# Passenger mutations in >2500 cancer genomes: Overall burdening & selective effects

A typical tumor has thousands of genomic variants, yet very few of these (<5/tumor[1]) are thought to drive tumor growth. The remaining variants, termed passengers, represent the overwhelming majority of the variants in cancer genomes, and their functional consequences are poorly understood. Furthermore, the bulk of these passengers fall within noncoding regions of the genome, making these the main product of whole-genome sequencing of tumors. Passengers can be subdivided into neutral and impactful based on their predicted functional impact on the genome. Low-impact passengers are thought to be inconsequential for tumor progression. However, impactful passengers can alter gene expression or activity, and while some of these changes may be irrelevant, others may promote or inhibit tumor cell growth and survival, as has been suggested for *latent driver variants* [2,3] ("mini-drivers") and *deleterious passengers*[4], respectively.

Here, we explore the landscape of passenger impact in various cancer cohorts by leveraging extensive pan-cancer variant calls from ~2700 uniformly processed whole cancer genomes. More specifically, we annotate and evaluate the impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset. Subsequently, we integrate their annotations and impact scores to quantify the overall burdening of various elements in cancer genomes. Furthermore, we also show how overall functional burdening correlates with age at cancer diagnosis, patient survival time, and tumor clonality.

In order to substantiate the presence of various categories of passenger variants, we surveyed the functional impact distribution of somatic variants in the pan-cancer dataset, comparing it to the impact of natural germline ones. Based on canonical classification of somatic variants as passenger and drivers, one might expect their functional impact score distribution to be unimodal and centered around 0 (as a result of a large number of neutral passengers), along with a tail in the high-impact score regime, corresponding to putative drivers. However, inspection of impact scores for somatic variants across cancer cohorts reveals a very different picture: passengers can be broadly classified into three distinct subgroups. The upper and the lower extremes, which comprise ~23 and ~13,500 thousand noncoding variants per patient, fall under traditional definitions of high-impact putative driver variants and neutral passengers. In contrast, the intermediate functional impact regime comprises of *impactful passengers* (~3,500 thousand noncoding variants per patient), which can further influence cancer progression by acting as latent drivers or through aggregate burdening of functional elements. This score distribution is very different for the germline. There is a significant depletion of mildly deleterious, intermediate functional impact variants in the cancer cohorts. As expected, the germline genome comprises of mostly neutral variants with a very few high-impact alterations.

One might further expect that presence of impactful passengers varies among different genomic elements as well as different cancer cohorts. Consequently, we comprehensively analyzed the overall burdening of various genomic elements, including TF (transcription factor) motifs in the pan-cancer somatic variant dataset. The presence of a variant within a TF binding site can lead to either the creation or destruction of binding motifs (gain or loss of function). In both cases, we observe significant differential burdening of *impactful variants* among different cancer cohorts. For instance, we observe significant enrichment of high impact variants creating new motifs in various TFs such as CCNT2 and HNF4 in lymphatic and myeloid cancer meta-cohorts. Similarly, high impact variants influencing gene expression by breaking already existing TF motifs, were highly enriched in STAT and SP1 TFs in both

these cohorts. In contrast, we observed a significant depletion of large impact SNVs creating new TF motif in YY1 (known to regulate activity of various promoters) for these meta-cohorts. This selective enrichment or depletion suggest distinct alteration profiles associated with different components of regulatory networks in various cancers.

