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 divergent 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.

Deleted: contrasting

Deleted:

Formatted: Normal (Web), Justified

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 insertions and 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 explored the landscape of *putative passengers* in various cancer cohorts by leveraging extensive pan-cancer variant calls in PCAWG. More specifically, we built 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 differential impact on coding genes and regulatory elements, as elucidated by our signature analysis. We observed differential impact of putative passengers that may impact tumor progression. However, these *putative passenger* mutations may be **driven purely** by background processes or suggest non-neutral effect among *putative passengers*. Hence, we also considered ways of assessing a

Formatted: None

Deleted: An

Deleted: impact tumor progression, or have a

Formatted: Font:Italic

Formatted: Font:Bold

Deleted: dependent relationship to driver mutations. For example, a mutation in a DNA repair gene may contribute to cancer, while also generating many downstream passenger mutations that are merely correlated with cancer

possible non-neutral role for *putative passengers* on cancer progression. We found that the molecular functional impact burden of various genomic elements correlates with patient survival time and tumor clonality. We also found that *putative passengers* provide significant predictive power beyond common driver mutations to distinguish cancer phenotypes from non-cancerous ones, even after controlling for known mutational signatures and background mutation rate as possible confounders in our analysis. We determined that this effect is likely prominent among tumors without known drivers, or with fewer driver variants than expected. Although the effect of these possible driver variants can only be detected in aggregate by our model, it motivates future search for these variants among *putative passenger* variants.

Deleted: This suggests that putative passengers exert a modest but non-negligible effect on cancer development in aggregate.

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 {}) 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 range of this distribution corresponds to *putative passenger*s with intermediate molecular functional impact (**Fig 1a**).

Deleted: distribution

Deleted: peak

Subsequently, we investigated whether the frequency of medium- and high-impact noncoding *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).**

Formatted: Font:Not Italic

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 a 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 the 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 *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 seven categories of cancer-related functional annotation (**Fig 2a**). Driver LoF variants showed significant enrichment in six categories of cancer-related genes (cell cycle, immune response, cancer pathway, apoptosis, DNA repair and essential) relative to random expectation ($p < 0.001$). Conversely, non-driver LoFs displayed a small but significant depletion relative to random expectation in each of these categories except in metabolic and immune response genes, for which they showed slight enrichment compared to random expectation ($p < 0.001$) (supplement Fig. X).

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**

Deleted: six

Deleted: As expected, driver

Deleted: four

Deleted: and

Deleted: a

Deleted: (shuffled-variant) control

Deleted: ,

Deleted: ($p < 0.001$). However, non-driver LoFs

Deleted: essential

Deleted: were slightly enriched

Deleted: the

Deleted: .

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 a greater enrichment of germline SVs that engulf an entire functional element rather than for those that break a functional element partially[13]. Moreover, we observed the same pattern for somatic SVs.

Mutational process analysis

The differential burdening of various genomic elements may be attributed to the underlying stochastic but biased mutational processes. Thus, we closely inspected the underlying mutational signatures generating SNVs in both coding and non-coding regions of cancer genomes. First, we looked into the most impactful event, loss-of-function in coding regions. We would expect premature stop codons would show strong mutational contextual bias due to the nature of codon

Deleted: **Signature**

Deleted: either

Formatted: Tabs: 2.31", Left

Deleted: can

Deleted: random

Deleted: or selection on variants occupying these elements.

Deleted: For instance, one

Deleted: that mutational processes creating

Deleted: highly correlate with the number of LoF variants observed in a cancer sample. Indeed, we were able

Deleted: identify a high correlation between

Deleted: mutation spectrum and the number

Deleted: LoFs within some cancer types. However, these correlations are highly heterogeneous among different cancer cohorts, and the number of LoF

composition. Indeed, we found premature stop mutations carry specific mutational spectrum, which differs significantly from the overall tumor mutational spectrum. In particular, some mutations (e.g. T>Cs) cannot create premature stop as we expected. However, when compared with the pan-cancer premature stops, individual cancer shows a spectrum shift. For example, premature stops in RCCs shows a higher percentage of T>As compared to all cancers (18% versus 8%) (**Fig 3**). Our observation can be explained by the divergence of mutational processes in individual cancer types and implies mutational processes confer distinct effects in coding regions.

