

# **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 remaining mutations (termed “nominal passengers”) are considered inconsequential for tumorigenesis. In this study, we leverage the comprehensive variant data from PCAWG to predict the molecular impact of each variant, including nominal passengers. This allows us to decipher their overall molecular impact on different coding and noncoding genomic elements. The overall molecular impact distribution of PCAWG mutations shows that, in addition to high impact drivers and low-impact passengers, there is a group of medium-impact passenger variants predicted to influence gene expression or activity. Furthermore, we find that molecular impact relates to the underlying mutational signature. Thus, different signatures confer different extent of molecular functional impact. Moreover, burdening of variants is non-random in effect on different regulatory subsystems and for different categories of genes. In addition, we find that molecular functional impact varies based on subclonal architecture (i.e. early vs late mutations) and can be also related to patient survival. Finally, we speculate on how the differential burdening might be related to the existence of both weak positive and negative selection during tumor evolution.

## **Introduction**

Previous studies have focused on characterizing variants occupying coding regions of cancer genomes \cite{24071849}. However, the extensive Pan-cancer Analysis of Whole Genomes (PCAWG) dataset, which includes variant calls from ~2500 uniformly processed whole cancer genomes, offers an unparalleled opportunity to investigate the overall molecular functional impact of variants influencing different non-coding genomic elements. Given that the majority of cancer variants lie in non-coding regions \cite{26781813}, this variant dataset serves as a substantially more informative resource than the many existing datasets focused on exomes. Moreover, it also contains a full spectrum of variants, including copy number variants (CNVs) and large structural variants (SVs), in addition to single-nucleotide variants (SNVs) and small insertion & deletions (INDELS).

Of the 30 million SNVs in the PCAWG data set, a few thousand ( $< 5/\text{tumor}^1$ ) \cite{26559569} can be identified as driver variants – positively selected variants that favor tumor growth. The remaining ~99% of SNVs are termed nominal passenger variants, with poorly understood molecular consequences and fitness effects. Furthermore, the bulk of these nominal passengers fall within noncoding regions of the genome, making these the main product of whole-genome sequencing of tumors. Recent studies have proposed that, among variants that have not been found to be driver variants (i.e. nominal passenger variants), some may weakly affect tumor cell fitness by promoting or inhibiting tumor growth. In prior studies, these variants have been described as “mini-drivers” \cite{26456849} and “deleterious passengers” \cite{23388632}.

Conceptually, variants can be classified into three categories based on their impact on tumor cell fitness: positively-selected driver variants, neutrally-selected neutral passenger variants, and negatively-selected deleterious passenger variants. This broad classification can be further refined by considering ascertainment-bias and the putative molecular impact of different variants (**Fig 1**). Previous power analyses \cite{24390350} suggest that existing cohort sizes support the identification of strong positively-selected driver variants, but that many weaker drivers, and even some moderately strong driver variants would be missed. However, these moderately strong and weak driver variants can also provide potential fitness advantage to tumor cells. As for the functional-impact-based-classification: the philosophy of molecular reductionism holds that any positively or negatively selected variants have some functional impact (i.e. effect on gene expression or activity). The relevance of molecular functional impact is firmly established for driver mutations - positively-selected variants promoting tumor growth. However, rapid accumulation of weak and strong deleterious passengers, which undergo negative selection, could adversely affect the fitness of tumor cell \cite{23388632}. Moreover, majority of low impact and some high functional impact variants may alter tumor gene expression or activity in ways that are not ultimately relevant for tumor fitness; hence, these variants will be under neutral selection.

