**Measuring the overall functional burdening of passenger mutations in cancer genomes -- in the PCAWG project.**

**Measuring the overall functional burdening & selective effect of passenger mutations in cancer genomes -- in the PCAWG project.**

**passenger mutations in 2800 cancer genomes :overall functional burdening & selective effects**

**Extended Abstract**

A typical tumor has thousands of genomic variants, yet very few of these (<5/tumor\cite{26559569}) are thought to drive tumor growth. The remaining variants, termed passengers, represent the majority of all variants in cancer, but their functional role is poorly understood. Furthermore, bulk of these passenger variants occupy the noncoding region of the genome. These passenger variants can be further subdivided into neutral and impactful based on their functional impact on the genome. As expected, low-impact passenger variants are inconsequential for tumor’s progression. However, impactful passenger variants can alter gene expression or activity, and while some of these changes are irrelevant, others may promote or inhibit tumor cell growth and survival, as is the case in *latent driver variants* and *deleterious passenger variants*, respectively. In this work, we explore the landscape of passenger variants in various cancer cohorts by leveraging extensive pan-cancer variation data from ~2500 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 functional elements in cancer genomes. Furthermore, we also show how overall functional burdening correlates with age at cancer diagnosis, patient survival time, and tumor clonality.

 One might expect that the functional impact score distribution for PCAWG variants will be unimodal distribution centered around 0 due to large amount of neutral passenger variants, along with a tail in high impact score regime corresponding to few putative driver variants. However, our observation depicts an alternative landscape of passenger variants. Inspection of variant functional impact scores across cancer cohorts, reveals that variants can be broadly classified into three distinct subgroups. The upper and the lower extreme of scored variants comprises of ~23 and ~13.5K noncoding variants per patient, which fall under traditional definitions of high impact putative driver variants and neutral passenger variants, respectively. In contrast, the intermediate functional impact regime comprises of *impactful* *passenger variants* (~3.5K noncoding variants), which can further influence cancer progression by acting as latent drivers or through large burdening of various functional elements of the genome. Interestingly, there is a significant depletion of such mildly deleterious variants in the germline of these cancer cohorts. As expected, the germline genome comprises of mostly neutral variants with very few high impact alterations. Consequently, we comprehensively analyzed the overall burdening of various genomic elements including TF (transcription factor) motifs in the PCAWG dataset. Presence of a variant in TF binding site can lead either to creation or destruction of TF motifs. In both cases, we observe differential burdening of *impactful variants* among different cancer cohorts. Furthermore, the selective enrichment/depletion of impactful *variants* in different TF motifs highlight distinct alteration profile associated with different component of regulatory networks.

Moreover, we integrated sub-clonal information along with functional burdening measure to investigate the role of impactful passenger variants in cancer evolution. Intuitively, one would expect that earlier variants in the cancer will be mostly enriched in mildly impactful passenger variants, while the later phase of cancer progression will involve few randomly created high impact driver variants. Our observation is consistent with the above hypothesis. We observed that high functional impact passenger variants along with LOF variants have a higher allelic frequency and a higher prevalence in parental subclones, signifying their important role in the early phases of cancer progression. We next sought to examine whether these functional mutations might exert a clinically meaningful impact on tumor progression. Therefore, we performed survival analysis to see if functional mutation burden predicted patient survival within individual cancer subtypes. Surprisingly, we discerned a statistically significant correlation between functional burden and patient survival for few cancer cohorts. However, these correlations varied substantially for different cancer types. For instance, we observed that functional mutation burden predicted substantially earlier death in CLL and substantially prolonged survival in RCC, respectively. These results lend support to the hypothesis that functional passenger mutations are clinically meaningful, at least in some cancer subtypes. More specifically, the results suggest that latent drivers are more important than deleterious passengers in CLL but that the situation is reversed in RCC. In conclusion, our work highlights an important subset of the coding and noncoding variants that originally identified as “passenger variants”, nonetheless show biologically and clinically relevant functional roles across a range of cancers.

[[this is particularly important for noncoding]]

*Selected v functional - impactful*

*Formally the passengers can be divided up thus....*

[[One might expect that distribution of impact scores to be uni-centered around 0 for the mostly neutral variants with a tail represnting the drivers.... However, this is not hte case ....]]