Similarly, the disproportionate functional load on certain TFs in cancers can be related to the underlying mutational spectrum influencing their binding sites. Different transcription factors have varying nucleotide context in their binding sites (TFBS). These variations in TFBS may facilitate the role of different mutational processes and will be reflected in their mutational spectrum. For instance, the mutational spectrum of motif breaking events observed in *SP1* TFBS suggests a 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{23456}). 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 a different proportion of signature 5 and signature 1(**Fig 3c**). We also scrutinized LoFs in coding regions, which carry the highest molecular function impact. Compared to *putative passengers*, signature 1 and 23 together contribute a relatively higher fraction to premature stop. 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 3d**). Collectively, these findings suggest that various mutational processes shape and disproportionally burden cancer genomes.

Deleted: 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

Deleted: sample (r=0.55 and 0.46 respectively) (see supplement table X). Other cancer cohorts such

Deleted: colorectal adenocarcinoma and non-Hodgkin lymphomas were able to withhold the majority of their LoFs with the ratio of observed vs

Deleted: close to 1

Deleted: 3a).

Deleted: This can be partially explained by the different nucleotide context among TF biding sites (TFBS).

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Highlight

Subclonal architecture and cancer progression

Deleted:

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 *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 by GERP score here) of each variant with their corresponding cellular prevalence measured by 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 (**Fig 4c**). 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 of 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

The results we have found may be explained in relation to underlying mutational processes. However, they may also be indicative of selective effects among subset of these mutations, whether or not they are generated by a neutral mutational process. If indeed a subset of putative passengers possess fitness effects, 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 molecular functional impact (i.e. effect on gene expression or activity). The relevance of molecular functional impact is firmly established for 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

Deleted: 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, if all putative passengers in a cancer cohort had completely neutral fitness effects, one would expect their molecular functional impact to be distributed uniformly. On the contrary, the multi-modal functional impact distribution of non-coding mutations indicate that a subset of mutations among *putative passengers* might confer potentially weak fitness effect to tumors. Nonetheless, this observation can be also attributed to underlying mutational signatures. Similarly, in a completely neutral putative passenger model, we would expect no correlation between molecular functional impact and patient survival. In contrast, our analysis suggests strong correlation between differential molecular functional impact of putative passengers and patient survival in certain cancer cohorts. This can be interpreted as the presence of weak selection among putative passengers in these cohorts. However, the presence of distinct cancer subtypes within these cancer cohorts could be an alternative explanation. In addition, differential burdening of distinct genomic elements in cancer can be associated with the operation of various signatures, which in itself is interesting. Nevertheless, 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]. ... [1]

Formatted: Indent: First line: 0"

in ways that are not ultimately relevant for tumor fitness; hence, these variants will undergo neutral evolution.

An initial step towards identifying the presence of variants with effects on tumor fitness is to compare observed mutation distributions with ones generated by simulating or modeling neutral processes. This approach has been extensively leveraged in the context of individual driver discovery using element burden testing. Such an approach is potentially powerful since it allows the use of complex background mutational models, although the possibility of detecting artifacts due to the inadequacy of current models of neutral mutational processes remains, since unmodeled mutation process may result in confounding effects. With this caveat, we explore such an approach below in an attempt to quantify non-neutral aggregate effects among putative passengers, using a variety of recent neutral models and an additive model which combines both positive and negative fitness effects. As in the case of individual driver discovery, validation of such effects requires follow-up experimentation.

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]. Although, we do not currently have estimates of 'missing heritability' at the subclone level for tumorigenicity, which may depend on both genetic and epigenetic factors. However, the fact that many tumors lack a known driver (\cite{Nuria's paper}) suggests that some driver mutations remains to be discovered. The models above suggest the importance of investigating the cumulative effect of *putative passengers* in this context.

To address this, we adapted an additive effects model[6,10], originally used in complex trait analysis, to quantify the relative size the aggregated effect of *putative passengers* 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

Deleted:

Deleted: A recently proposed "omnigenic model" takes this logic a step further, arguing that the majority of complex traits are influenced by thousands of variants with individually small effects[9]. Although we do not currently have estimates of 'missing heritability' at the subclone level for tumorigenicity (which may depend on both genetic and epigenetic factors), the fact that many tumors lack a known driver (\cite{Nuria's paper}) suggests that at least a portion of this heritability remains to be discovered, while the

Deleted: of these

Deleted: effects

ability of this model to predict cancerous from null samples as a binary phenotypic trait (**Fig 5b**). Briefly, we created a balanced dataset of observed tumor and matched neutral (null) model samples, using a 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 datasets as well as our custom randomization correcting for various covariates (see supplemental method).

