our functional analyses of putative passengers showed that different mutational processes are associated with extensive differences in impact on cellular subsystems, irrespective of whether these are caused by, cause, or have negligible impact on subclonal fitness differences in an evolving tumor. These observations further motivate follow-up experiments and additional whole-genome analyses to explore the role of *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: ... [3]

Formatted: Pattern: Clear, Highlight

Deleted: 3

Formatted: Not Superscript/ Subscript, Not Highlight

Deleted: We observed that functional impact distribution has a multi-modal characteristic with a significant number of nominal passengers exhibiting intermediate functional impacts. Furthermore, contrary to simple expectation, we observe fewer 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. Furthermore, 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 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

Formatted: Indent: First line: 0"

Moved up [1]: negative selection in different cancers[5].

Formatted: Font color: Auto, Pattern: Clear

Formatted: Font color: Auto, Pattern: Clear

Deleted: Additionally, 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 TSGs may potentially have 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. ... [4]

Deleted: weak

Formatted: Indent: First line: 0"

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

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

Formatted: Pattern: Clear, Highlight

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Deleted:).

... [5]

Deleted:

... [6]

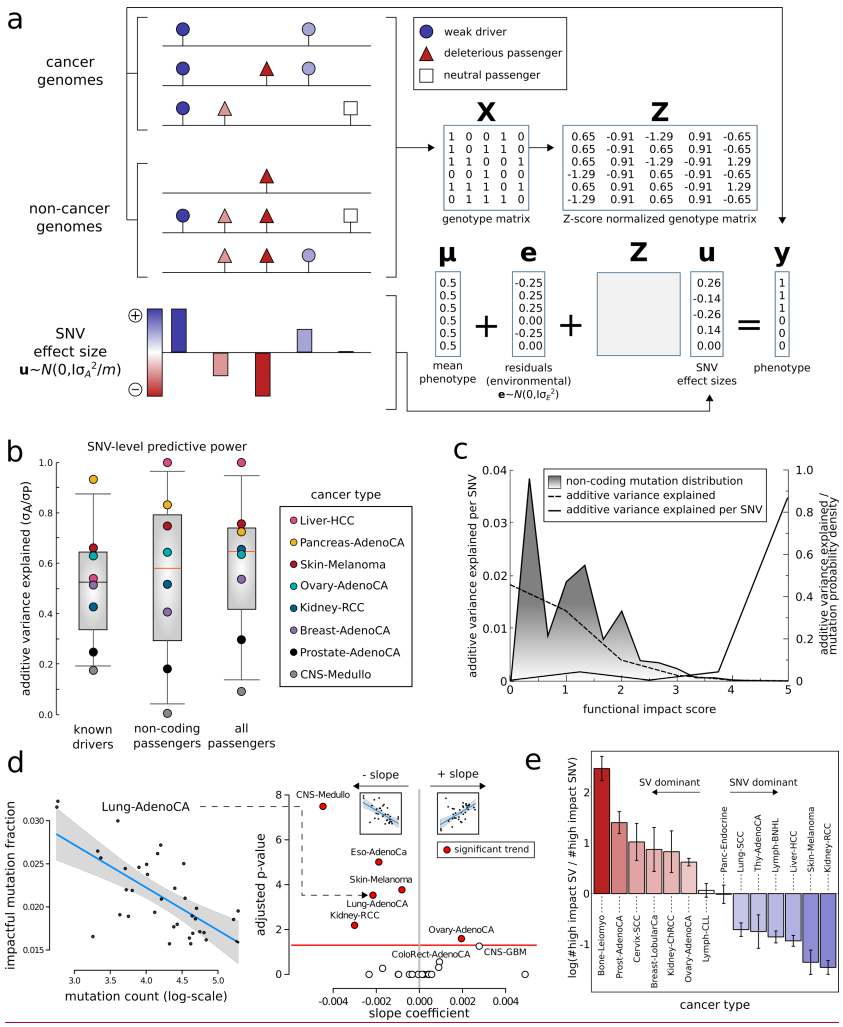


Figure 1: Additive effect and overall functional impact of PCAWG variants: *Additive effects model for nominal passengers*: The combined effects of many nominal passengers are modeled using a linear model, which predicts whether a genotype arises from an observed cancer sample or from a null (neutral) 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 $\boldsymbol{\mu}$ and $\mathbf{e}$. **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)**

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:12 pt

Formatted: Font:Cambria Math, 12 pt

Formatted: Font:Cambria Math, 12 pt

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.