### **Impactful passengers and their prevalence**

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

One would expect that if a nominal passenger variant does indeed impact tumor cell fitness, its effect should be mediated through its molecular functional impact. Therefore, in order to relate the presence of different categories of nominal passenger SNVs to their role in cancer progression, we surveyed the putative molecular functional impact distribution of somatic variants in different cancer genomes. The molecular functional impact distribution varies across different cancer types and different genomic elements. For instance, impact score distributions of non-coding variants in different cancer genomes demonstrate three distinct peaks. The upper and the lower extremes of this distribution correspond to traditional definitions of high-impact putative driver variants and low impact neutral passengers respectively. In contrast, the middle peak in the intermediate molecular functional impact regime corresponds to what we term *impactful nominal passengers*. This intermediate functional impact category could include undiscovered drivers (strong & weak) as well as potentially deleterious passengers (**Fig 2a**). Conceptually, fitness effects of mutations can be positive or negative for tumor cells. Although fitness effects can be directly established through specialized genetic functional experiments \cite{} , one powerful statistical approach for detecting the fitness effects of variants is to identify discrepancies between observed mutation feature distributions and appropriate null models of neutral mutation. A uniform distribution of mutation across the genome is a useful null model for making descriptive statements about the functional properties of the human genome, and about the functional impact of mutational processes in cancer. A null distribution formed by shuffling the location of variants identified in cancer genomes has the potential to provide suggestive evidence of selection and is described in more detail in supplemental Method X.X.

According to a uniform null expectation, we might assume that the overall burden of variants in a cancer genome will be uniformly distributed across different functional elements and among different gene categories. In contrast, we observe that the molecular impact burden in certain cancers is concentrated in particular gene categories. This is easiest to understand in terms of coding loss-of-function (LoF) variants, where the molecular impact is most intuitive. For instance, as a measure of the molecular impact of both driver and non-driver loss of function SNVs, we examined the fraction of deleterious LoFs affecting genes across four categories of cancer-related functional annotation (**Fig 2d**).

Driver LoF variants, which are well understood high impact variants, showed significant enrichment in each category of cancer-related functional annotation compared to a random (shuffled-variant) control ( $p < 0.001$ ). Conversely, non-driver LoF SNVs displayed depletion in each of these categories ( $p < 0.001$ ). Driver, non-driver, and random loss of function mutations were all enriched in comparison to germline LoF mutations ( $p < 0.001$ ). Given the high selective pressure presumed to act against germline deleterious loss of function mutations *in utero*, our observations suggest that both driver and non-driver LoF mutations exert molecular functional impact. Similarly, compared with the uniform null distribution, we observe that *impactful variants* (nonsynonymous & promoter SNVs) tend to occur in essential genes more often compared to low impact variants (**Fig 2b**). Conversely, low impact passengers constitute larger fractions of variants influencing non-essential genes. This observation is consistent with underlying functional properties of the human genome.

### **TF binding landscape and overall impact of variants**

Similar to LoF variants, we can also quantify the overall burden of the noncoding region of the genome. However, for majority of noncoding variant, functional impact is less easy to gauge. For instance, noncoding and coding variants occupying the terminal region of the gene or undergoing alternatively splicing, will have little functional consequence. In contrast, transcription factor binding site (TFBS) variants are among the noncoding variants with functional impact that is more straightforward to measure. Similar to the effect of a LoF variant on a gene, their molecular impact is clearly manifested through the creation or destruction of TF binding motifs (gain or loss of motif). In both cases (gain or loss), we observed significant differential burdening of TFBS among different cancer cohorts. For instance, we detect significant enrichment of high impact variants creating new motifs in various TFs including GATA, PRRX2 and SOX10 (**Fig 3b**) across major cancer types, compared with uniform expectation. Similarly, high impact variants breaking motifs, were highly enriched in TFs such as IRF, POU2F2, NR3C1 and STAT (**Fig 3f**) in the majority of cohorts. This selective enrichment or depletion suggests distinct alteration profiles associated with different components of regulatory networks in various cancers. A gene-centric analysis of these alteration patterns highlights genes with bias towards creation or disruption of specific motifs in their regulatory elements (promoters and enhancers). For instance, TERT shows the largest alteration bias for ETS motif creation across a variety of cancer types (Fig 3d), with other genes (such as NEAT1) showing a similar bias, albeit in fewer cancers. Interestingly, ETS motifs appear to show a systematic bias towards motif creation, whereas MYC-family motif alterations show alteration biases in both directions (Fig. 3d). Furthermore, enrichment of SNVs in selective TF motifs leads to gain and break events in promoter that significantly perturb the overall downstream gene

expression (**Fig 3g**). For example, among lung adenocarcinoma patients, target gene expression for TFs undergoing motif breaking events was significantly lower than for target genes without TF motif loss. Moreover, in lung adenocarcinoma, we found gain events in three TFBSs (ZBTB14, E2F and HNF4) that significantly increase downstream expression level ( $p < 5e-7$ ,  $3e-6$  and  $2e-4$  respectively) (**Fig 3c**). Similarly, ETS family transcription factor at the regulatory region of IRF4 and PSIP1 gene displayed a strong motif creation bias and a significant change in gene expression (with p-value IRF4=0.001 and p-value PSIP1=0.019).