We compared several versions of the additive variance model above in 8 cancer cohorts having a 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. 5ci**). *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). This approach effectively isolates the effect of *putative passengers* in tumors without driver mutations. 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. 5cii**). Here, we observed that the coding mutations accounted for by far the largest overall proportion of the variance (~50.7% averaged across cohorts), while promoters and other non-coding also contributed much lesser, but still 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. 5ciii**) 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 repeating these analyses for two other randomization schemes, with quantitatively similar results (Supp Fig. X).

By including a binary predictor for known driver SNVs in the above model, we expect the contribution of the *putative passengers* to be higher among samples without known drivers (as well as all null samples). To confirm that the *putative passengers* were indeed contributing to the discrimination of samples without known drivers, we further calculated the additive variance exclusively for such samples in PCAWG. For these samples, we observed an average of 12.5% additive variance across cohorts (Supp Fig. X), which was higher than the 9.5% additive variance estimates based on *putative passengers* among all samples (with and without known drivers). This observation is consistent with a more important role for the *putative passengers* among samples without know driver, since they may have partially redundant effects in the samples harboring known drivers. Furthermore, we calculated the additive variance after excluding samples with driver SVs and CNAs alterations in addition to samples with known driver SNVs. This analysis was performed only for pan-cancer meta-cohort which pools all such samples (Supp Fig. X). We observed lower amount of additive variance (6.1%) for the pan-cancer meta cohort. The lower estimate for the pan-cancer cohort may be due to tissue-specific effects which are lost at the meta-cohort level. Finally, we estimated the Best Linear Unbiased Predictor (BLUP) for each cohort, and used this to derive an estimate of the number of weak drivers among samples with all discovered known drivers excluded (details in Online Methods). Using this approach, we estimated an average of 8.4 weak drivers per cohort. We expect that these estimates are limited by sample size, and thus represent lower bounds.

Deleted: . ... [2]

Discussion

Certain key alterations in the tumor genome, often identified through the detection of strong signals of positive selection on individual variants, have been shown to play a 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

Deleted: 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.

functional consequences are poorly understood. In this work, we comprehensively characterized *putative passengers* in the PCAWG dataset. We came across multiple lines of evidence, which suggested the 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 cumulative effect of *putative passengers* on tumor progression. 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 cause, are indirectly associated with, or are independent of 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 currently identified as *putative passengers* nonetheless shows biologically and clinically relevant functional roles across a range of cancers.

Deleted: As described earlier, we

Deleted: .

Deleted: weak positive and negative selection among

Deleted: originally

Formatted: Font color: R,G,B (0,0,10), Not Highlight

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

Formatted: Indent: First line: 0"

References

1. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
2. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17,** 93–108 (2016).
3. Vogelstein, B. & Kinzler, K. W. The Path to Cancer — Three Strikes and You're Out. *N. Engl. J. Med.* **373,** 1895–1898 (2015).
4. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15,** 680–685 (2015).
5. McFarland, C. D., Korolev, K. S., Kryukov, G. V, Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2910–5 (2013).
6. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–9 (2010).
7. International Schizophrenia Consortium, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–52 (2009).
8. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29,** 150–159 (2013).
9. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
10. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
11. Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15,** 480 (2014).
12. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–33 (2013).
13. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
14. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–13 (2005).
15. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
16. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* (2017). doi:10.1038/nature22992

