RESPONSE LETTER

-- Ref 1.1 – Description of survival analysis --

Reviewer Comment	1. Could the CLL results be explained by appeal to the two types of CLL known to have different signatures and prognosis? 2. Explain Survival Analysis in context of, i) how old the tumor was and ii) how well the person is when it was discovered. 3. Why does somatic passenger burden seem to effect only CLL and RCC, and in reverse directions?
Author Response	1. This comment may be referring to CLL's subcategorization into IGH mutant and unmutated classes. To address this comment we first considered a simple survival model predicting CLL patient survival as a sole function of whether IGH was mutated or unmutated. This analysis showed that mutated IGH status was associated with greatly prolonged survival (HR 0.31, p=0.0016) in keeping with trends reported in the literature. Nonetheless, when we include IGH mutation burden as an additional covariate into our full model, the association of somatic passenger impact with shortened survival remains significant (HR 1.40, p=0.035). 2. i) As tumors age, they accumulate mutations, and so tend to have more mutations of all types including putatively impactful and low-impact mutations. Patients with older tumors may have fewer years remaining before they succumb to their disease. We have addressed this confounding relationship in two ways. First, we include low-impact passenger mutation load as a covariate. Second, we define somatic impact burden in relation to corresponding randomized mutation sets, which ensures that an older tumor will not obtain a higher somatic passenger impact burden simply in virtue of its number of mutations. ii) Information regarding patient clinical status at time of sequencing is generally lacking in the PCAWG data-set. Even were this information available, however, it would not be clear that it would be appropriate to correct for patient clinical status – presumably impactful passenger variants might have already contributed to patient clinical status at the time of sequencing. 3. Nine cancer subtypes have sufficient patient events (>=20 patient deaths) for survival analysis with our model. Of these nine cancer subtypes, three have significant VAF-based evidence of neither LD nor DP and are thus unlikely to have a strong association between passenger burden and survival; two have significant VAF-based evidence of both LD and DP, which may tend to balance each other out; and four have "pure effects" (evidence of either LD or DP but not both) for which we might expect the greatest association of somatic

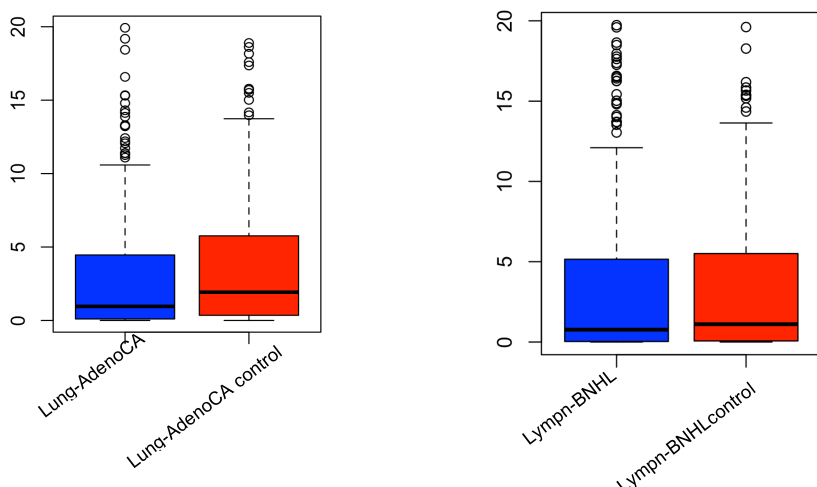|	passenger burden with survival. Indeed, of the four survival-eligible cancer subtypes with "pure effects, CLL has the highest predominance of LD over DP and RCC has the strongest evidence for DP over LD, which could help explain why only CLL has an LD-like survival curve and only RCC has a DP-like survival curve. More fundamentally, why cancer subtypes differ in terms of their relative enrichment of DP vs LD is an intriguing question raised by this analysis which may inspire future investigation

-- Ref 1.2 – Signature related analysis --

Reviewer Comment	We need to incorporate signature information for our analysis.
Author Response	Considering the absence of signature data available from PCAWG-7, we performed our signature analysis by generating a custom signature data for each cancer cohort. We will update our result once PCAWG-7 releases the finalized version of the signature data for PCAWG variants. However, we don't expect significant changes in our observation. Briefly, we use an in-house signature pipeline that is similar to Alexandrov et al., (NComms, 2015, http://www.nature.com/articles/ncomms9683). Namely, our pipeline solves a Frobenius norm minimization linear problem while promoting sparsity for each sample. Furthermore, our pipeline also takes the previously identified signatures in various cancer types as priors (http://cancer.sanger.ac.uk/cosmic/signatures). We exclude signatures that contributes to less than 2% of the explained somatic mutations which is an approach similar to Hong et al., (NComms, 2015, PMID: 25827447) but more prudent. The dominant signature we identified in in Kidney-RCC is Signature 5, which is in concordant with previous studies (Alexandrov et al., 2013). The bar for each patient are sorted by the number of mutations explained by the signatures we identified. Similarly, for the spectrum analysis related to TFmotifs, we identified most common motif breaking events in the Kidney-RCC cohort. Subsequently, mutation spectrum was generated by normalizing each mutation by the number of each trinucleotide triplet in the genome.

-- Ref 1.3 –Randomization set as baseline --

Reviewer Comment	Use simulation data instead of germline to perform comparison.
Author Response	We are utilizing multiple iterations of randomized neutral simulations data generated by sanger group to perform comparisons and enrichment analysis.

-- Ref 1.4 – Gene expression to validate burdening --

Reviewer Comment	Incorporate gene expression data to validate observations related to functional burdening of TF motifs
Author Response	In our work, we provide overall functional burdening observed in the TF binding landscape of various cancer cohorts. We observe that binding motifs of certain transcription factor are highly enriched in somatic variants leading to either gain or loss of motif. There was a suggestion to incorporate gene-expression data to further validate this observation. We performed a simple calculation as part of this validation effort. We identified target genes of TFs undergoing motif breaking event in samples belonging to a particular cancer cohort. Subsequently, we obtained expression value of these genes for appropriate samples where SNV induced motif breaking event was observed. In addition, we also extracted expression value of the same set of genes for samples, where TFs regulating these genes didn't undergo motif breaking events. This complementary set of gene expression values served as our control. As expected, expression distribution of case (target gene set in samples where TF binding motif was broken) was lower compared to control (target gene set for samples where TF motif was not unbroken) set. This observation was statically significant and was observed for multiple cancer cohorts. Below, we highlight this difference for the Lung-Adenocarcinoma and mature B-cell lymphoma.

-- Ref 1.5 – Additional description on impactful passengers --

Reviewer Comment	Extra validation or description suggesting presence of latent driver and deleterious passenger.
Author Response	We have newly performed a further validation of the existence of latent drivers and deleterious passengers using variant allele frequency. We hypothesized that passenger genes enriched in high-impact variants (relative to randomized mutations in the same genes) would be associated with extreme variant allele frequency – exceptionally low VAF genes representing deleterious passengers, and exceptionally high VAF genes representing latent drivers. Through this approach we detect the presence of deleterious passenger genes pan-cancer and in several cancer subtypes. We do not detect genes with pan-cancer VAF-based evidence of latent-driver variants, but we do see significant -and sometimes impressively significant- associations between high-impact enrichment and high-VAF variants in individual cancer subtypes (including $p=1.7E-8$ in Prost-AdenoCA).