### **Signature Analysis**

The disproportionate functional load on certain TFs in cancers can be related to an underlying mutational spectrum (signature) influencing their binding sites. For instance, the mutational spectrum of motif breaking events observed in SP1 TF binding sites (TFBS) suggests major contribution from C>T and C>A mutation (**Fig 4b**). In contrast, motif-breaking events at TFBS of HDAC2 and EWSR1 have relatively uniform mutational profiles. Similarly, comparing the signature composition of low and high impact SNVs in certain cancer-cohorts can help us to distinguish between mutational processes that generate distinct variant impact classes. For instance, we observed distinct signature distributions for the low and high impact non-coding passengers in the renal cell carcinoma cohort. While the majority of passenger variants can be explained by signature 5, high impact passengers have a higher fraction of SNVs explained by signature 4 (**Fig4a**). Moreover, we observed cancers showing microsatellite instability (MSI) due to failure of DNA mismatch repair, have a higher percentage of high impact non-coding passengers (**Fig4c**). Our findings suggest that various mutational processes shape and disproportionately burden cancer genomes.

### **Overall variant impact**

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

Additionally, we sought to examine whether cumulative molecular impact of variants can be associated with tumor initiation and progression. Using a cox proportional hazard model, we used the mean GERP of somatic variants per patient to predict patient survival, after including patient age at diagnosis and total number of mutations as covariates in the 9 tumor subtypes with 10 or more patient deaths. We obtained significant results in two subtypes after multiple test correction. In both cases, higher mean GERP predicted better patient survival (p-value 0.00023 in Lymph-CLL, 0.0020 in Ovary-AdenoCA). High mean GERP is driven by the mean GEPR of passenger variants, and the preceding sections indicate that high mean GERP may imply deleterious passenger variants. Therefore, the prolonged survival of high mean GERP patients in these subtypes is consistent with the possibility that mutations that are deleterious to tumor cells benefit the patient. In addition, we observed similar correlation between patient age at cancer diagnosis and impactful germline mutation burden. We found that patients harboring a greater number of high-impact rare germline alleles were diagnosed with cancer at earlier ages in three cancer subtypes including breast adenocarcinoma, CNS medulloblastoma and pancreatic endocrine cancer.

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

### **Subclonality and impact score**

Furthermore, we also explored the role of impactful variants in cancer evolution by analyzing variants in the context of their associated tumor sub-clone. One might hypothesize that high impact mutations achieve greater prevalence in tumor cells if they are advantageous to the tumor, and a lower prevalence if

deleterious. Interestingly, there is evidence to corroborate this hypothesis. We observed that high impact passenger variants in coding regions have greater prevalence among parental subclones (**Fig 5a**) – an effect driven by high impact nominal passenger SNVs in tumor suppressor and apoptotic genes (**Fig 5a**). In contrast, high impact passenger SNVs in oncogenes appear slightly depleted. Similarly, impactful SNVs in DNA repair genes and cell cycle genes are depleted in early subclones (**Fig 5a**). Furthermore, we also observe low heterogeneity in prevalence among higher impact variants compared to lower impact variants. This observation is consistent for both coding and non-coding variants (**Fig 5c**).