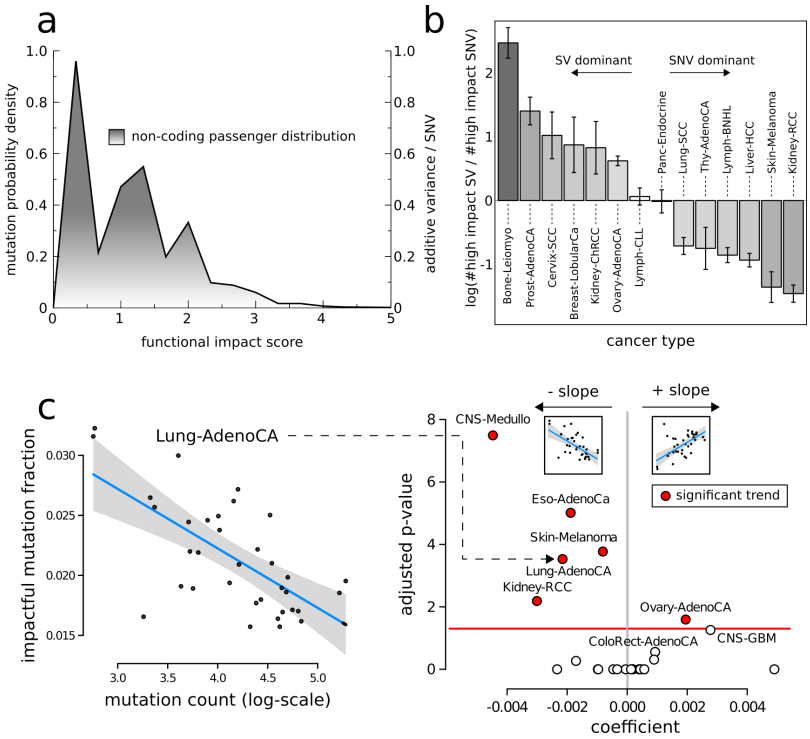


Figure 1: Overall functional impact of PCAWG variants: a) Functional impact distribution in noncoding region: three peaks correspond to low-, medium- and high-impact variants. **b)** log ratio of high-impact structural variants(SVs) and SNVs in different cancer cohorts. **c)** Correlation between number of impactful and total SNV frequencies for different cohorts.

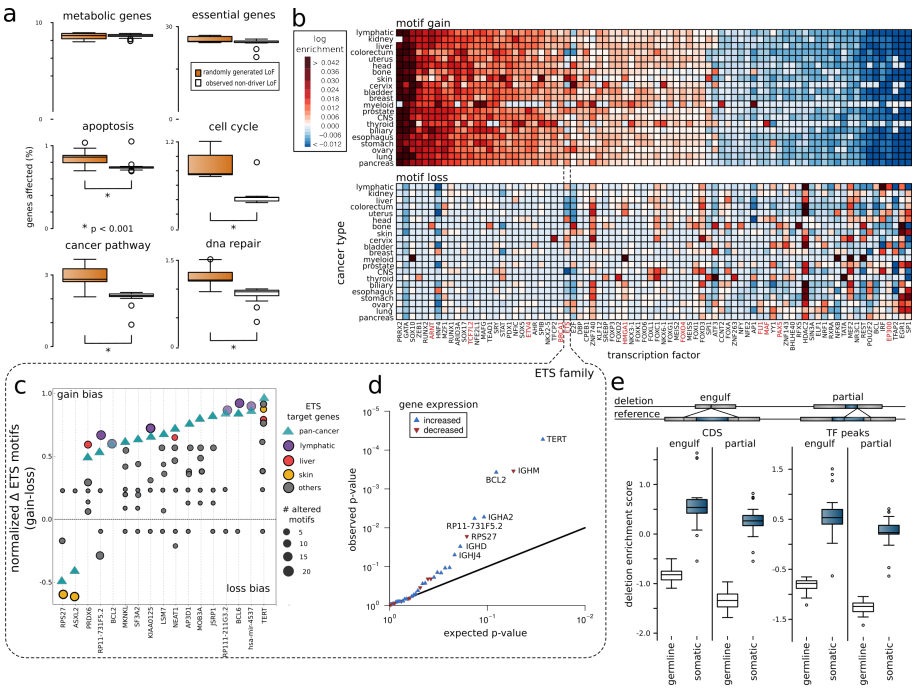
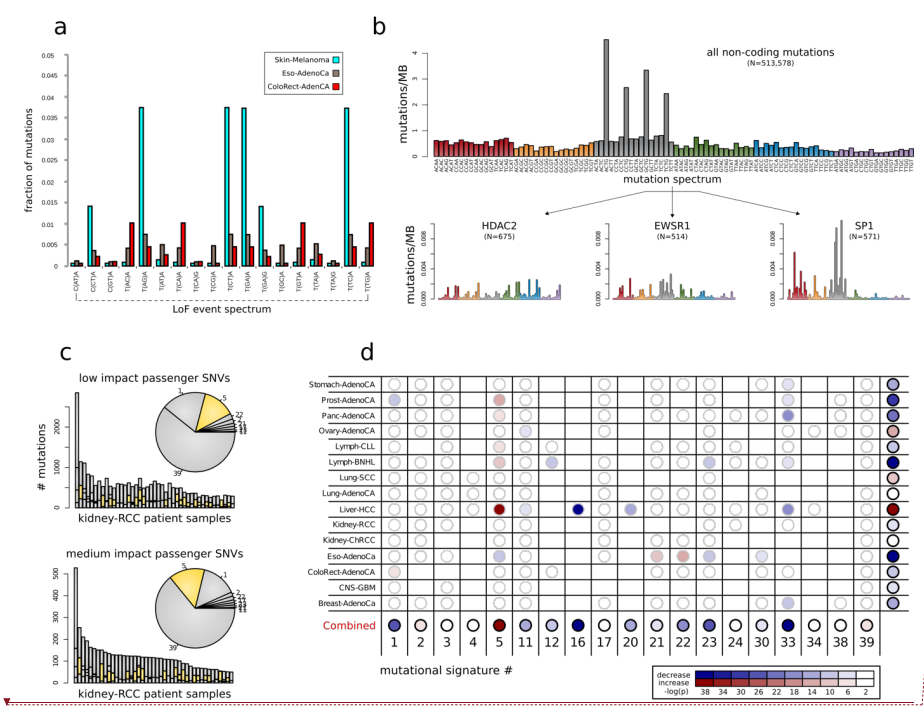