[[comprising XXX & YYY variants]]]

[[One might expect that earlier mutations in teh cancer are more impactful and later ones are more enriched in randomly created drivers... we do, in fact, see this]]

[[Surprisingly ,we observed a statistically signif pattern but it differented substantially betw cancers... ]]

**\* passenger burdening is obviously not uniform: diff parts of the regulatory network are burdened differently by passengers in cancers**

**\* the distribution of functional impact is not simple... there appears to be subgroups, one of which could correspond to latent drivers .... this is very different the functional impact in for germline**

**\* we can divide passengers into high and low impact... if we do this, in general, we can find that there are more high impact passengers in the early passengers than the late**

**\* furthermore, if we divide a given cancer, into two parts, one with lots of high impact passengers and another without, we can find differential survival**

* Avg cancer has ~10 drivers & ~5000 mutations. What is the overall burdening of the many passengers in different cancers?
* Overwhelming number of mutations uncovered by pcawg are in the passa...

[[ there are various hypothesis about these - neutral...

* [[ We attempt to address these questions....]][`
* Look at Overall variation burden observed in various genomic elements (coding & noncoding) in different PCAWG cohorts.
	+ Comparison between real and simulated data to highlight genomic elements with significant burden from passengers in different cohorts
		- A (coding)
		- B (coding LOF) - PDM

 ALoFT is a tool for classifying protein truncating events according to their probable impact. We used ALoFT to identify key gene loss of function events (LOFs) across all cancers, and in particular cancer types. Germline and somatic LOFs are correlated, clarifying how loss of gene function in germline influences subsequent somatic burdening in cancer. We show that cancer LOFs are enriched in genes essential to organismal survival and cellular processes. Examination of individual tumor samples with high burden of LOF events, gives insight into how cancerous processes test the limits of cell biology.

* + - C (non-coding)
		- D (SV enrichment) - YZ

 Overlap enrichment statistic between DNA sequence variants (SVs and SNPs) and genomic elements is a way to measure the impact of variants on the genome. We have set up a framework on high-performance computing clusters for overlap enrichment analysis, and has applied it to a normal human cohort (PMID: 26432246), in which we showed that SVs rarely overlap with conserved, functional genomic elements, and the results are in line with purifying selection. In this study, we will perform overlap enrichment analysis of SVs in various cancer categories (indolent vs. aggressive cancers), and investigate the impact of SVs on functional coding regions (under purifying selection), pseudogene regions (under neutral selection), and disease sensitive regions (ultraconserved regions and ultrasensitive regions). We will compare results in cancer samples and normal samples, and study the role of germline SVs and somatically acquired SVs in cancer propensity and development. [[Figure place holder: Panel A: Normal samples, Panel B: Cancer samples]]

* + This work will provide **comprehensive functional annotations across all of pcawg**  (FunSeq & aloft score)
		- A (deleterious passenger mutations)
		- B (LOF score distribution) -PDM

* Coding and noncoding functional impact score distribution across pan-cancer cohorts.
	+ Enrichment/depletion of high impact passengers (other than drivers) in gene block/neighborhood
		- TF gain/loss of motif/promoter/enhancer
	+ Correlation of passenger burdening with downstream gene expression changes
		- To-do list
	+ Framework to evaluate structural variation impact score
		- Large Del score distribution - AH

* Comparison between somatic and germline variation burdening
	+ Investigate influence of germline mutational burden on the somatic genome variation profile
		- A (germline/somatic burden)
		- B (co-occurrence of rare germline/somatic variants)

* Decipher the the differential passenger burdening in various cohorts (how it relates to mechanism)
	+ Relate to different Signature, Ageing, sub-clonality & other clinical information
		- A (sub-clonal analysis) - LS
		- B (survival analysis) - WM
			* Defined a functional burdening score for coding and noncoding regions of samples.
			* Applied a cox proportional hazards model with functional burdening score and patient age as covariates
			* Validated proportionality assumptions and robustness of coefficients to outliers
			* Observed a statistically significant and clinically relevant {protective effect / harmful effect} in {select subtypes}
			* This analysis was limited by the extent of available clinical information but nonetheless tends to support the hypothesis that functionally-significant passenger mutations {impose a fitness cost to tumor cells / can serve as latent drivers}, at least in {select subtypes}
			* <insert lovely KM-survival curve showing separation of patients into short-lived and long-lived cohorts upon stratification by functional burdening score>
		- C (signature – to do list)