Reviewer’s report

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

Version: 2 Date: 24 February 2014

Reviewer: Dirk Paul

Reviewer’s report:

In this article, Chen and colleagues investigated non-protein coding sequence variation associated with various cancer types. The authors introduced a new methodological approach to prioritising candidate functional SNPs by using reference epigenomic data sets in relevant cell lines in combination with allelic transcription factor binding affinity scores from position weight matrices. The authors demonstrated the potential of their method by studying a SNP, rs12087869, associated with lung cancer in greater detail. The presented framework may be applied to other disease traits.

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’.

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.

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.

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).

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.

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.

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.

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?

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?

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?

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’?

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.

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.

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.

13. Lastly, to improve the writing style of the article, I suggest the authors consistently keep the Methods and Results section in past tense.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.