Author's response to reviews

Title: On the identification of potential regulatory variants within genome wide association candidate SNP sets

Authors:

Chih-yu Chen (juliec@cmmt.ubc.ca)
I-Shou Chang (ischang@nhri.org.tw)
Chao A Hsiung (hsiung@nhri.org.tw)
Wyeth W Wasserman (wyeth@cmmt.ubc.ca)

Version: 3
Date: 19 April 2014

Author's response to reviews: see over
18 April 2014

BMC Medical Genomics

Re. Revision of rSNP informatics manuscript

Dear Dr. Sleiman and colleagues,

We are pleased to submit the revision of the manuscript entitled “On the identification of potential regulatory variants within genome wide association candidate SNP sets” for consideration as a research article in BMC Medical Genomics. We believe the methodology and the results will be of significant interest to our research community.

In this report, we introduce a novel approach to the study of non-coding variants within GWAS identified loci, with an applied focus on cancer studies. We demonstrate that regulatory features are enriched in cancer susceptibility GWAS loci SNPs relative to the genome. The general informatics approach is applied to lung cancer susceptibility regions, revealing a striking candidate variant that impacts a Myc transcription factor binding site.

We appreciate the referee feedback provided for the previous version of our manuscript, and have adjusted the methods and the manuscript in response. The outcome is a significantly improved manuscript, which will help others working on the interpretation of regulatory alterations using GWAS data. A statement detailing the actions we have taken is provided for each of the referees’ comments.

I look forward to your considered response.

Yours sincerely,

Wyeth W. Wasserman, Ph.D.
Referee #2

Major compulsory revisions

#1

Table S3 has to be completed to be able to see the PWM name that have been used to score the SNP region (some TFs have multiple PWM in JASPAR) and the scores of each allele (e.g. 78.1 & 90.5 for rs12087869). Also it would be interesting to extend table S3 to less significant results or to provide a raw file containing scores & score differences for each SNP / PWM. It will help to understand why when looking at Figure 6, Myc PWM score difference is significant whereas other E-boxes that are present in JASPAR (BHLHE40 and Max) didn’t give a score difference significant enough to be present in table S3.

Done as requested. Please refer to Additional file 6 for the updated list and Additional file 7 for the extensive list of PWM scores for Lung.Meta set. The PWM IDs from JASPAR are listed in Additional file 7. Note that we adjusted our TF affinity scoring approach to better account for the distribution of scores, as requested below. With the improved procedure, Myc remained a significant motif, and the MAX and TLX1::NFIC score at rs12087869 are also found to be significantly altered.

#2

Points that should be explained / discussed:

- Figure 4: It is unclear to me if the A549 cell line used to compute the differential regulatory potential (log2((RPIcancer+1)/(RPInormal+1))) contains the SNPs displayed in figure 4. If not, it would be nice to better explain what we can learn from figure 4.

We did not restrict the SNPs to be present in A549, as we are limited by availability of high throughput sequencing data sets. A549 does not have the risk allele, but does show the regions to be active and accessible in support of a role for the segment in transcriptional regulation in the lung cancer cells.

- Figure 6: Similar point as in figure 4, if I understood well, neither the NHLF nor the A549 cell line contain the SNP increasing DNA sequence affinity to myc.
Then the apparition in A549 cell line of a DNAseI mark and of histones methyla
tion/acetylation which are not present in NHLF cell line is not linked to rs12087869-C allele. So what can be concluded?

We do not expect a binary (on/off) characteristic for the TFBS. We are proposing that this region is an open/accessile regulatory region and that Myc/MAX/TLX1::NFIC will have strong binding affinity in those individuals with the risk allele. Altered binding affinity has been shown to have substantial impact on the expression mediated by TFBS. We have clarified this concept in the manuscript.

#3

In the “Methods, Differential TF binding affinity analysis using PWMs” section, can the author confirm that the score differences are following a normal distribution to justify the use of a Z-score?

We appreciate the referee’s observation about the distribution of scores. While some of the distributions approximate normal, we conclude that it would be more appropriate to use an alternative measure. We have thusly replaced the Z-scores with empirical p-values. We report the two-tailed empirical p-values for differences in TF binding affinity by comparing to that of 10000 matched random SNPs. We obtain the probability of getting random score differences higher than the true score difference for assessing an increase in TF affinity, and the probability that is lower than true score difference for a decrease in TF affinity. We multiple the minimum of the two probabilities by two to be the two-tailed p-value of each PWM for each SNP.

