Passenger mutations in >2500 cancer genomes: Overall functional impact & consequences

Summary

- **This is a unique study which evaluate impact of each variant in a cancer including passengers to decipher the overall functional burdening of different genomic elements in cancer genome.**
- **Overall functional impact distribution of PCAWG SNVs shows that in addition to high impact drivers and low-impact passengers there are group of medium impact passengers variants.**
- **We observed that the functional burdening relates to the underlying signature and thus different signatures contribute to the functional burdening to different extent.**
- **we observed that burdening is non-random in terms of the functional subsystems and for different categories of genes.**
- **we find that functional burdening varies based on subclonal architecture and further can be related to survivability of patients.**
- **Finally, we speculate on how the differential burdening might be related to both weak positive and negative selection during tumor evolution.**

To a first approximation, all clinically significant consequences of genomic variants in cancer are mediated through their functional impact, such as changes in gene expression or gene activity. Certain key alterations in tumor genome, often identified through the detection of strong signals of positive selection on individual variants, have been shown to play pivotal role in tumor progression. Although a typical tumor has thousands of genomic variants, 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. Recent studies have proposed that some passenger variants may impact tumor cell biology along a range of dimensions and weakly affect tumor cell fitness by promoting (*latent driver variants* [2,3] or inhibiting (*deleterious passengers*[4]) tumor growth. If there are indeed any yet-unidentified weak latent driver mutations buried among the nominal passenger variants, or if there are any passenger variants that are deleterious to tumor cell fitness, we would expect their effects to be mediated through their functional impact on tumor cell biology.

In this work, we explore the functional landscape of passenger variants in various cancer cohorts by leveraging extensive pan-cancer variant calls from ~2500 uniformly processed whole cancer genomes. More specifically, we build on and apply existing tools to annotate and score the predicted functional 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. We show that disruption of genetic regulatory elements in the noncoding genome

correlates with altered gene expression. Furthermore, we also show how overall functional burdening of various genomic elements correlate with age at cancer diagnosis, patient survival time, and tumor clonality. Finally, we observe statistical signals consistent with the notion that aggregated subsets of passenger variants – particularly those we predict to be functionally impactful- confer weak selective effects. We offer our variant-level predictions of functional impact as a uniformly-annotated resource.

In order to substantiate the presence of impactful passenger and their role in cancer progression, we surveyed the functional impact distribution of somatic variants in different cancer genomes. The functional impact distribution varies among different cancer types and different genomic elements. For instance, impact score distributions of non-coding variants occupying DHS region across different cancer genomes indicate 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 functional impact regime corresponds to what we term *impactful passengers*.

According to a simple random expectation, one would assume that the overall functional burdening in a cancer genome will be uniformly distributed across different functional elements and among different gene categories. In contrast, we observe that the functional burdening in certain cancers is concentrated in particular gene categories. In particular, in figure 3B, we show that medium-impact variants tend to occur in essential genes more often compared to low impact variants. Conversely, low impact passengers constitute larger fractions of variants influencing non-essential genes. We observe similar trend among various key gene categories, such as metabolic and immune response genes.

Furthermore, in the uniform model, we would expect that the fraction of impactful variant will remain constant as one accumulate large amount mutation in certain cancer sample. In contrast, we observe that as we accumulate more mutations in cancer, the fraction of impactful mutations decreases suggesting that a tumor can accommodate impactful variants to a certain extent. This trend is particularly strong and in CNS medulloblastoma ($p < 4e-8$, Bonferroni's correction), lung adenocarcinoma ($p<3e-4$, Bonferroni's correction), and a few other cancers (Figure3 B).

One might further expect that passenger variants will be uniformly distributed contributing uniform functional burden across the genome. Consequently, we comprehensively analyzed the overall mutational burdening of various genomic elements, including TF (transcription factor) binding motifs in various cancer genomes. The presence of a variant within a TF binding site(TFBS) 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 TFBS among different cancer cohorts. For instance, we observe significant enrichment of high impact variants creating new motifs in various TFs such as GATA, PRRX2 and SOX10 across major cancer types analyzed in this study. Similarly, high impact variants influencing gene expression by breaking TF motifs, were highly enriched in TFs such as YY1, BCL, RAD21 and CTCF in a majority of cohorts. This selective enrichment or depletion suggests distinct alteration profiles associated with different components of regulatory networks in various cancers. Furthermore, enrichment of SNVs in selective TF motifs leading to gain and break events in promoter significantly perturb the downstream gene expression. For instance, in lung adenocarcinoma, we found three TFBSs gain events (ZBTB14, E2F and HNF4) significantly increase downstream expression level ($p<5e-7$, 3e-6 and 2e-4 respectively).

