Author's response to reviews

Title: Dark Matter RNA Illuminates the Puzzle of Genome-Wide Association Studies

Authors:

Georges St. Laurent (georgest98@yahoo.com)
Yuri Vyatkin (yuri@nprog.ru)
Philipp Kapranov (philippk08@gmail.com)

Version: 3 Date: 20 May 2014

Author's response to reviews: see over
Dear Sabina,

Please find the revised manuscript and the detailed response to the reviewers’ comments. We hope you find them satisfactory and we are eagerly awaiting your decision.

Sincerely,

Phil

Reviewer: Leslie Biesecker

Comment: If the authors and editors wish for this review to be accessible and useful to the wider readership, it will need extensive editing so that the numerous specialty terms are defined, assumptions explained, etc. …

Response: We have included a Table with Glossary of 25 technical terms and their explanations. We feel it would represent the best compromise between maintaining the flow and readability of the text for a more savvy reader and helping the less knowledgeable reader with terms and assumptions exactly as requested by the reviewer.

Comment: For example, I think the non-specialist reader would need to see some discussion of the general concept of gene and so-called “junk DNA” and how our understanding of that has changed over the years.

Response: We included the concept of “junk DNA” in the beginning of the review – bottom of page 2. This would indeed help a non-specialist and facilitate the discussion of the change in the concept of the junk DNA later in the review.

Comment: The authors need to be more careful with their slang – this is an even more serious problem than formally correct arcane terminology.

Response: We went through the manuscript and edited it accordingly.

Comment: “These large and well validated datasets combined now provide a strong basis for a strategic discussion of non-coding RNA (ncRNA) and GWAS.” I have no idea what a strategic discussion is, who should be having it, and what the outcome or objective of such a discussion would be.
Response: We have edited the sentence accordingly. In addition, we have edited the entire manuscript to improve the flow and the meaning.

Comment: The conclusion paragraph needs to be rewritten – again, the authors need to think about the audience they are writing for and what they need to take away from this.

Response: We have modified the conclusion paragraph and added information on clinical utility of ncRNAs that should also address the reviewer’s point on clinical vs research sequencing.

Comment: The authors fail to properly distinguish research sequencing from clinical sequencing. Since the latter is now going on in a big way, this will confuse the readership. Research sequencing of non-coding elements is essential for researchers to understand the relationship of variation to disease, but it is not useful at this stage for clinical sequence interpretation.

Response: The information requested has been added to the conclusion paragraph.

Reviewer: Eduardo Reis

Comment: In this well-written and timely short review, the authors elaborate the argument that disease-associate sequence variants identified by GWAS studies that lie in unannotated genomic regions may exert their effects by targeting regulatory long noncoding RNAs.

Response: We wish to thank the reviewer for kind words.

Comment: However, I would disagree that there is consistent evidence to support the claim that complex diseases are better explained by mutations in ncRNAs than in protein-coding mRNAs (pg. 7). The ability of GWAS studies to reveal the genetic basis of complex diseases (e.g. cancer, neurological) has been reduced by limited sample sizes, which is likely to affect the identification of both coding and noncoding variants.

Response: The reviewer brings up an important point – sample size is indeed critical for identification of any variant, coding and non-coding associated with disease. However, all GWAS variants reported here were downloaded from UCSC GWAS database (please see Supplementary Text) that only contains SNP-trait associations that have p-values < 1.0 x 10^{-5}. Therefore, all GWAS hits reported here come from studies with sufficient power to detect both coding and non-coding variants. Overall, 93% of 11,774 such GWAS SNPs (Supplementary Table 3) from a total of 1,462 different publications (Supplementary Table 1) correspond to non-coding regions of the genome – outside of coding exons of known genes (CDSs+UTRs).
**Comment:** To allow reproducibility, the analysis shown in Figure 1 and discussed in the text should be accompanied by more detailed information regarding the primary data used (what was the GWAS catalog?) and the analytic method (What is the impact of counting the same variant multiple times if it appears in more than one publication? Could it introduce biases?).

**Response:** We have now included three Supplemental Tables with the original classification of each GWAS variant into a disease category and the final numbers that went into making of the Figure 1. We have also included Supplementary text with detailed explanation of the analysis.

In terms of counting the association between a variant and a disease category more than once if it appeared in different publications, the reasoning for it was as follows. Reproducibility is of paramount importance in GWAS, for example please see discussion on this topic here: [http://www.nature.com/nature/journal/v447/n7145/full/447655a.html](http://www.nature.com/nature/journal/v447/n7145/full/447655a.html). Therefore, we thought that we had to give some statistical weight to the associations between a variant and a disease category independently found more than once. After all, each GWAS study surveys entire genome and each of the 100’s of thousands or millions of SNPs in a platform used in a study has an equal chance to be found. The fact that a variant was found again adds a significant weight to that variant’s importance in a disease and therefore should be factored in the analysis.

**Comment:** Pg. 2: It would be nice to mention what kind of errors are potentially involved in GWAS studies and/or provide appropriate references to this issue.

**Response:** We have now included the information requested on Pg.3: “Three possible explanations exist for the large preponderance of non-coding GWAS hits. They could arise from methodological errors such as imprecise measurements of phenotypes [3], differences in population structures [4] or DNA quality issues [5, 6] between cases and controls.”