Minor essential revisions

#1

In the “Results, Compilation of GWAS SNP data” section, can you remove “referred to in the same order in Figure2”, or put the same order as in Figure 2?

Done

#2

In “Results, Prioritizing functional SNPs using regulatory potential and TF binding affinity”, shouldn’t it be Lung.meta instead of Landi.meta?

Done. The referee is correct that this was an error.

Discretionary revisions
The authors of the paper are using only the JASPAR motif database whereas multiple others exist (e.g. Transfac database, elemento motifs etc...). It would be interesting to discuss limitation and advantage of such a choice?

We have provided our perspective about the selection of TF binding model collections to the discussion section of the manuscript. Within this initial study, we utilize the JASPAR database as we understand it well as part of the development team. However, future studies could appropriately evaluate the alternative sources, such as HoCoMoCo, which aggregates profiles from multiple resources.

In Figure 3, what can explain the enrichment for SNPs in regulatory sequences also for non-cancer cell lines?

This relates to the lack of clarity around the enriched presence of the SNPs in open/accessible regulatory regions. Such regions may well be open/accessible in both normal and cancer cells, but in the presence of the TFBS altering change the regulation conferred would be altered. As stated above, we have clarified this perspective in the revised manuscript.

Referee #3

Major Compulsory Revisions:

(1) The “Results” part starts with a definition of cancer-associated SNPs collection and a calculation of LD SNPs. The exact number of SNPs for each set should be mentioned in the main text.

We have mentioned the size of each LD95 SNPs in the main text.

(2) The scale for rpkm values of demonstrated tracks should be added (Figure 6A). It would be nice to provide not only tracks for this epigenetic data, but called peaks as well. The whole region containing the SNPs of interest should be zoomed in.

We have revised figure 6 according to the request, but we elected not to include the zoomed-in figure, as in our opinion it does not enhance the presentation of information and the manuscript is already burdened with an excess of figures.
Minor Essential Revisions

(1) In the current version the signatures under Figure 3E and 3G look cluttered. We have replaced Figure 3 with a heatmap representation.

(2) Figure 6A shows not H3K4m1 and H3K4m3 tracks, but H3K4me1 and H3K4me3 respectively. We have corrected this typing error.

Referee #1

Major Compulsory Revisions:

1. SNPs associated with several cancer types through GWAS were obtained from different sources (e.g. original discovery paper, NIH GWAS catalogue, etc.). As the authors noted in the Methods section, these sources each applied their specific methodology and significance thresholds for processing the data sets. The difference in strength of association has a substantial effect on the enrichment analysis using epigenomic data sets (e.g. as recently shown by Maurano et al. (2012), Science 337, 1190-5). The selection of GWAS SNP data sets appears inconsistent and selective. For example, why was only one lung cancer GWAS considered, but several for prostate and colorectal cancer GWAS? Therefore, I advise the authors (1) to set a consistent threshold of significance for association, e.g. P<5x10-8 (‘genome-wide significance’), (2) to have a consistent approach to obtaining SNP data, e.g. entirely via the NIH GWAS catalogue and setting discrete filtering criteria, (3) to have only one cumulative SNP data set per cancer type. The latter may also improve the current nomenclature, such as ‘Lung.Meta’ (also referred to as ‘Landi.Meta’), ‘Lung.Adeno’ and ‘Breast.20852631’.

We recognize the variability in thresholds of significance as well as the statistical approaches among the GWAS studies that we included. Taking the reviewer’s advice, we have included and
analyzed the set of lung and breast cancer studies provided via UCSC/NHGRI and use a threshold consistent with the previously included prostate and colorectal sets (p-values < 1.0 × 10⁻⁵). To avoid distraction, we have removed the individual GWAS studies (Breast.* and Lung.Adeno), but have retained the meta-study from Landi et al. (Lung.meta) for comparison.

2. For the enrichment analysis of SNPs in regulatory sequences, the authors obtained random background SNP sets by accounting only for genomic functional category. In addition, the authors should account for minor allele frequency, distance to the TSS of the nearest gene, and GC content. This procedure accounts for potential positional bias on the SNP array, and has been described in the literature, e.g. Maurano et al. (2012), Science 337, 1190-5; ENCODE Project Consortium (2012), Nature 489, 57-74; or online: http://www.1000genomes.org/forge-analysis. This background set should also be used for the differential TF binding affinity analysis. A consistent number of permutations should be used, e.g. n=1000.