The disproportionate burdening of certain TFs in different cancers can be further related to the underlying mutational spectrum of variants influencing their binding sites. For instance, mutation spectrum of motif breaking events observed in SP1 TF binding sites (TFBS) suggest major contribution

from C>T and C>A mutation. In contrast, motif breaking events at TFBS of HDAC2 and EWSR1 have relatively uniform mutation spectrum profiles. Similarly, comparing signature composition of low and high impact SNVs in certain cancer-cohort can help us to distinguish between mutational processes that generate these distinct classes of variants. For instance, we observed distinct signature distributions for the low and high impact non-coding passengers in the Kidney-RCC cohort. While the majority of passengers can be explained by signature 5, high impact passengers have a higher fraction of SNVs explained by signature 4. This suggests that certain fraction of high impact passenger SNVs show shifts in mutational signatures compared to the lower ones.

Additionally, we explored the role of impactful variants in cancer evolution by integrating their sub-clonality information. Intuitively, one might hypothesize that high impact mutations should either achieve higher prevalence in tumor cells if they are advantageous to the tumor, or a lower prevalence if deleterious. Interestingly, one finds suggestive evidences corroborating this hypothesis. In particular, we observe that high functional impact passenger variants in coding regions have higher pervasiveness among parental subclones, signifying their important role in the early phases of cancer progression by providing a higher fitness advantage to the cancer cell. In addition, we closely inspected the subclonal ratio (ratio of SNVs in early subclones compared to late subclones) of high impact passenger variants with respect to low impact passenger variants for different gene categories. We observed that high impact passenger SNVs in tumor suppressor and apoptotic gene regions show enrichment in early subclones and thus promote fitness of tumor cells leading to cancer progression. In contrast, high impact passenger SNVs in oncogenes appear slightly depleted, suggesting that presence of these variants might have an adverse effect on oncogene functionality, while burdening the tumor cells. Moreover, these impactful SNVs in DNA repair and cell cycle genes show depletion in early subclones, suggesting that a high impact variant might eventually provide a critical burden for the survival of tumor cell. This observation is consistent with prior studies highlighting role of deleterious passengers inhibiting cancer progression.

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. One measure of the fitness impact of a set of variants is their mean variant allele frequency (VAF), which reflects the degree to which the involved variants have spread through their respective tumors. We would expect for variants that enhance tumor cell fitness to achieve high mean VAF, variants that reduce tumor cell fitness to occur at low mean VAF, and neutral variants to occur at intermediate mean VAF. We observe that driver SNVs occur at high mean VAF, non-silent coding mutations and noncoding variants in sensitive regions occur at low mean VAF, and synonymous variants along with variants in intergenomic regions occur at intermediate mean VAF. The most straightforward interpretation of these findings is that, in aggregate, non-silent passenger variants and noncoding variants in sensitive regions impair cancer cell fitness.

We further generalize our observations of the VAF distribution among functional classes by computing the VAF-GERP correlation among variants in driver genes and among variants not in driver genes. GERP is a nucleotide-level measure of the inter-species sequence conservation at genomic positions and provides a continuous alternative to binning variants into functional impact classes. Highly conserved positions (i.e. those with high GERP) are believed to be important for organismal fitness, in some cases because polymorphisms at those positions would hurt cellular fitness and in other cases because polymorphisms at those positions would 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).