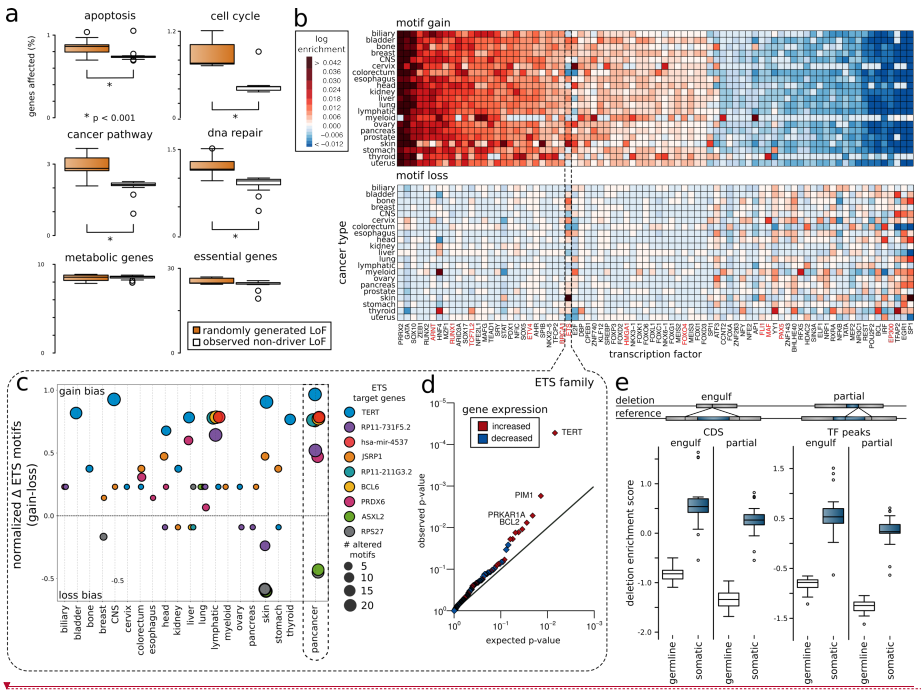
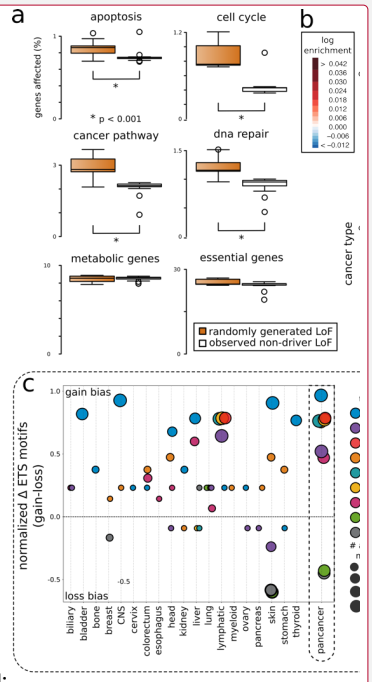


Figure 2: Overall functional burdening of different genomic elements: a) Percentage of genes in different gene categories (apoptosis, cell cycle, cancer pathway, dna repair, metabolic and essential genes) affected by non-driver LoFs in observed and random model, **b)** *Pan-cancer overview of TFs burdening:* 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.