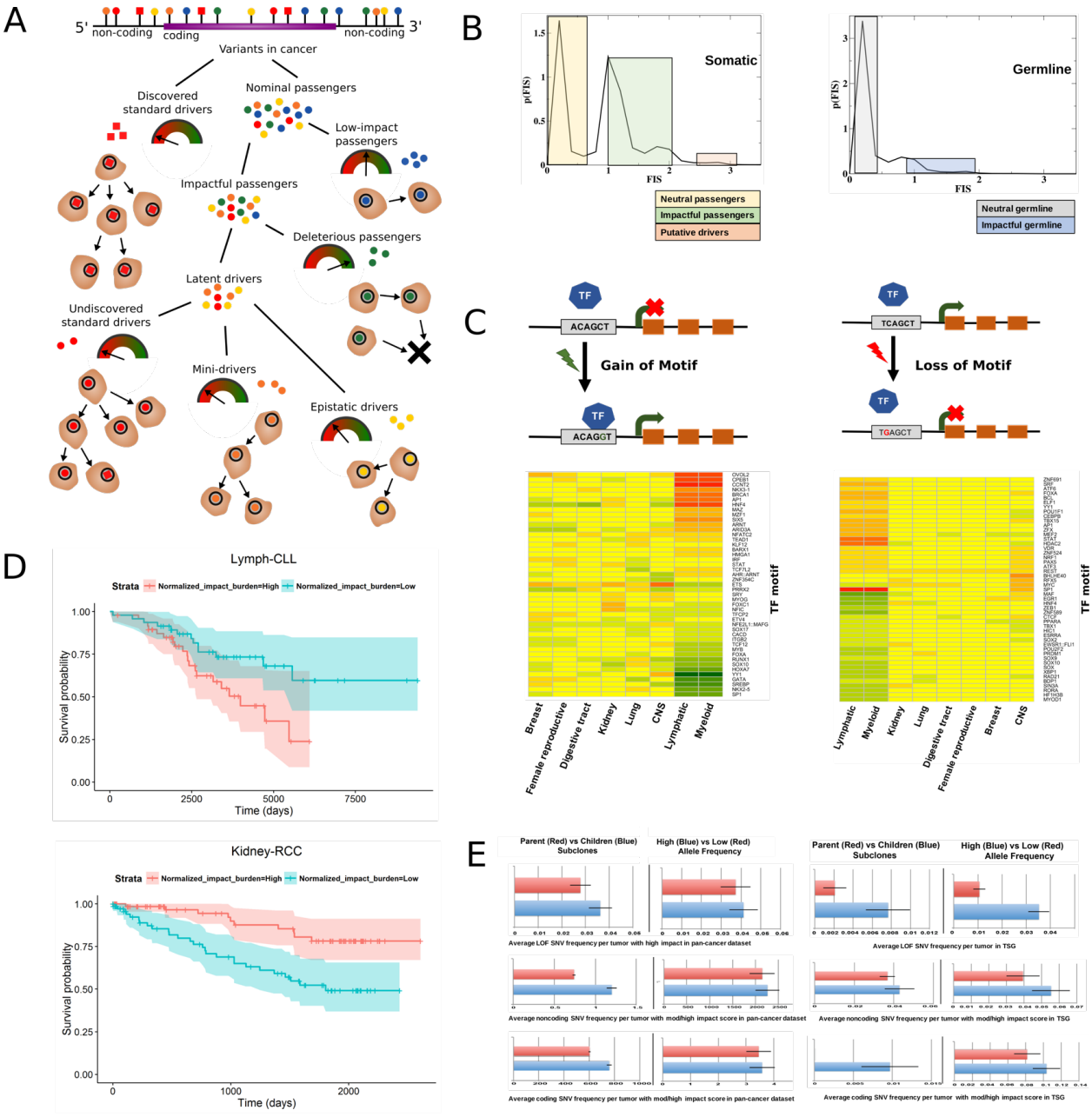
Additionally, we explored the role of impactful variations in cancer evolution by integrating them with sub-clonal information and allele frequencies. Intuitively, one might hypothesize that high impact mutations should either achieve higher frequency if they are advantageous to the tumor, or a lower frequency if deleterious. Interestingly, one finds suggestive observations that this is the case. In particular, we observe that high functional impact non-coding variants (along with high impacting coding LOF variants) have a higher allelic frequency and a higher prevalence in parental subclones, signifying a potential important role in the early phases of cancer progression or providing a higher fitness advantage.

Finally, we sought to examine whether impactful passengers might exert a clinically meaningful effect on cancer initiation and progression. To study tumor initiation, we correlated patient age at diagnosis with their number of high impact coding germline variants, finding that each high impact coding germline variant led to an earlier diagnosis of cancer by 2-5 years in five tumor subtypes, suggesting that an important share of germline variants can serve as epistatic drivers. Surprisingly, there was also one subtype –squamous cell lung cancer– in which germline impact burden correlated with later tumor diagnosis. To study tumor progression, we performed survival analysis to see if impact burden in non-driver genes predicted patient survival within individual cancer subtypes. These correlations varied substantially in different cancer types. For instance, we observed that somatic mutation burden predicted substantially earlier death in chronic lymphocytic leukemia (CLL) and substantially prolonged survival in renal cell carcinoma (RCC), respectively. These results lend support to the hypothesis that the aggregate amount of impactful passengers is clinically meaningful. More specifically, the results suggest that latent drivers are more important than deleterious passengers in CLL, but that the situation is reversed in RCC. This can be explained by the large share of missing drivers in CLL, which suggests a greater role for latent drivers in CLL.