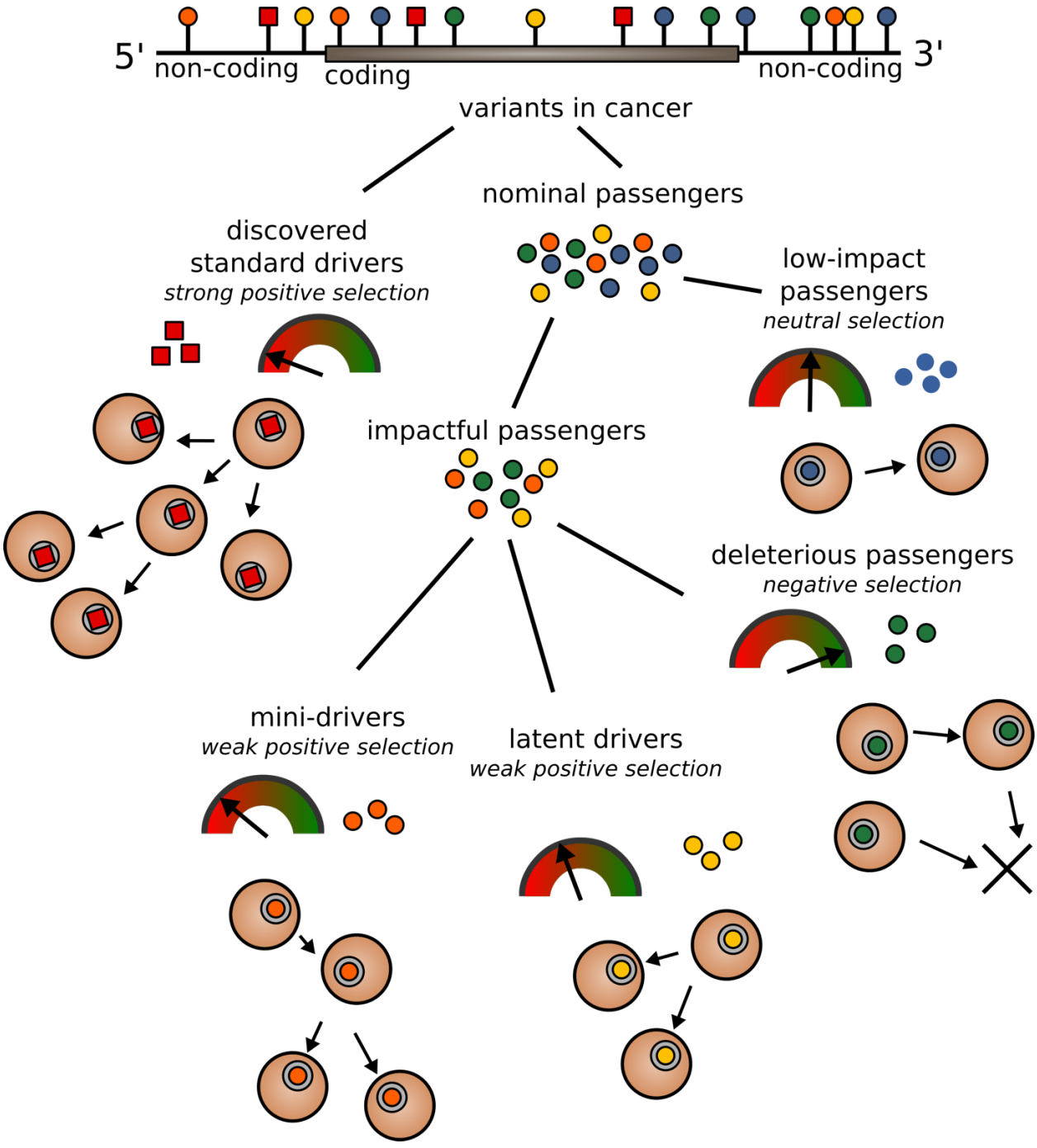
Interestingly, VAF and GERP have a correlation of similar magnitude but in opposite direction among variants not in driver genes, with very high significance (coefficient -0.0034, p.value < 2.2e-16). The observed trend for passenger variants at more conserved positions to occur at lower VAF is consistent with the deleterious passenger hypothesis.

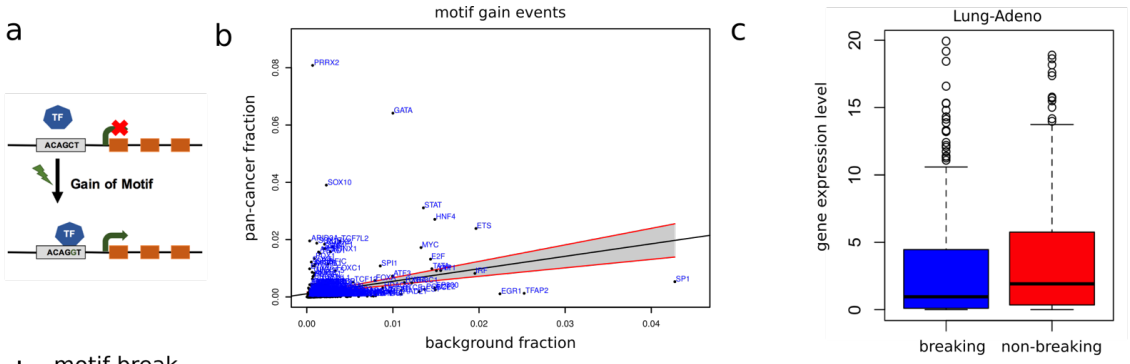
Finally, we sought to examine whether impactful passengers might be associated with tumor initiation and progression. Therefore, we correlated patient impactful somatic mutation burden with patient survival and age at diagnosis with their impactful germline mutation burden. We observed that patients harboring a larger number of high-impact rare germline alleles were diagnosed with cancer at earlier ages in three cancer subtypes. We then performed survival analysis to see if somatic impact burden –the ranked sum of the impact scores of coding and noncoding variants – 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 observations remained after redefining somatic impact burden in relation to the burdening of corresponding randomized sets. Furthermore, these patterns remained after adjusting for patient age at diagnosis, low-impact mutation load, and –in the case of CLL, including a covariate for IgVH mutation status. These results lend support to the hypothesis that the aggregate amount of impactful passengers is clinically meaningful. More specifically, these results suggest that latent drivers are more important than deleterious passengers in CLL, but that the situation is reversed in RCC.