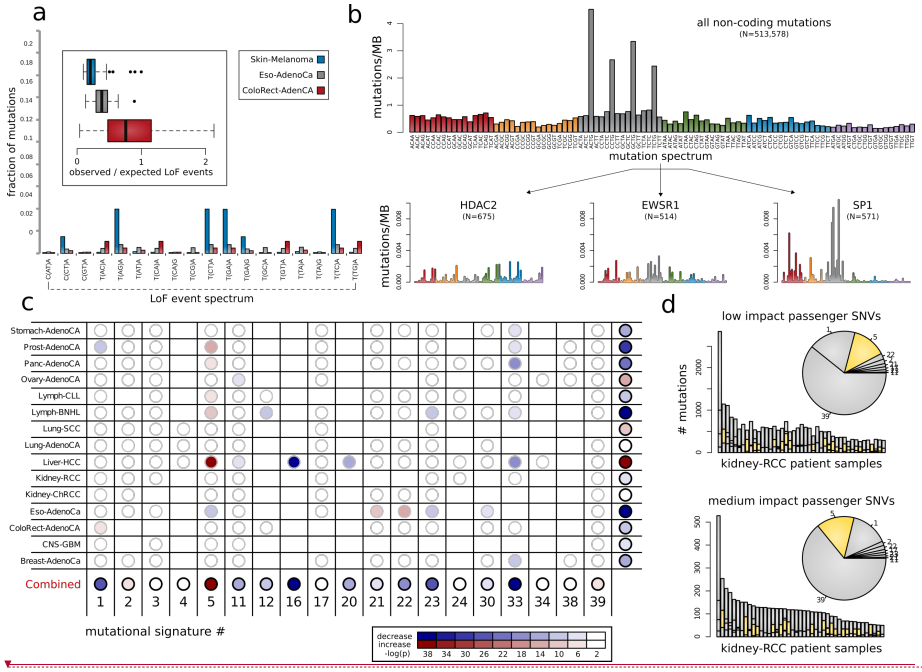
Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

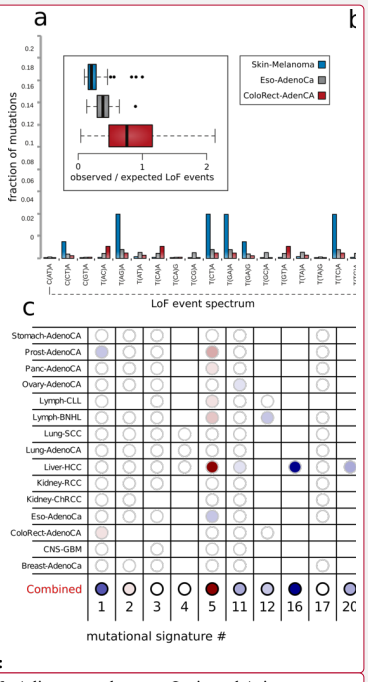


Deleted:

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers





Deleted:

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

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).

