RESPONSE LETTER

Reviewer 1:

-- Ref1.1 – Figure clarification --

Reviewer Comment	As such, Figure 1 has some weaknesses, which should be addressed. SMAP consists of two modules - creation of data context for variants and a variant prioritization pipeline. Please clarify the role of the weighted scoring scheme in the creation of the "Data Context", as well as the type of information associated with it. Indicate what the dashed horizontal line corresponds to.
Author Response	We thank the referee for the comment. We have modified figure legends to incorporate more details and added a new Figure 2 to illustrate the weighted scoring scheme. The dashed rectangle in Figure 1 shows how the data context is created. Variant prioritization pipeline will take user-input variants and score them using the weighted scoring scheme. Features used in the scoring scheme are shown in Figure 2.
Excerpt From Revised Manuscript	Please see Figure 1 & 2 and corresponding legends. Here is the excerpt from the legend of Figure 1. "SMAP consists of two modules: creation of data context and variant prioritization. We processed large-scale genomics (such as 1000 Genomes and ENCODE data) and cancer resources to create the small-scale informative data context, as shown within the dashed rectangle. The variant prioritization pipeline will take user-input cancer variants and then annotate and score them against the data context. All features are used to annotate variants (shown in Table S2), whereas a fraction of features highlighted with red asterisk are used to score variants (details in Figure 2 and Table S3) with the weighted scoring scheme (shown in the 'variant prioritization'; details described in the main text). 'Process' contains scripts to analyze data, which can be downloaded from our website."

-- Ref1.2 – Background section--

Reviewer Comment	The second and third paragraphs of Background should be reworded in places; in particular, the sentences beginning with "A number of tools" and "These include". The sentence beginning with "To explore the functional impact" seems to conflict the methods of analysis with their results and should be clarified.
Author Response	We thank the referee for the suggestion. We have modified the background section to make it clearer.
Excerpt From Revised Manuscript	Please refer to 'Background' section. "… Studies have shown that disease-associated single nucleotide polymorphisms (SNPs) identified by Genome-wide Association Studies (GWAS) are significantly enriched in ENCODE regions. A number of tools have been developed using these data to annotate potential regulatory variants or to suggest most likely causal variants in linkage disequilibrium with GWAS SNPs, such as Haploreg, RegulomeDB, ANNOVAR, GEMINI, FunciSNP and VEP …"

YAO FU 7/26/14 1:58 PM
Deleted: suggestions

YAO FU 7/26/14 2:01 PM
Deleted: We have modified Figure 1 and its legend to incorporate details of the figure and the scoring scheme. The dashed rectangle shows the data context, which combines large-scale genomics and cancer resources. Prioritization pipeline is a collection of scripts that use the data context to calculate motif-breaking and motif-gaining, recurrence among samples and to annotate and score (with the weighted scoring scheme) variants. The weighted scoring scheme is a prioritization step using the data context. Features used in the scoring scheme are highlighted with '*', which consist features directly from data context (e.g. in functional annotations) and also output features from prioritization pipeline, such as motif-gaining scores (detailed features are listed in Figure 2).

	"…Key features of our method include - 1) we integrated functional annotations to identify potential regulatory variants and predicted nucleotide-level loss-of and gain-of function events; 2) we examined whether variants occurred in noncoding regions that are less likely to tolerant mutations through analyzing both evolutionary and human population-level conservation; … "

YAO FU 7/27/14 3:44 PM
Deleted:

YAO FU 7/27/14 3:43 PM
Deleted: To explore the functional impact of regulatory variants, 1) we examined whether variants occurred in noncoding regions that are less likely to tolerant mutations through analyzing both evolutionary and human population-level conservation; 2) our method predicted nucleotide-level loss-of and gain-of function events;

Reviewer 2:

-- Ref2.1 – General comments--

Reviewer Comment	It is important to mention that this work builds up from a previous work by the authors (Khurana et al., Science 2013) in which the some of the ideas in the manuscript were already described. In this manuscript further details and novel elements are introduced to the prioritization system and the code and a web service of the framework is provided. The framework presented here (SMAP) represents a substantial evolution from the prioritization approach presented in the Science paper (FunSeq)
Author Response	We thank the referee for providing us these invaluable comments. In the revision, we provided more details and performed additional analysis as suggested.
Excerpt From Revised Manuscript	Please see the 'Background' section in main text. "…Through analyzing the variation patterns of inherited polymorphisms, we have published a prototype approach (FunSeq) to identify potential noncoding drivers. Here, we report a more elaborate and flexible framework - SMAP, built up from the previous work, to annotate and prioritize somatic alterations integrating various resources from genomic and cancer studies…"

-- Ref2.2 – Clarify the features --

Reviewer Comment	From one side it annotates with multiple features the variants, and next, some of the features are used to construct a prioritization score. However through the text and also in figure 1 it is not completely clear what features are used to obtain the prioritization score, and which are just used to annotate the variant in the final output to the user (variant reports). It would be helpful if authors make this clearer in the text and Figure 1.
Author Response	We thank the referee for pointing this out. We have added a new Figure 2 to clarify how features are used to score and

YAO FU 7/27/14 3:48 PM
Deleted: .