### **Functional impact and variant allele frequency**

Finally, we employed a similar analysis using variant allele frequency (VAF) to explore whether passenger variants with high functional impact also conferred a fitness impact to tumor cells. We would expect for variants that enhance tumor cell fitness to achieve an overall higher than average mean VAF, while variants that reduce tumor cell fitness to occur at an overall lower mean VAF. Indeed, driver SNVs occur at higher mean VAF, non-silent coding SNVs and noncoding variants in sensitive regions occur at lower mean VAF, and synonymous variants along with variants in inter-genomic regions occur at intermediate mean VAF (**Fig 5b**). This suggests that in aggregate, non-silent passenger variants and noncoding variants in sensitive regions impair cancer cell fitness. Additionally, we generalize our observations among functional classes by correlating their respective variant frequency with the degree of conservation. Highly conserved positions (i.e. those with high GERP) are expected to be important for organismal fitness, as polymorphisms at those positions could hurt cellular function and in other cases because polymorphisms at those positions could promote undue cellular fitness (i.e. cancer) at the cost of organismal fitness. As expected, we observe that in PCAWG driver genes, VAF and GERP have a small but statistically significant positive correlation (with coefficient 0.0040 and p-value 0.0046). Interestingly, VAF and GERP have a correlation of similar magnitude but in opposite direction among variants not in driver genes, with very high significance (coefficient -0.0034, p-value  $< 2.2e-16$ ). The observed trend for passenger variants at more conserved positions to occur at lower VAF is consistent with the deleterious passenger hypothesis.

### **Discussion**

Previous studies [\cite{20562875}](#) related to the missing heritability problem in GWAS, indicate that SNVs of low level association with a complex trait, usually cannot be identified through a stringent statistical test. However, the cumulative effect of SNVs can explain the majority of this missing association in a GWAS study. Similar to this, here we investigated the hypothesis that cumulative molecular impact of many weak somatic SNVs can have a meaningful impact on cancer progression. This

hypothesis stands in contrast to the classical model of cancer, which holds that a few driver variants promote tumor growth, while the thousands of remaining mutations are of no significance to tumor fitness. Intuitively, tumor cells must maintain function of some minimal set of essential genes in order to achieve homeostasis. It is plausible that the aggregate effect of functionally impactful passenger variants on these essential genes would be deleterious to tumor cells \cite{23388632}. For instance, radiation therapy and some chemotherapies are believed to kill tumor cells by causing DNA damage \cite{PARP inhibitors are a good example...}. Similarly, increased mutation counts in coding genes or regions relevant for splicing increase the antigenic cross-section of tumor cells, potentially making them vulnerable to immune surveillance \cite{ }. Conversely, any variant that optimizes cell-division at the expense of organism-supporting functions is expected to have a small positive effect on tumor fitness that may be challenging to detect. Moreover, certain variants through their complex genetic regulatory interactions may moderately increase the expression levels of canonical oncogenes. It has been proposed that these weak undiscovered driver variants benefit tumor growth and have a small associated positive selection.

In this work, several observations support the notion that some nominal passenger variants undergo weak selection. First, we observed overall enrichment and depletion of nominal passenger variants among TSGs and oncogenes, respectively. An interpretation of this finding is that passenger variants in tumor suppressor genes have weak driver activity, while passenger variants in oncogenes impair oncogenic activity to the detriment to tumor fitness. Similarly, our finding of depletion of nominal passenger variants among DNA repair and cell cycle genes may indicate that high impact variants affecting these genes decrease tumor cell survival in relation to greater mutational burden. We also found that variants at more conserved positions have lower VAF, suggesting that impactful passenger variants can encumber the tumor cells they inhabit. Consistent with a possible deleterious effect of passenger variants on tumor growth, in some cancer subtypes, the most mutated tumors have a lower fraction of impactful variants. This may relate to the aggregate impact of passenger variants becoming more deleterious at higher mutation loads. Alternatively, a fixed number of undiscovered drivers may become diluted by neutral passengers at higher mutation counts. Our LoF mutation analysis indicates that driver LoF mutations exert a positive selective effect, whereas non-driver LoF mutations apparently exert a net negative selective pressure. This observation is consistent with prior evidence of net negative selective effect among nominal passenger missense mutations \cite{ }. The aggregate fitness impact of nominal passenger variants may help explain why patient survival times are correlated with functional impact load in select cancer subtypes. In conclusion, our work highlights that an important subset of somatic variants originally identified as passengers nonetheless show biologically and clinically relevant functional roles across a range of cancers.



## **References**

1. Vogelstein, B. & Kinzler, K. W. The Path to Cancer --Three Strikes and You're Out. *N. Engl. J. Med.* **373**, 1895–8 (2015).
2. Nussinov, R. & Tsai, C. J. 'Latent drivers' expand the cancer mutational landscape. *Current Opinion in Structural Biology* **32**, 25–32 (2015).
3. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15**, 680–685 (2015).
4. McFarland, C. D., Korolev, K. S., Kryukov, G. V, Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2910–5 (2013).