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

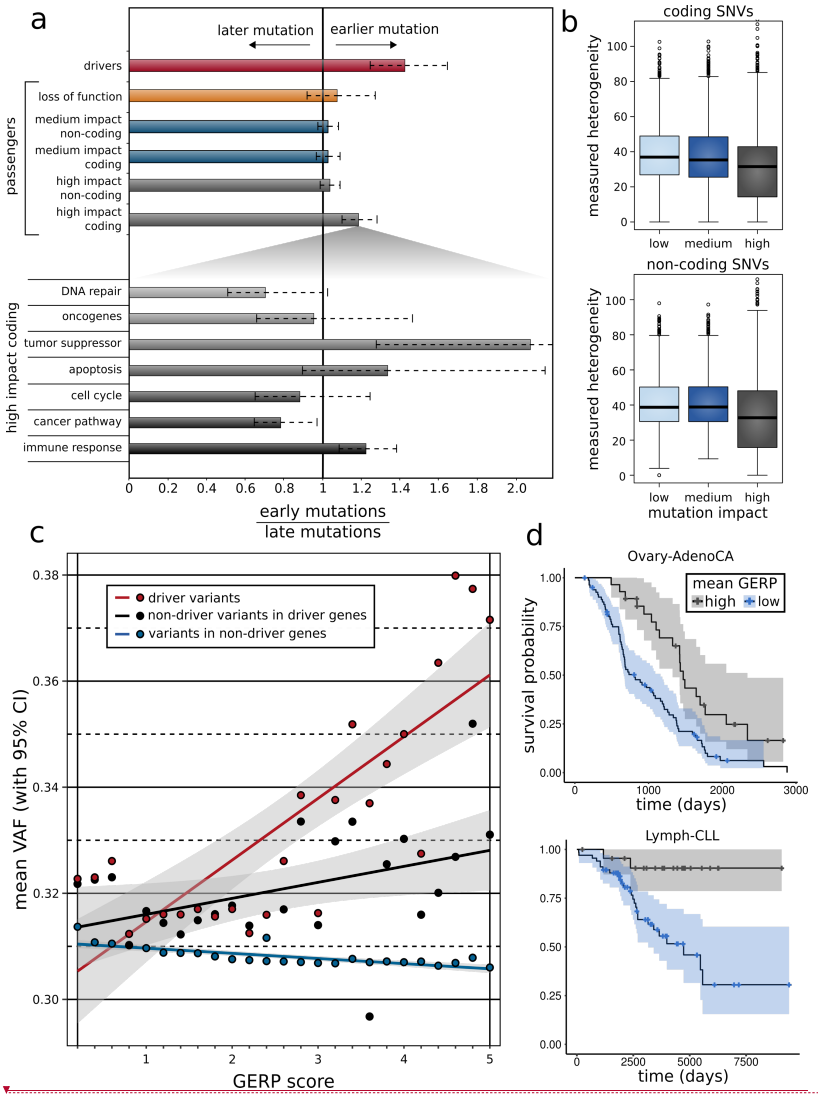


Figure 4: Correlating functional burdening with subclonal information and patient survival: a) Subclonal ratio (early/late) for different categories of SNVs (coding/non-coding) based on their impact score. Subclonal ratio for high-impact SNVs occupying distinct gene sets. **b)** Mutant tumor allele heterogeneity 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.

Deleted:

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

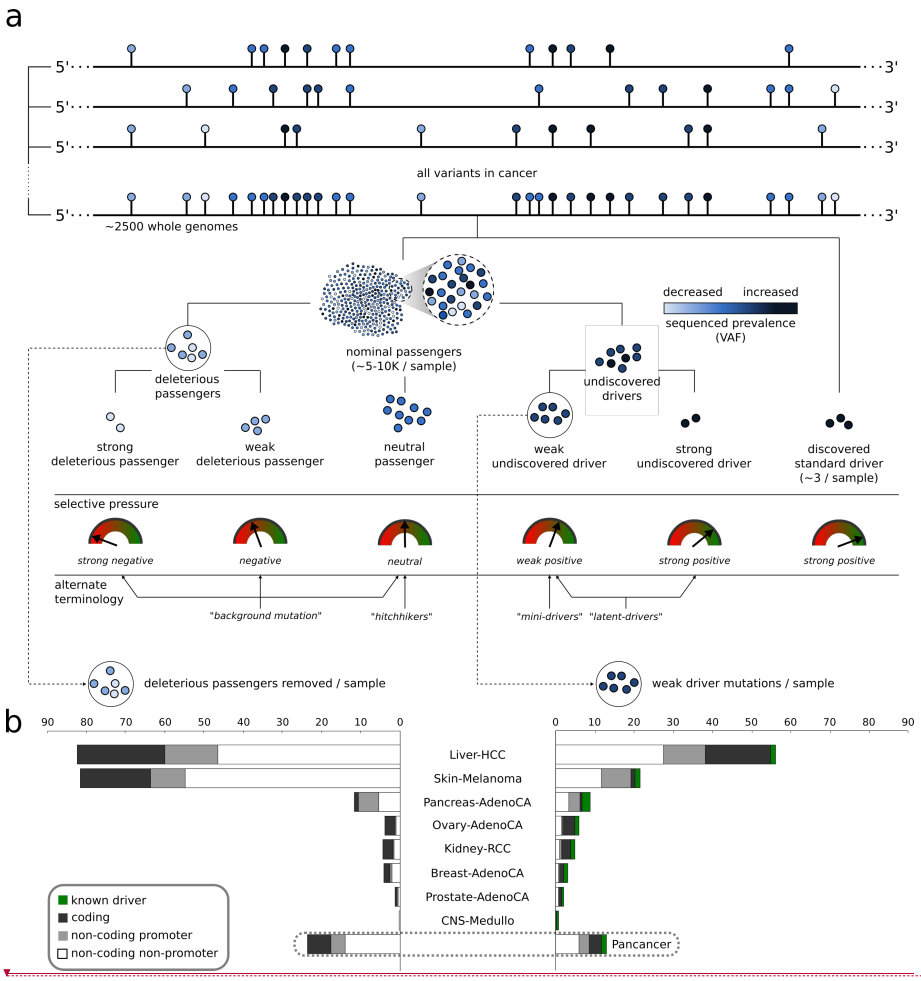


Figure 5. Conceptual classification of somatic variants into different categories based on their functional impact and selection characteristics: a) Both coding and non-coding variants can be classified as drivers and passengers based on their impact and signal of positive selection. Among nominated 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.