YAO FU 7/26/14 2:16 PM
Deleted: made changes to

YAO FU 7/26/14 2:16 PM
Deleted:

	annotate variants. We also restructured the main text to first describe features used in the weighted scoring scheme and then talk about additional features. Features used to score and annotate variants are further listed in Figure 3 and Table S2, respectively.
Excerpt From Revised Manuscript	Please refer to Figure 1 & 2 and Table S2 & S3.

YAO FU 7/26/14 2:16 PM
Deleted: in the manuscript, Figure 1 and 2

YAO FU 7/27/14 3:47 PM
Deleted: .

YAO FU 7/27/14 3:45 PM
Deleted: shown in Figure 2 and

YAO FU 7/27/14 3:45 PM
Deleted: Table S3

YAO FU 7/27/14 3:48 PM
Deleted: . We also added Table S2 to show all the features used to annotate variants.

YAO FU 7/26/14 2:20 PM
Deleted: In main text, 'Weighted scoring scheme' . ... [1]

-- Ref2.3 – Knowledge of genes --

Reviewer Comment	For instance, it is not clear to me how the "differentially gene expression analysis" module is used within the framework. Is this information used in any way to obtain the prioritization score of the variant? If not, how it is used? Similarly, how the prior knowledge of genes is used? I assume that it is not used in any way for the prioritization score, and it is only used to annotate the "gene info" column in the output? Is this the case?
Author Response	For 'differential gene expression analysis', we test for differentially expressed genes from RNA-Seq and then use those genes to annotate coding and non-coding variants associated with them. In the scoring scheme, we don't add additional scores for those variants, considering user-input samples are not always coupled with RNA-Seq data. But this information does help to further prioritize variants and we highlight those variants in the output. The prior knowledge of genes is used in the similar way as differentially expressed genes. We made this clearer by showing that these features are used as additional features to highlight variants.
Excerpt From Revised Manuscript	Please refer to Figure 1 & 2 and Table S2 & S3. In main text, we added section - '**Highlighting variants using prior knowledge of genes and user annotations**' Interpretation of the functional impact of noncoding variant can be greatly enhanced if the function of its target protein-coding gene is known. Many cancer genes are known to play a crucial role in cell proliferation and DNA repair. We incorporated prior knowledge of genes, such as known cancer-driver genes,...'

YAO FU 7/26/14 2:23 PM
Deleted: didn't

YAO FU 7/27/14 3:52 PM
Deleted: DNA-seq

YAO FU 7/26/14 2:22 PM
Deleted: ose

YAO FU 7/25/14 1:20 PM
Deleted: file 'Candidates.Summary'

YAO FU 7/26/14 2:25 PM
Formatted: Font:10 pt

YAO FU 7/26/14 2:26 PM
Formatted: Font:12 pt, Not Bold

YAO FU 7/26/14 2:25 PM
Formatted: Font:10 pt

YAO FU 7/26/14 2:25 PM
Formatted: Font:10 pt

YAO FU 7/26/14 2:24 PM
Deleted: Weighted scoring scheme' . ... [2]

YAO FU 7/26/14 2:25 PM
Formatted: Font:10 pt

-- Ref2.4 – Mention regulatory mutations --

Reviewer Comment	In the title and through the text authors talk about noncoding somatic variants, however most of the features they compute are specific for regulatory noncoding variants, and this is what this framework is able to prioritize. There are other types of noncoding variants different to regulatory variants, such as those in non-coding RNA genes, which would not be well prioritized by SMAP as they will have missing values in several features. If authors agree with this point they should clarify in the title, abstract and through the manuscript that SMAP is a framework to prioritize

	regulatory noncoding somatic variants.
Author Response	We agree with the reviewer's comment. The majority of our features focus on regulatory mutations. We have modified the text as suggested.
Excerpt From Revised Manuscript	Please see the title, abstract and main text. "**A flexible framework to annotate and prioritize regulatory somatic variants from cancer whole-genome sequencing**" "We have developed a method integrating various genomic and cancer resources to prioritize cancer somatic variants, especially regulatory noncoding mutations…."