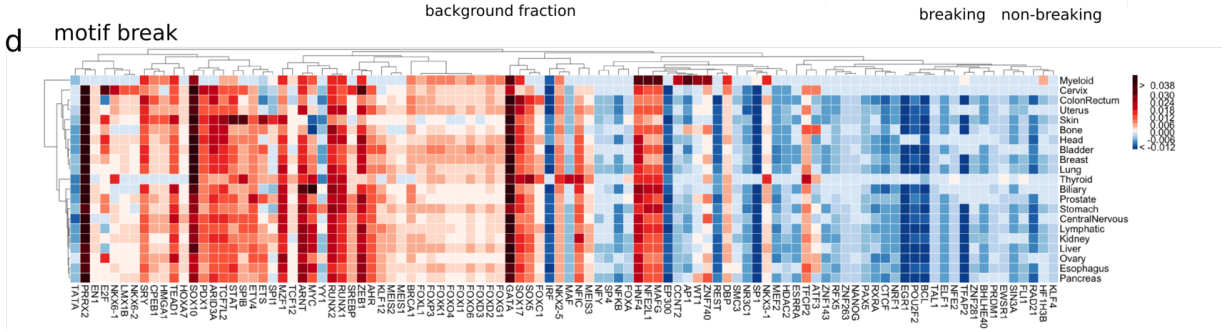
In conclusion, our work highlights that an important subset of somatic variants originally identified as passengers nonetheless show biologically and clinically relevant functional roles across a range of cancers.

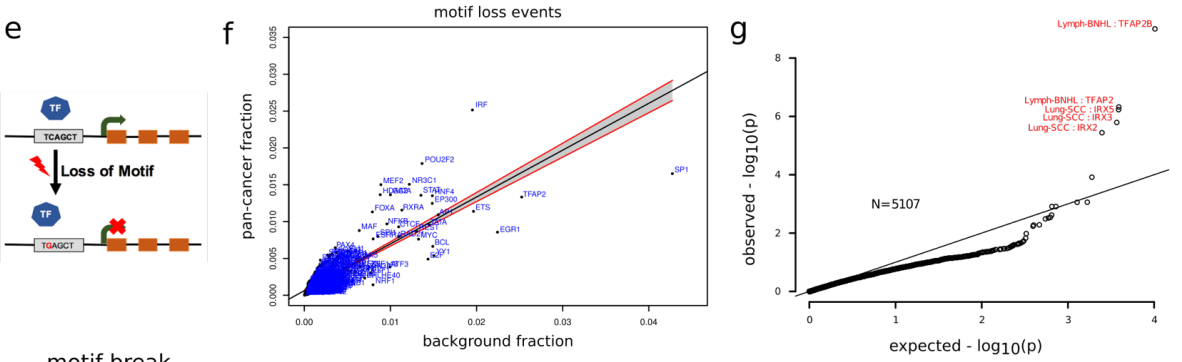
References

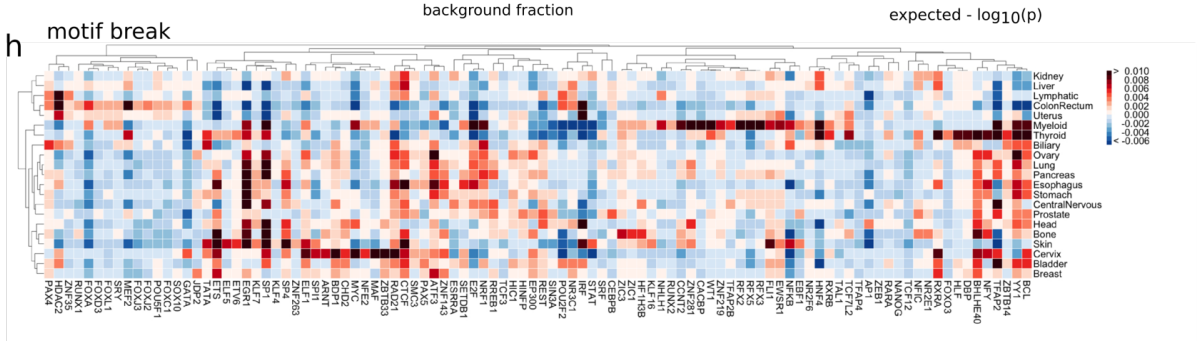
1. Vogelstein, B. & Kinzler, K. W. The Path to Cancer --Three Strikes and You're Out. *N. Engl. J. Med.* **373,** 1895–8 (2015).
2. Nussinov, R. & Tsai, C. J. 'Latent drivers' expand the cancer mutational landscape. *Current Opinion in Structural Biology* **32,** 25–32 (2015).
3. Castro-Giner, F., Ratcliffe, 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).