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Deleted:

Formatted: Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font color: R,G,B (0,0,10)

Page 3: [1] Deleted	Kumar, Sushant	11/3/17 9:21:00 PM

Page 3: [2] Deleted	Kumar, Sushant	11/3/17 9:21:00 PM

Page 11: [3] Deleted	Kumar, Sushant	11/3/17 9:21:00 PM

Estimating number of weak drivers and deleterious passenger variants

Finally, we used the additive variance model to attempt to provide an estimate of the numbers of nominal passengers with positive and negative effects on fitness in each cancer across mutation types. To estimate a lower bound on the number of the nominal passengers with non-neutral effects, we find the size of the smallest subset of SNVs needed to explain the same amount of the phenotypic variance as when using all nominal passengers collectively (See Supplemental Note). Further, we use the signs of the maximum a-posteriori estimate for the effect of each individual SNV to then predict the number of weak drivers and deleterious passengers removed and retained per tumor across the smallest subset. Using the SNVs directly as predictors as in the model above, we predict a pan-cancer average of ~9 weak drivers per tumor (in addition to the PCAWG drivers) and ~2 deleterious passengers removed (Supp Fig. X). We also investigated a version of the additive model in which we pooled mutations at the gene level to form the predictors (which we then normalized to form the z_{ijk}'s), and optimized over the functional impact score for the variants to include in the model. Using this model, we predicted slightly higher numbers of variants with negative effects removed (~8 per tumor) and also a significant number of retained deleterious variants (~2.4; Supp Fig. X). We note however that the pooled model may be more susceptible to systematic inaccuracies in the null model. As above, we observed quantitatively similar results when we compared the sensitivity of these results to changes in the background null model.

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. 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 estimates are higher than those predicted from the additive variance model, and may point to a lack of power in the additive variance analysis to detect non-neutral effects *ab initio*, which may be resolved through larger sample sizes.

Page 11: [4] Deleted **Kumar, Sushant** **11/3/17 9:21:00 PM**

Additionally, 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 TSGs may potentially have 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 have applied an accurate randomized model to perform the comparison. In this work, we utilized multiple local background models including PCAWG-wide randomized as well as our custom randomization correcting for various covariates. Our results are robust with respect to these different background model. However, our understanding of the underlying mutational processes and genome structure of a tumor sample is limited, which can be a hindrance to achieving the accurate null model. Nonetheless,