We appreciate the concerns raised by the reviewer, and have reanalyzed the enrichment analysis of SNPs in regulatory sequences in a manner that accounts for the indicated properties. These revisions, more fully discussed below, have been incorporated into the manuscript.

The minor allele frequency, distance to the TSS of the nearest gene, and GC content +/- 500bps were binned into 20 percentile bins, and the attributes were used to match to generate a matching background set of random SNPs from the Illumina 660K SNP array. The functional category constraint was removed, as it is correlated to distance to the TSS. The enrichment test results were comparable.

We also reanalyzed the TF binding affinity differences using randomly drawn SNPs with matching allele differences, distance to the TSS of the nearest gene and GC content +/-500bps.

We maintained the numbers of permutations for a better precision in computing the p-values for TF binding affinity differences. As motivated by comments from another reviewer, we note that the distributions of PWM score differences for TFs are not consistently normal. Therefore, we replaced the Z-score (based on normal distribution) and reported the two-tailed empirical p-values for differences in TF binding affinity.

3. The authors claimed that the SNP rs12087869 ‘lies within segments observed as open chromatin for multiple cancer cell lines such as LNCaP, MCF-7, PANC-1, and the PrEC prostate epithelial cells, but not in normal cell lines such
as NHLF and HMEC.’ This claim was not supported by functional data. Further, Chen and colleagues stated that ‘the A549 cell line does not display elevated ROR1 expression relative to NHLF cells’. Again, no data was shown in support. Generally, the RNA-seq data, as outlined in the Introduction and Methods section, was not discussed in the Results section.

We have provided the data requested in Additional file 8.

The RNA-seq data outlined in introduction and methods was used to identify the nearest differentially expressed genes of SNPs in the supplementary table (Additional file 6). We have attempted to make this more clear in the revised manuscript.

Minor Essential Revisions:

1. The authors selected SNPs in LD with the cancer GWAS index SNPs using the SNAP web tool. The authors should point out that the tool uses SNPs from the Pilot 1 of the 1000 Genomes project. The authors should clarify whether the list of proxy SNPs were obtained from the 1kG Project Pilot 1 and the HapMap Project (different releases) and then combined (as indicated in the Methods section), or obtained from either source (as indicated in the Results section). A consistent approach would be best. Further, I would recommend retrieving proxy SNP data from the latest release of the 1kG Project (e.g. via the ‘FunctiSNP’ tool).

We have clarified the text to better describe our consistent practice for compilation of LD95 SNPs. We recognize that there are more than one tool to retrieve proxy SNP data, and will explore options in the future.

2. The authors claimed that ‘similar significance in enrichments tests of regulatory sequences’ was found using an r2 threshold of 0.80 compared to a threshold of 0.95 (Methods section, 2nd paragraph). However, no evidence of this observation was given. We have provided an enrichment heatmap for R² threshold of 0.8 in Additional file 4 to address this concern.
3. The description of the selected cell lines could be improved. For example, it is important to note that HMEC cells are Human Mammary Epithelial Cells and thus the ‘healthy equivalent’ to MCF-7 breast cancer cells. The same is true for Human Prostate Epithelial Cells (PrEC) and LNCaP prostate cancer cells, etc. Done.

4. The cell line Panc-1 was featured in Table S1, but not discussed in the Results section. Further, the authors should discuss why H1 embryonic stem cells were included in this study, and/or discuss results obtained with these cells. Panc-1 was included by mistake; it has been removed. The datasets from H1 embryonic stem cell were included, as GWAS SNPs associated with various cancers have been reported to be enriched in ES cell enhancers by Teng et al. 2011 NAR. We have clarified that in the text.

5. I suggest replacing Figure 3 with Figure S2, as the latter is much more informative. The current Figure 3 shows a subset of only significant enrichment, whereas non-significant enrichment of GWAS SNPs in potential regulatory regions in ‘non-relevant’ cell types is equally important. Further, the authors should discuss the enrichment of GWAS SNPs in normal cells. For example, the closest genes to SNPs within regulatory regions may be functionally characterised, and compared to the genes proximal to SNPs within regulatory regions in cancer cells. Are different molecular processes highlighted? Are common cancer pathways enriched across different cancer types? We agree with the reviewer and have replaced Figure 3 with the original figure S2. As SNPs in high linkage disequilibrium are frequently close to the same gene, given the small numbers of SNPs in each LD95 set we could not find any conclusive evidence of (differential) functional enrichment in regulatory sequences of cancer and/or normal cells.