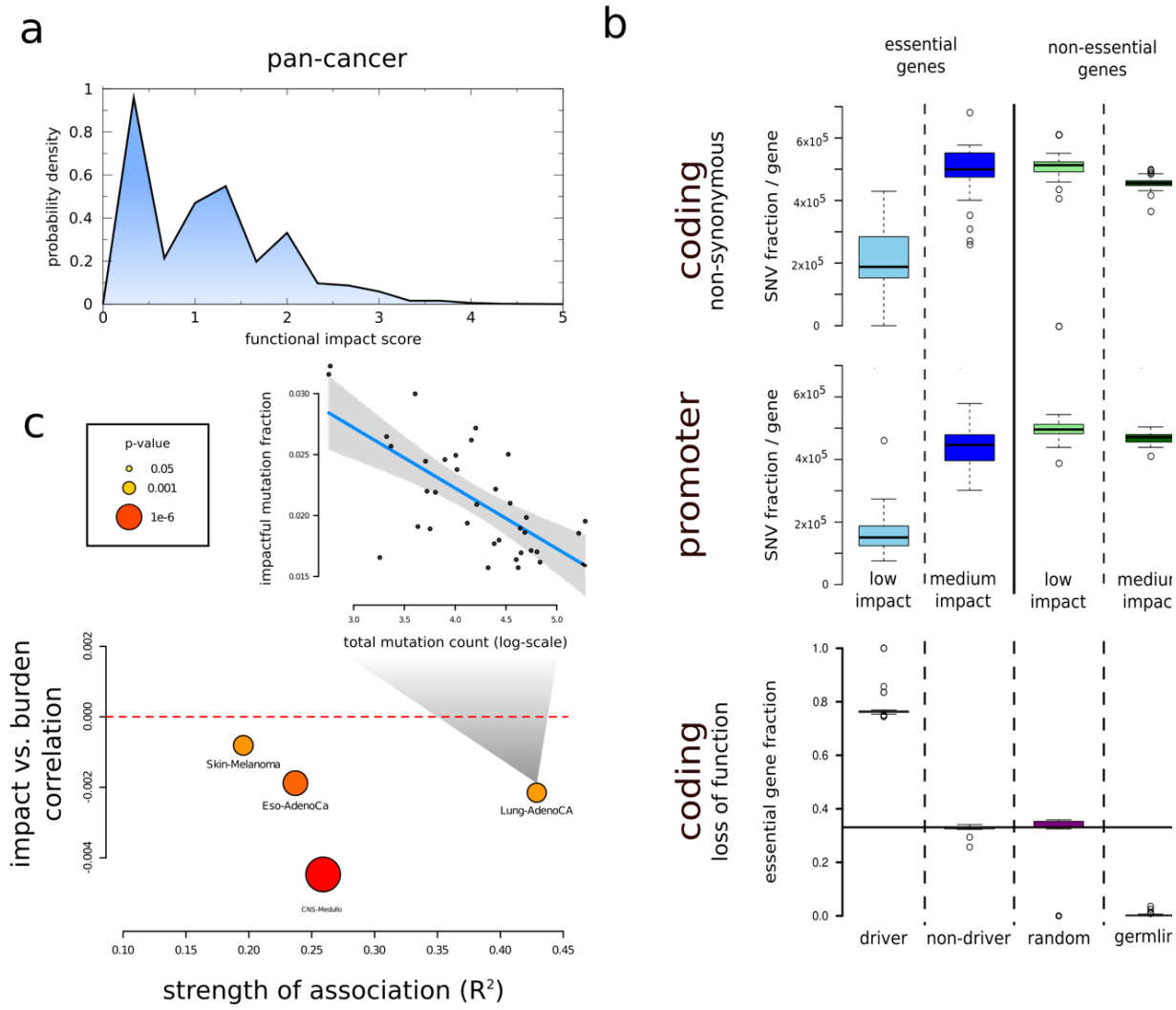

5'
non-coding
coding
non-coding
3'
variants in cancer
nominal passengers
discovered
standard drivers
strong positive selection
low-impact
passengers
neutral selection
impactful passengers
deleterious passengers
negative selection
mini-drivers
weak positive selection
latent drivers
weak positive selection