In conclusion, our work highlights 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).




F

| Cancer Subtype | Change in age at diagnosis, per high-impact germline SNV (years) | p-value |
|---|---|---|
| *PiloAstro* | -2.07 | 0.012 |
| *ColoRect-AdenoCA* | -5.00 | 0.0083 |
| *Lung-AdenoCA* | -4.52 | 0.039 |
| *Lung-SCC* | 2.93 | 0.017 |
| *Ovary-AdenoCA* | -2.11 | 0.030 |
| *Stomach-AdenoCA* | -2.52 | 0.022 |

**Figure A. Classification of somatic variants into different categories based on their functional impact on cancer genome.** *Discovered standard drivers* are variants discovered through canonical driver identification methods employing statistical enrichment and/or functional impact of variants in cancer; *nominal passengers* are rest of the variant dataset. Most nominal passengers are *low-impact*, but some are predicted to be *impactful*, based off features such as the evolutionary conservation of the affected genomic positions, gains or losses of functional motifs, and measures of the biological importance of the affected

noncoding element or gene. Impactful variants divide into *latent drivers*, which promote cancer progression, and *deleterious passengers*, which inhibit tumor growth. Latent drivers further subdivide into three types: *Mini-drivers* incrementally promote tumor growth but are unnecessary for tumor survival. *Epistatic drivers* confer a meaningful selective advantage to tumors only in synergistic combination with other specific somatic or germline variants. *Undiscovered standard drivers* have a functional impact similar to the known driver variants reported in literature, but they have not yet been detected as drivers due to limited statistical power/cohort size.

**Figure B. Functional impact score distribution of non-coding SNVs present in the PCAWG pan-cancer variant set for a) somatic and b) germline genome**. We clearly observe three distinct subgroups based on somatic SNV impact distribution. Low and high impact score regions consist of neutral passenger mutations and putative driver mutations, respectively. In addition, we also observe *impactful passengers* with intermediate functional impact score. In contrast, functional score distribution for germline SNVs indicate presence of mostly neutral SNVs with very few mild/high impact variants.

**Figure C. Overall functional burdening of TF motifs:** TFs can undergo functional burdening via a) gain-of-motif (left panel) and b) loss-of-motif (right panel) events upon somatic mutation in various cancers. Functional burdening of TF motifs varies between different cancer types as well as within same cancer cohort. Further based on these functional burdening we can deduce influence of somatic variants on various regulatory networks. Red and green color in the heat-map signifies enrichment & depletion of high impact SNVs in various TF motifs, respectively. Columns in the heat map correspond to different cancer meta-cohorts, while the rows indicate different TF motifs.

**Figure D. Correlating functional impact and patient survival**: Survival curves in CLL (*top panel*) and RCC (*bottom panel*) with 95% confidence intervals, stratified by normalized impact burden. Patients in the top half of impact burden (in red), normalized by low-impact mutation load, die sooner in CLL but live longer in RCC, plausibly reflecting the different balance of latent drivers versus deleterious passengers in these two subtypes.

**Figure E. Impact of coding and non-coding mutations in tumor progression**: We divide each variant i) based on their prevalence in a parental-early subclone (in blue) vs child-late subclone (in red), and ii) based on their respective variant allele frequency (VAF). A high VAF (in blue) suggests early mutations or mutations in dominant/parental subclone, while low VAFs represent late mutations or mutations in low-fitness populations. In *left panel*, we show that the average number of LOF (*top*), non coding mutations (*middle*)and coding mutations (*bottom*)with moderate and high functional impact score per tumor sample. In *right panel*, we depict the average number of LOF(*top*), non-coding(*middle*) and coding mutations(*bottom*) with moderate and high impact that occurred in tumor suppressor gene (TSG) regions. We observe higher prevalence of impactful passenger and LOF variants in the early phases of cancer progression.

**Figure F. Correlating high impact germline variant with patient age**: Linear regression coefficients for predicting age in years at diagnosis of cancer from number of high-impact germline coding variants for six tumor subtypes. Negative coefficients represent a cancer predisposing effect; positive coefficients predict later cancer onset.