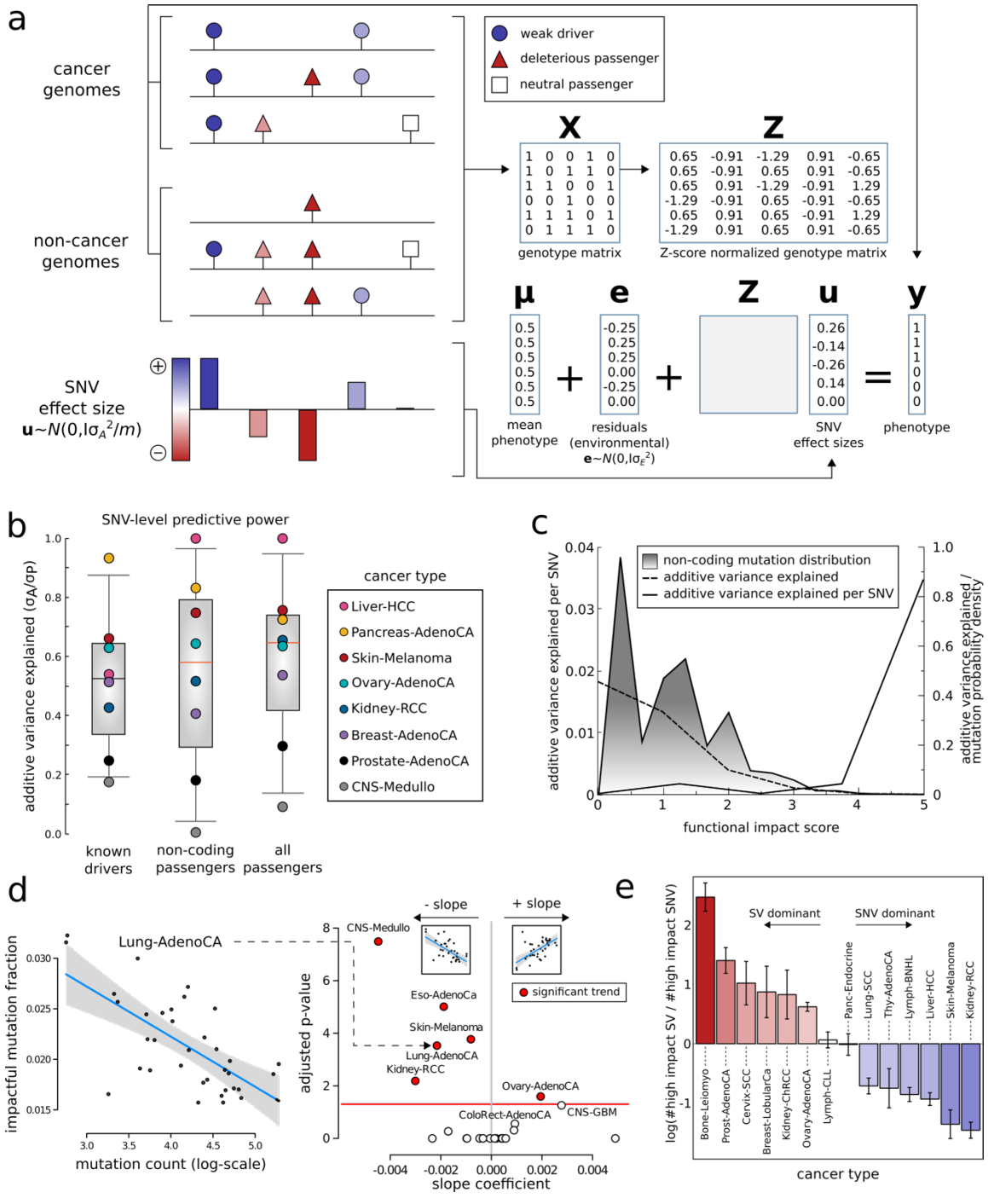
our additive variance analysis suggest potential role of weak positive and negative selection among *putative passengers*.

Page 12: [5] Deleted	Kumar, Sushant	11/3/17 9:21:00 PM

).

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

Page 12: [6] Deleted	Kumar, Sushant	11/3/17 9:21:00 PM


a
cancer genomes
non-cancer genomes
SNV effect size
u~N(0,Iσ_A²/m)
weak driver
deleterious passenger
neutral passenger
X
genotype matrix
Z
Z-score normalized genotype matrix
μ
mean phenotype
e
residuals (environmental)
e~N(0,Iσ_E²)
Z
u
SNV effect sizes
y
phenotype
b
SNV-level predictive power
additive variance explained (σ_A/σp)
known drivers
non-coding passengers
all passengers
cancer type
Liver-HCC
Pancreas-AdenoCA
Skin-Melanoma
Ovary-AdenoCA
Kidney-RCC
Breast-AdenoCA
Prostate-AdenoCA
CNS-Medullo
c
non-coding mutation distribution
additive variance explained
additive variance explained per SNV
additive variance explained per SNV
additive variance explained / mutation probability density
functional impact score
d
Lung-AdenoCA
impactful mutation fraction
mutation count (log-scale)
- slope
+ slope
significant trend
adjusted p-value
slope coefficient
CNS-Medullo
Eso-AdenoCa
Skin-Melanoma
Lung-AdenoCA
Kidney-RCC
Ovary-AdenoCA
ColoRect-AdenoCA
CNS-GBM
e
SV dominant
SNV dominant
log(#high impact SV / #high impact SNV)
Bone-Leiomyo
Prost-AdenoCA
Cervix-SCC
Breast-LobularCa
Kidney-ChRCC
Ovary-AdenoCA
Lymph-CLL
Panc-Endocrine
Lung-SCC
Thy-AdenoCA
Lymph-BNHL
Liver-HCC
Skin-Melanoma
Kidney-RCC
cancer type