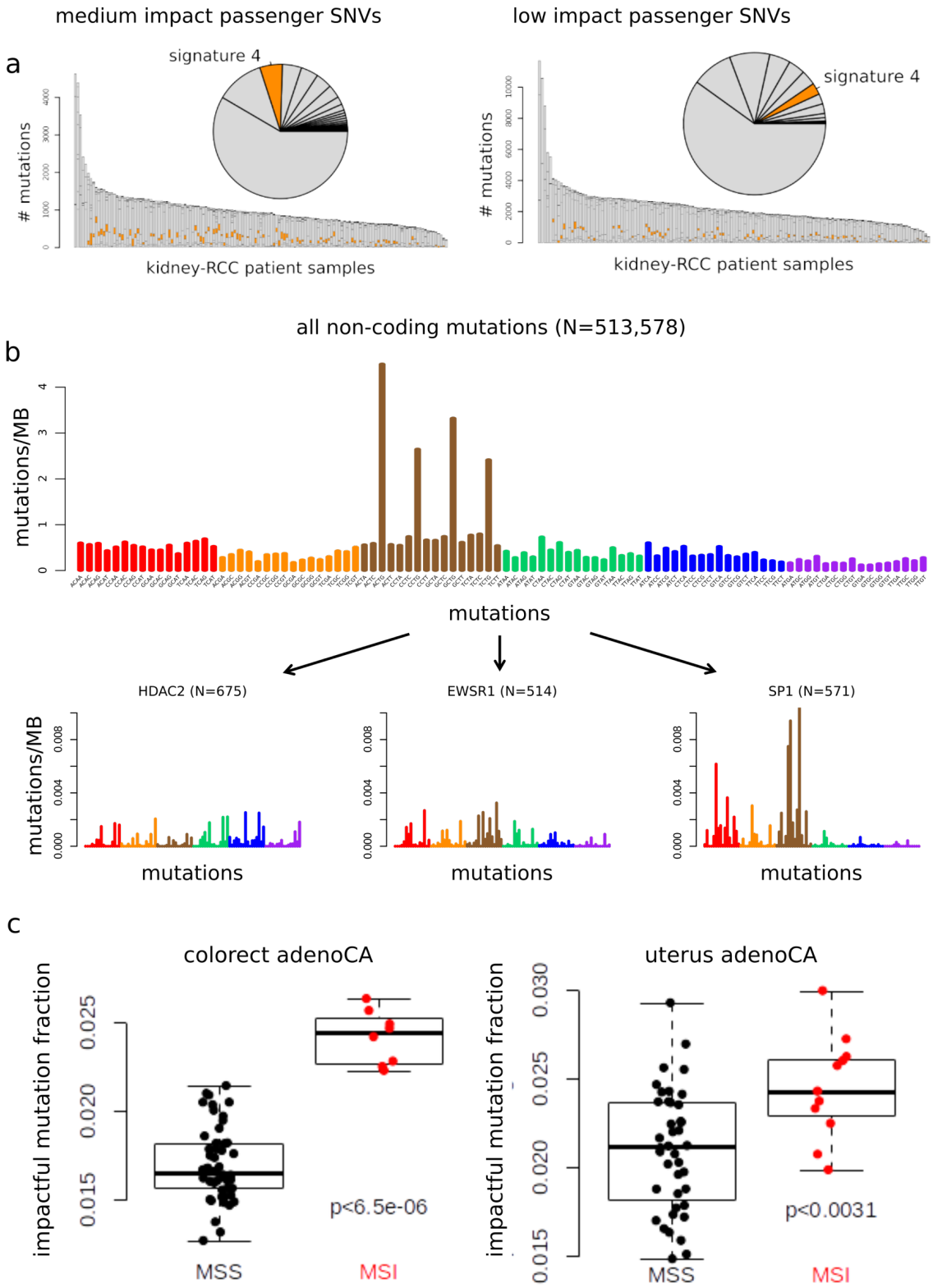

a
b
c
TF
ACAGCT
Gain of Motif
ACAGGT
motif gain events
pan-cancer fraction
background fraction
PRRX2
GATA
SOX10
STAT
HNF4
ETS
MYC
E2F
SP1
EGR1
TFAP2
Lung-Adeno
gene expression level
breaking
non-breaking