Figure 2: Overall functional burdening of different genomic elements: a) Percentage of genes in different gene categories (apoptosis, cell cycle, cancer pathway, dna repair, metabolic and essential genes) affected by non-driver LoFs in observed and random model, **b)** *Pan-cancer overview of TFs burdening:*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



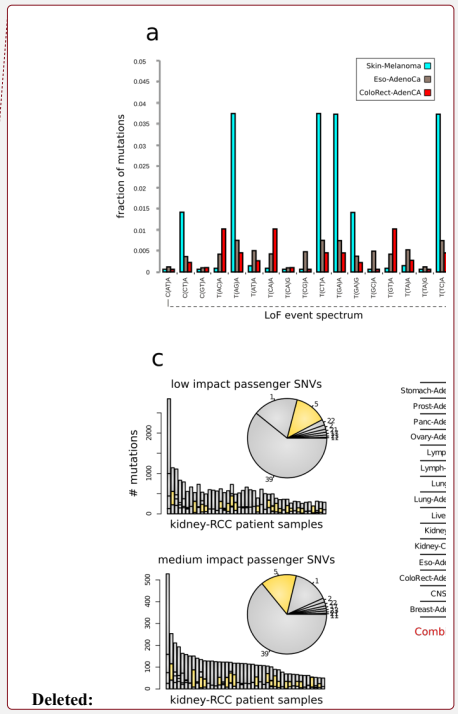


Figure 3. Mutational signatures associated with different categories of impactful variants: a),b) Mutation spectra associated with motif breaking events observed in HDAC2, EWSR1 and SP1 in the kidney-RCC cohort. **c)** Distribution of canonical signatures in the kidney-chRCC cohort for premature stops (top), impactful (middle) and low-impact SNVs (bottom). **d)** Differences in underlying signatures between high-and low-impact nominal passengers in different cancer cohorts.

Deleted: 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.