6. For the assessment of the potential impact of SNPs on TF binding affinity, the authors used PWM scores. It is not clear to me which SNPs were used for this analysis, e.g. all SNPs of the LD95 set that are located within putative regulatory regions identified, and by which chromatin marks? All SNPs of the LD95 set were used for the analysis, and only the ones with strong TF affinity
differences and with >85 PWM score cutoff in either major or minor alleles were presented in Figure 4. We believe the restriction to the top scoring subset improves the presentation of results. We have clarified the legend and text to better highlight this restriction.

7. In the section ‘Prioritizing functional SNPs using regulatory potential and TAF ChIP-seq data’, the authors should indicate why the cell lines LNCaP, HMEC, PrEC were not studied (alongside the cell line NHLF). I assume ENCODE ChIP-seq data was not available for these cell types?

As we found the Lung.meta set to be more significantly enriched in regulatory sequences and with more HTS datasets available, we focus the rest of the analysis on this set. We now provide the additional results in (Additional file 8), for the readers interested in other cell lines.

8. The y-axis in Figure 5 is unclear, and attention should be paid concerning the nomenclature ‘TAF’ vs ‘TF’.

9. In the last Results paragraph, the authors highlighted a particular SNP, rs12087869, and discussed it in greater detail. Why was this SNP chosen, and not any of the other SNPs that had a greater ‘relative regulatory potential’?

We highlighted rs12087869 as it is among the highest rankings across three criteria: regulatory potential differences, predicted TF binding affinity differences and overlap with TAF ChIP-seq binding sites (in 8 datasets) in A549 cells. We have modified the text to clarify this motivation.

10. The authors state that there was a ‘significant difference’ between the PWM scores of the alleles of rs12087869 and c-Myc binding (i.e. 78.1/90.5). It would be useful if the authors could provide the standard deviation or results of an appropriate statistical test. In addition, I suggest showing all eight TFs as PWMs and their PWM scores in Figure 6B, as well as their PWM score statistics.

Not all TAF ChIP-seq peaks we identified have a corresponding PWM. We included the raw output of PWM scores and the empirical p-values for each TF at each SNP in Lung.Meta analysis in Additional file 7 (as requested by Referee #2).

11. In the Discussion, the authors should touch on the possibility that other target genes than the suggested ROC1 may be regulated by the putative enhancer harbouring rs12087869. As no evidence is shown, e.g. through allele-specific expression or chromatin conformation capture, the selection on ROC1 as target gene merely on genomic proximity is speculative.
We have modified the discussion of potential target genes. It has been reported that enhancers can target multiple TSSs, and 40% of inferred TSS-associated enhancers (from FANTOM5 CAGE data) target the nearest TSS at least (Anderson et al. 2014 Nature). We now include additional topological domain data (i.e. HiC chromatin interaction data from Dixon et al. 2012 Nature) to further support the potential gene targets for the variants. These TAD boundaries have been shown to be consistent between cell types and developmental stages. We added this analysis to the result (Figure 7) to infer potential gene targets, and showed that ROR1 to be one of the genes co-localizing with the putative enhancer in a topological domain.

12. Over the past few years, several SNPs associated with various cancer types via GWAS have been prioritised and their functional effects been validated in vitro and in vivo. A recent review (Edwards et al. 2013, Am. J. Hum. Genet. 93, 779-97) summarised such potential causal variants. I would be interested to see whether these variants have also an increased regulatory potential as determined by the authors’ method. This analysis would serve as a validation of the candidate SNPs prioritised by the approach.

We greatly appreciate the helpful suggestion, and thus added Additional file 9 to summarize the comparison of our results to the potential causal variants listed in the review. We note that some SNPs that were identified in studies reviewed were not included in the input to our analysis. All SNPs included in our study met one or two out of our three criteria in regulatory potential, TF binding affinity or TAF ChIP-seq binding.

13. Lastly, to improve the writing style of the article, I suggest the authors consistently keep the Methods and Results section in past tense. We have converted the Methods section to past tense, while some sentences in the results section remain in present tense as appropriate to their content.