-- Ref2.5 – Missing features --

Reviewer Comment	Related to that I also have a question regarding missing features. How they are treated? I assume that if a variant do not have any information on one of the features (eg. Motif-breaking or gaining score) it does not sum up anything in the final score, does it? This could be clearly stated in Formula 3.
Author Response	The referee is right. If a variant does not have a particular feature, there is nothing added-up in the final score. We clarify this in Formula 3.
Excerpt From Revised Manuscript	Please see the Methods, 'Weighted scoring scheme'. "…If a particular feature is not observed, it is not used in the scoring."

-- Ref2.6 – Distribution of distal linkages--

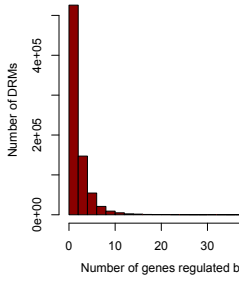
Reviewer Comment	Authors give the number of distal regulatory elements (~769K) and the number of associated genes (~17K), but it would help to also know the total number of interactions and the distribution of interactions per gene and per distal regulatory element so to have an idea of how many interactions are observed per gene and regulatory element.
Author Response	We thank the referee for this suggestion. We incorporated the numbers in the Methods section and showed the distributions in Figure S1.
Excerpt From Revised Manuscript	In Methods, 'Associating regulatory elements to likely target genes' "… we further expanded the method to all ENCODE non-coding regulatory elements and identified ~2,225K significant associations between ~769K regulatory elements and ~17K genes (see below). The distributions of regulatory element-gene associations are shown in Figure S1. The median number of associations is 22 and 2 for per gene and per regulatory element, respectively... "

-- Ref2.7 – Region within 10kb--

Reviewer Comment	DRMs are defined as those regulatory regions at least 1kb from the closest gene and associations with tssEUs

YAO FU 7/26/14 2:26 PM
Deleted: d

YAO FU 7/25/14 1:25 PM



Deleted:

YAO FU 7/25/14 1:25 PM
Deleted: . ... [3]

	were computed from all tssEUs beyond 10kb but within 1Mb from it. What happens with the regions between 1kb and 10kb? They are not tested?
Author Response	We thank the referee for pointing this out. For regulatory regions within 10kb of genes, we also test for associations between them and adjacent genes. ….. [will come soon…. From Shaoke]
Excerpt From Revised Manuscript	????????? ??

-- Ref2.8 – TSS cut-off--

Reviewer Comment	Page 11 "For each tssEU, we defined its expression level as the number of RNA-seq reads aligned to the [TSS- 50, TSS+50] window." Why from TSS-50?
Author Response	The main reason is to allow potential small errors in the annotation of the TSS. 50bp is a small window size to avoid running into another TSS.
Excerpt From Revised Manuscript	

-- Ref2.9 – Add number of variants --

Reviewer Comment	Authors should provide the number of variants in each group (in Methods and/or Figure 2).
Author Response	We thank the referee for the suggestions. We provided the numbers in both the Methods and Figure 2.
Excerpt From Revised Manuscript	Please refer to the Methods and Figure 2. 'Application to regulatory pathogenic and somatic cancer variants' "…We obtained noncoding somatic variants form COSMIC (version 68). Recurrent variants (10,041) are defined as identified in whole-genome sequencing and observed in at least 2 samples. All other variants (1,311,389) are non-recurrent ones. After excluding variants in coding regions (GENCODE 16) and mitochondrion, there are 956 variants occurred in more than 2 samples, 8,932 variants in 2 samples and 1,305,699 non-recurrent variants…"

-- Ref2.10 – Clarify COSMIC and 570 samples--

Reviewer Comment	COSMIC includes somatic mutations observed in tumors and many of the most recent data is from whole genome/exome projects. There is the possibility that the datasets of COSMIC and those from the recurrence database of 570 samples of 10 tumor types are overlapping quite a lot. Could authors clarify this?
Author Response	We have checked the COSMIC database and our 570 samples. 468 out of 570 (82.1%) are not in COSMIC. Majority of the cancer samples (from Alexandrov's paper) are newly sequenced and are

	not submitted to COSMIC.
Excerpt From Revised Manuscript	

-- Ref2.11 – Comparisons--

Reviewer Comment	1. For the sake of comparison with existing methods, the ability to separate between these groups of variants by CADD and GWAVA should also be shown, including boxplots and p-values. 2. The authors could also compute AUC and/or other performance metrics for how well SMAP, CADD and GWAVA are able to separate between the two extreme groups (non-recurrent variants Vs. >2 sample variants).
Author Response	We thank the referee for the suggestions. We made the comparisons with GWAVA and CADD using boxplots and AUC calculations. The results are shown in Figure S5. As mentioned by the referee, recurrence is not a good criterion to define functional and non-functional sites. As expected, none of the three methods could separate recurrence from non-recurrence well, with AUCs around 0.5. Generally speaking, our method performs better than others.
Excerpt From Revised Manuscript	Please see the Additional file 1: Figure S5. "… Results from CADD and GWAVA are shown in Additional file 1: Figure S5. "