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

Deleted: RCC

Deleted: bottom

Deleted: top

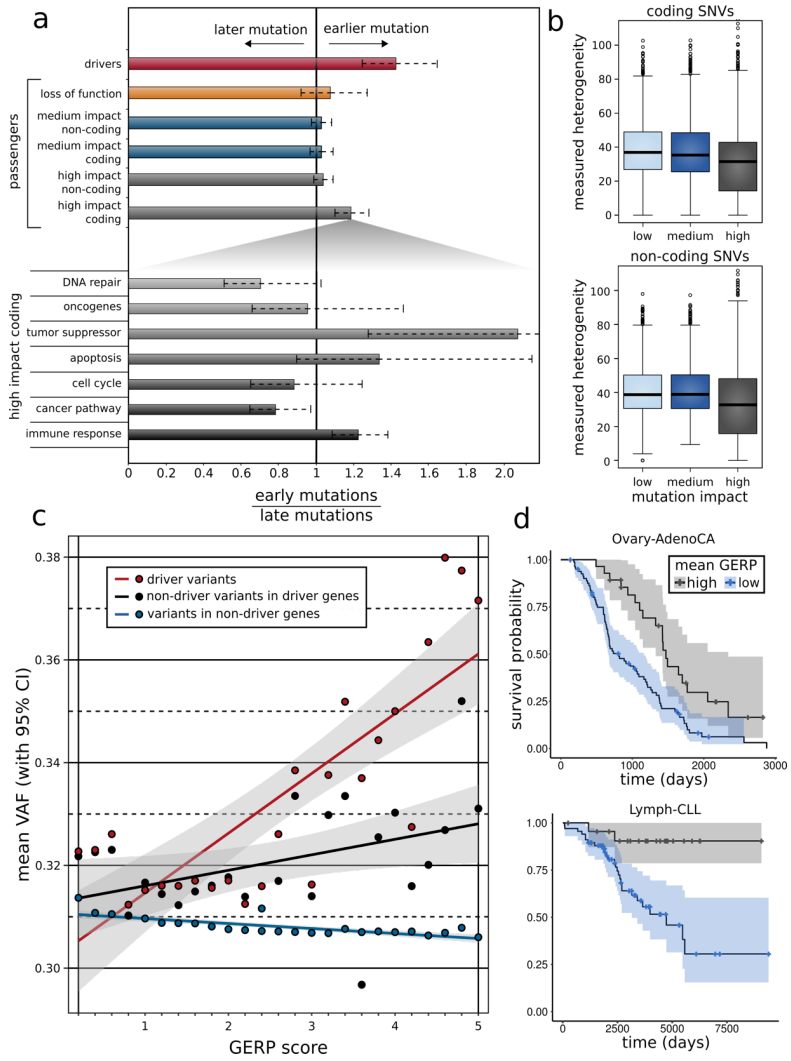


Figure 4: Correlating functional burdening with subclonal information and patient survival: a) Subclonal ratio (early/late) for different categories of SNVs (coding/non-coding) based on their impact score. Subclonal ratio for high-impact SNVs occupying distinct gene sets. **b)** Mutant tumor allele heterogeneity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