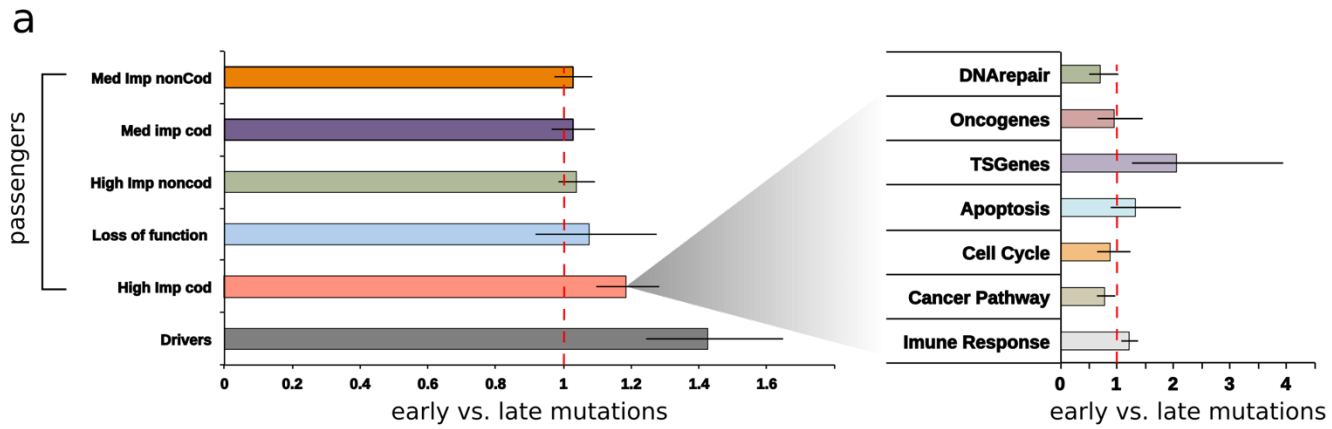

d
motif break
Myeloid
Cervix
ColonRectum
Uterus
Skin
Bone
Head
Bladder
Breast
Lung
Thyroid
Biliary
Prostate
Stomach
CentralNervous
Lymphatic
Kidney
Liver
Ovary
Esophagus
Pancreas

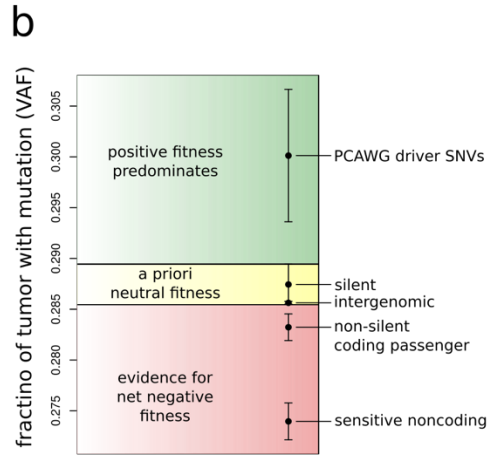

e
f
g
TF
TCAGCT
Loss of Motif
TGAGCT
motif loss events
pan-cancer fraction
background fraction
IRF
POU2F2
SP1
TFAP2
ETS
EGR1
Lymph-BNHL : TFAP2B
Lung-SCC : IRX5
Lung-SCC : IRX3
Lung-SCC : IRX2
N=5107
observed - log10(p)
expected - log10(p)

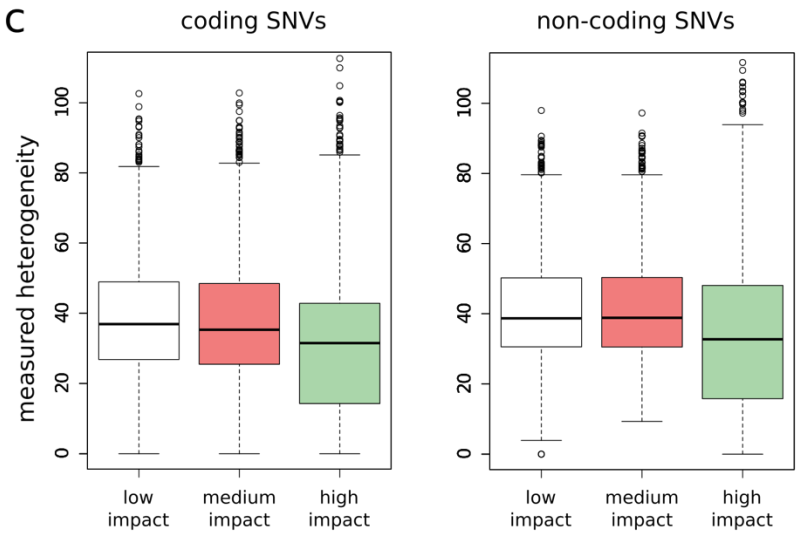

h
motif break
Kidney
Liver
Lymphatic
ColonRectum
Uterus
Myeloid
Thyroid
Biliary
Ovary
Lung
Pancreas
Esophagus
Stomach
CentralNervous
Prostate
Head
Bone
Skin
Cervix
Bladder
Breast