-- Ref2.12 – Clarify recurrent elements--

Reviewer Comment	How are defined recurrent regulatory elements? For instance, it is required that more than one sample has mutations in the same TFBS motif or in the same promoter?
Author Response	Recurrent regulatory elements are regulatory regions mutated in more than one sample. For example, the same TFBS motif with mutations from two or more samples. We added the description in the Methods section.
Excerpt From Revised Manuscript	Please refer to Methods, 'Noncoding somatic variants in recurrent regulatory elements'. "Regulatory regions mutated in more than one sample are defined as recurrent regulatory elements, such as the same TF binding motif or the same noncoding RNA…."

-- Ref2.13 – *TERT* promoter mutation in other samples--

Reviewer Comment	1. As there are 7 samples with the TERT promoter mutation, why the authors do not test the ranking of the mutation in all 7 instead of only one? It would be more

YAO FU 7/21/14 10:31 AM
Deleted: 9

YAO FU 7/21/14 10:31 AM
Deleted: .

YAO FU 7/25/14 1:21 PM
Deleted: on both COSMIC (do we want to show results from COSMIC ??? our results are slightly worse than GWAVA…. Reasons see below…) and Breast cancer data (

YAO FU 7/25/14 1:21 PM
Deleted:)

YAO FU 7/25/14 1:24 PM
Deleted: As noted in COSMIC database, the same sample can have multiple ids if the sample has been entered into the database multiple times from different papers. Also COSMIC collects various studies, probably with various data qualities. Even though it claims that COSMIC contains only somatic mutations, we found 6.6% of recurrent and 2.7% of non-recurrent mutations occurred in 1000 Genomes. For those overlapping with 1000 Genomes, the allele frequency of recurrent ones (mean: 0.19) is significantly higher than non-recurrent ones (mean: 0.11) (p-value < 2.2 e-16). On the contrary, the Breast cancer mutations are carefully filtered against germline and natural variants. We considered that the results from Breast cancer were more reliable.

YAO FU 7/25/14 1:23 PM
Formatted: Indent: Left: 0"

YAO FU 7/25/14 1:23 PM
Deleted: . ... [4]

YAO FU 7/21/14 10:32 AM
Deleted: 0

YAO FU 7/21/14 10:32 AM
Deleted: 1

	information to see the ranking of this mutation in each of the 7 samples and also the ranking provided by GWAVA and CADD. 2. Does the prioritization in the case study use anyhow the recurrence information?
Author Response	We thank the referee for the comments. 1) We provided results for all 7 samples in Additional file 1: Table S4, together with results from GWAVA and CADD. 2) In the case study (sample MB59), we used the recurrence information. We also provided the ranking without recurrence in Table S4. Without recurrence, our method still performs better than GWAVA and CADD.
Excerpt From Revised Manuscript	Please refer to Additional file 1: Table S4. In main text: "… Results of additional 6 samples are shown in Additional file 1: Table S4."

YAO FU 7/25/14 1:25 PM
Deleted: Sample ... [5]

YAO FU 7/21/14 10:32 AM
Deleted: 2

-- Ref2.14 – Figure legends--

Reviewer Comment	Figure legends are short and not informative enough, they could be much longer to contain all the necessary information to understand the figure without having to go to methods section. For instance, Figure 2B Y and X-axis should be defined in the legend.
Author Response	We agree with the referee. We made modifications to the legends.
Excerpt From Revised Manuscript	Please refer to figure legends. Here is an excerpt for Figure 3. "Figure 3 - Weighted scoring scheme. A) Features used in the weighted scoring scheme. Features can be classified into discrete and continuous. Discrete features are binary, such as in ultra-conserved elements or not. For continuous features, taking 'motif-breaking score' as an example, the values would be the changes in PWMs. * only applicable when user input multiple genomes; B) We weighted each feature based on the mutation patterns observed in natural polymorphisms. Features that are frequently observed are less likely to contribute to the deleteriousness of variants and are weighted less (entropy based method, details described in Materials and Methods). For continuous feature, such as motif-breaking scores, we calculated weights for each observed value. The x-axis is the observed motif-breaking scores and y-axis is the corresponding weights. The black line show the values observed in natural polymorphisms. We then fitted a smooth curve (the red dashed line) to obtain continuous weights for all possible motif-breaking scores."

YAO FU 7/26/14 2:28 PM
Deleted: 2

YAO FU 7/26/14 2:28 PM
Deleted: 2