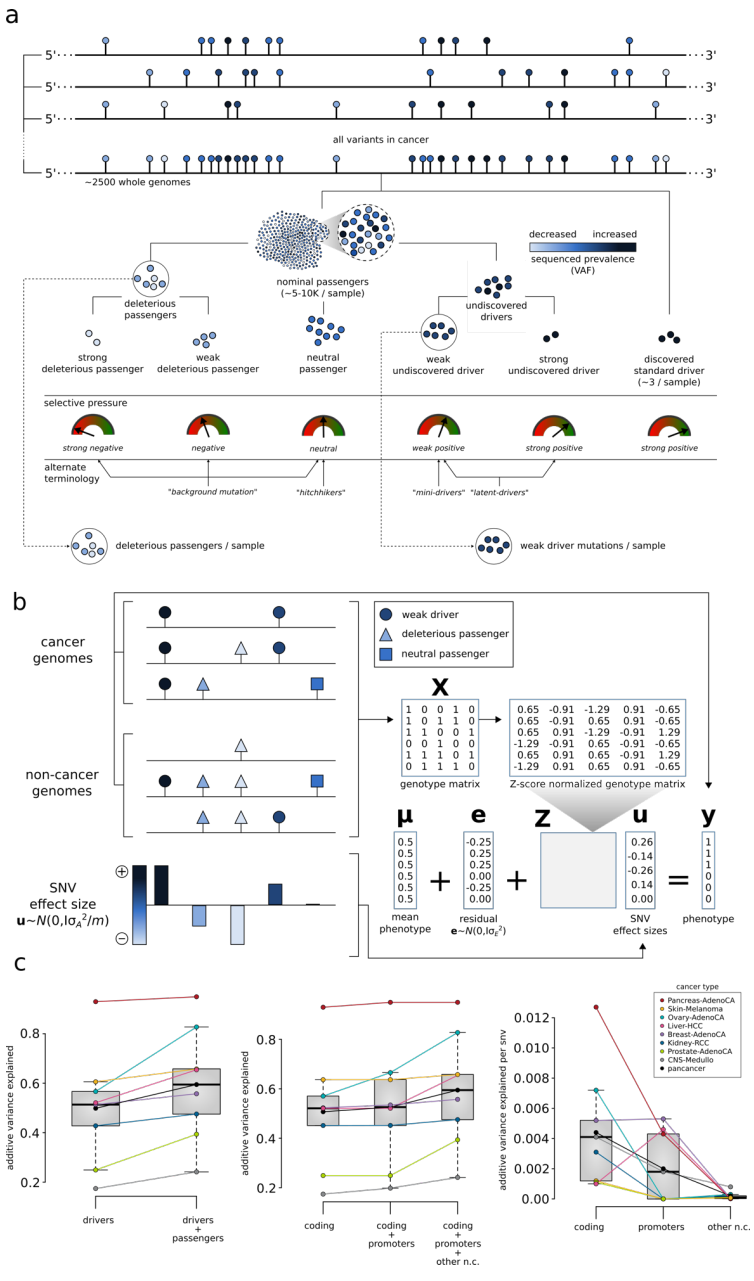

a
5'
3'
all variants in cancer
~2500 whole genomes
nominal passengers
(~5-10K / sample)
decreased
increased
sequenced prevalence
(VAF)
deleterious
passengers
undiscovered
drivers
strong
deleterious passenger
weak
deleterious passenger
neutral
passenger
weak
undiscovered driver
strong
undiscovered driver
discovered
standard driver
(~3 / sample)
selective pressure
strong negative
negative
neutral
weak positive
strong positive
strong positive
alternate
terminology
"background mutation"
"hitchhikers"
"mini-drivers"
"latent-drivers"
deleterious passengers / sample
weak driver mutations / sample
b
cancer
genomes
non-cancer
genomes
SNV
effect size
u~N(0,Iσ_A²/m)
weak driver
deleterious passenger
neutral passenger
X
genotype matrix
Z-score normalized genotype matrix
μ
e
z
u
y
mean
phenotype
residual
e~N(0,Iσ_e²)
SNV
effect sizes
phenotype
c
additive variance explained
drivers
drivers
+
passengers
coding
coding
+
promoters
coding
+
promoters
+
other n.c.
additive variance explained per snv
coding
promoters
other n.c.
cancer type
Pancreas-AdenoCA
Skin-Melanoma
Ovary-AdenoCA
Liver-HCC
Breast-AdenoCA
Kidney-RCC
Prostate-AdenoCA
CNS-Medullo
pancancer

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 *putative 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)** *Additive effects model for nominal passengers*: The combined effects of many nominal passengers are modeled using a linear model, which predicts whether a genotype arises from an observed cancer sample or from a null (neutral) model (notation defined in text). The model is fitted by optimizing the hyper-parameter σ_A^2, and a test for significant combined effects of the nominal passengers is made by performing a log-likelihood ratio test against a restricted model which includes only μ and e. **c)** *Predictive power of known drivers and nominal passengers using the additive effects model*: Left figure compares the maximum possible variance which can be explained using known drivers with the performance of the model with those from driver and putative passengers. Central figure further breakdown this into contribution from coding, coding & promoter and everything. Right figure presents normalized additive variance explained by *putative passengers* in coding only, promoter only and other non-coding only elements of the genome.

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, if all putative passengers in a cancer cohort had completely neutral fitness effects, one would expect their molecular functional impact to be distributed uniformly. On the contrary, the multi-modal functional impact distribution of non-coding mutations indicate that a subset of mutations among *putative passengers* might confer potentially weak fitness effect to tumors. Nonetheless, this observation can be also attributed to underlying mutational signatures. Similarly, in a completely neutral putative passenger model, we would expect no correlation between molecular functional impact and patient survival. In contrast, our analysis suggests strong correlation between differential molecular functional impact of putative passengers and patient survival in certain cancer cohorts. This can be interpreted as the presence of weak selection among putative passengers in these cohorts. However, the presence of distinct cancer subtypes within these cancer cohorts could be an alternative explanation. In addition, differential burdening of distinct genomic elements in cancer can be associated with the operation of various signatures, which in itself is interesting. Nevertheless, 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, a close inspection of putative passengers in early and late subclones suggests that similar to driver mutations, high impact putative passengers are slightly enriched among earlier subclones. This enrichment is particularly strong among TSGs. In contrast, high impact *putative passengers* affecting oncogenes are depleted in earlier subclones. An interpretation of this finding is that putative passengers in tumor suppressor genes may have potentially weak driver activity, while those in oncogenes impair oncogenic activity to the detriment of tumor fitness. Similarly, positive and negative correlation between the conservation score of *putative passengers* and their corresponding VAF, suggests the presence of weak positive and negative fitness effects among a subset of these mutations. However, we note that differences in signatures can potentially confound the presence of high impact putative passengers between early and late subclones.

If indeed a subset of putative passengers possess weak

Page 11: [2] Deleted	Microsoft Office User	11/18/17 2:56:00 PM