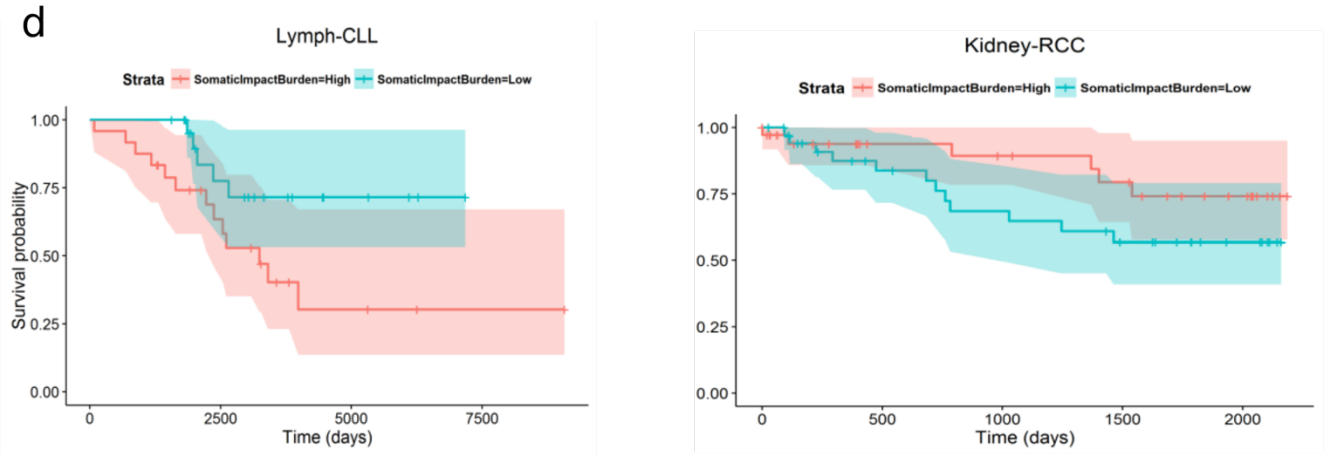

a
pan-cancer
probability density
functional impact score
b
essential genes
non-essential genes
coding
non-synonymous
SNV fraction / gene
promoter
low impact
medium impact
low impact
coding
loss of function
essential gene fraction
driver
non-driver
random
c
p-value
0.05
0.001
1e-6
impactful mutation fraction
total mutation count (log-scale)
impact vs. burden correlation
Skin-Melanoma
Eso-AdenoCa
Lung-AdenoCA
CNS-Medullo
strength of association (R²)


a
medium impact passenger SNVs
low impact passenger SNVs
signature 4
signature 4
mutations
mutations
kidney-RCC patient samples
kidney-RCC patient samples
b
all non-coding mutations (N=513,578)
mutations/MB
mutations
HDAC2 (N=675)
EWSR1 (N=514)
SP1 (N=571)
mutations/MB
mutations
mutations
mutations
c
colorect adenoCA
uterus adenoCA
impactful mutation fraction
impactful mutation fraction
p<6.5e-06
p<0.0031
MSS
MSI
MSS
MSI


a
passengers
Med Imp nonCod
Med imp cod
High Imp noncod
Loss of function
High Imp cod
Drivers
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
early vs. late mutations
DNArepair
Oncogenes
TSGenes
Apoptosis
Cell Cycle
Cancer Pathway
Imune Response
0
1
2
3
4
early vs. late mutations


b
fractino of tumor with mutation (VAF)
0.305
0.300
0.295
0.290
0.285
0.280
0.275
positive fitness predominates
a priori neutral fitness
evidence for net negative fitness
PCAWG driver SNVs
silent intergenomic
non-silent coding passenger
sensitive noncoding


c
coding SNVs
non-coding SNVs
measured heterogeneity
0
20
40
60
80
100
low impact
medium impact
high impact


d
Lymph-CLL
Strata
SomaticImpactBurden=High
SomaticImpactBurden=Low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
Time (days)
Kidney-RCC
Strata
SomaticImpactBurden=High
SomaticImpactBurden=Low
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
Time (days)