Author's response to reviews

Title: Diverse Convergent Evidence in the Genetic Analysis of Complex Disease: Coordinating Omic, Informatic, and Laboratory based evidence to prioritize findings for further study

Authors:

Timothy H Ciesielski (Timothy.H.Ciesielski@Dartmouth.edu)
Sarah A Pendergrass (sap29@psu.edu)
Marquitta J White (Marquitta.J.White@Dartmouth.edu)
Nuri Kodaman (Nuri.Kodaman@Dartmouth.edu)
Rafal Sobota (Rafal.Sobota@Dartmouth.edu)
Minjun Huang (Minjun.Huang.GR@Dartmouth.edu)
Jacquelaine Bartlett (Jacquelaine.Bartlett@Dartmouth.edu)
Jing Li ( Jing.Li.GR@Dartmouth.edu)
Qinxin Pan (Qinxin.Pan.GR@Dartmouth.edu)
Jiang Gui (Jiang.Gui@Dartmouth.edu)
Scott B Selleck (sbs24@psu.edu)
Christopher I Amos (Christopher.I.Amos@Dartmouth.edu)
Marylyn D Ritchie (mdr23@psu.edu)
Jason H Moore (Jason.H.Moore@Dartmouth.edu)
Scott M Williams (Scott.Williams@Dartmouth.edu)

Version: 2 Date: 5 March 2014

Author's response to reviews: see over
Dear Dr. Aguilar-Ruiz,

We are attaching a revised version of our manuscript, 2803529011084735 (important changes highlighted), which we have retitled: *Diverse Convergent Evidence in the Genetic Analysis of Complex Disease: Coordinating omic, informatic, and experimental evidence to better identify and validate risk factors*. We thank you for the opportunity to revise and resubmit what we think is an improved manuscript. We would also like to thank the reviewers for their insightful comments which have helped us to enhance our work and clarify its presentation. Specific responses to the reviewers’ comments are provided below.

We hope that the manuscript is now ready for publication. Thank you very much for your time and consideration.

Sincerely,

Timothy Ciesielski
Reviewer’s report
Title: Diverse Convergent Evidence in the Genetic Analysis of Complex Disease: Coordinating Omic, Informatic, and Laboratory based evidence to prioritize findings for further study
Version: 1
Date: 13 December 2013
Reviewer: David Reif

Reviewer’s report:
The authors present an eloquent case for combining multiple lines of evidence to characterize the level of support for putative associations in studies of complex disease. The rationale, illustrative anecdotes, and discussion are convincing that such “convergent lines of evidence” can be useful for prioritizing findings. The authors propose a semi-formal DiCE approach to incorporating evidence from omic (e.g. GWAS), informatic (e.g. KEGG), and laboratory (e.g. model organism experiments) sources.

- Major Compulsory Revisions
The literature cited supports the main thesis of the paper. However, the introduction of DiCE is not supported by the level statistical accompaniment typically presented in this journal. Examples of how this might be improved include:

We thank Dr. Reif for this comment which has helped us to better characterize DiCE. We agree that our approach is semi-formal and not statistically rigorous. However, we note that statistical rigor is to a large extent subjective itself (see excerpt from the text below). That said, we now examine DiCE relative to other ways of thinking, and this has added important discussion to the manuscript. See specific responses below.

(Page 10, Paragraph 3) R.A. Fisher, the father of p-value based inference, provides us with evidence that the application of a thoughtful yet subjective convention can be very productive. He did not view the 5% false positive rate threshold as an immutable postulate but rather as a convenient evidence benchmark that could guide scientific decision making. [30, 31] “If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05 . . .” [32] Thus, much of our biomedical research progress in the last 80 years has been based on a metric that is subjective and imperfect, but useful.

1. Add citation(s) and some exposition of relevant statistical literature on closely-related methods for meta-analysis, inter-rater reliability, and inferential ranking.
We have now added citations and text addressing these issues in the discussion (see the new section on page 13: Comparison of DiCE to existing procedures for knowledge integration). Comparing and contrasting DiCE with these more formal statistical methods helps to better contextualize its proposed role. DiCE is a semiformal, dynamic heuristic that provides an ordinal assessment of the available
diverse convergent evidence for a given genetic factor. It allows us to evaluate both highly significant and not so significant statistical findings in the context of more complete evidence. It is designed to be utilized as an adjunct to statistical validation procedures that leverages non-omic evidence to enhance the detection and vetting of biologically meaningful signals in GWAS data.

2. Provide empirical evidence (from simulation studies or a wider set of literature examples) to justify parameters such as the “>6” threshold. While it is appreciated that this will have an element of arbitrariness, possible alternatives should receive cursory exploration, at least.

We now provide a more thorough justification of the threshold (which should have been written as ≥ 6, see below). Six is used as a cutoff because it requires more than one domain of evidence but not necessarily all three. Therefore, if a domain is missing or flawed we can still find evidence of a relationship. This is a logical, albeit not necessarily quantitative approach.

(Page 6, paragraph 2) We suggest that a total composite score of ≥ 6 indicates strong evidence. Although the scores themselves are arbitrary, they convey ordinal information about the available diverse evidence, and there is a strong rationale for the relationship between the component scores and the chosen threshold. No single category of evidence is necessary or sufficient to achieve a score of 6. This threshold requires convergent evidence from at least two types of data, but protects the conclusion from being deleteriously affected if one category of evidence (out of the three) is missing or flawed. Overall the DiCE process yields a semi-formal dynamic heuristic that is based in logic and empiricism.

3. Discuss the dependence of DiCE on the availability/popularity of studies addressing the association in question. Relatedly, would the authors consider negative context scoring for instances where solid evidence for lack of association is provided in a convergent evidence source?

We thank the reviewer for making these points as they are critical to appropriate implementation of our system (see below, excerpts from page 10 and 11). DiCE can be expected to have a dynamic interplay with the literature. Where a given factor has not been very well or diversely studied it will highlight this. The prospect of negative scoring is enticing but it is problematic in the context of heterogeneity (see below, excerpt from page 6).

(Page 10, last paragraph that ends on page 11). Of course, as with any approach to evidence synthesis the efficacy of this method will depend on the quality of the available prior studies and their annotation as well as the technology used to access this information. The utility of this strategy will be limited where relevant information does not exist, is derived from flawed studies, or is difficult to access. Researchers with expertise in the relevant subject matter and methodologies should be consulted when the value of a piece of evidence is in question. Furthermore, Chanock et al 2007 provides a detailed list of considerations to help guide researchers when making study quality assessments. [1] These judgments may be particularly important in the context of low quality omic studies that could provide a poor foundation for directing further inquiry. Essentially, this approach will be useful where it is thoughtfully applied.
If a preliminary finding is exciting and diverse evidence has not been collected, a low DICE score should encourage researchers to collect the remaining evidence without delay, and thus the quality of the finding should be quickly ascertained. Thus DICE scores can be expected to have a dynamic and productive interplay with the literature.

One could consider developing a more nuanced DICE scoring rubric, by attempting to quantify the number of total validations or rate of validation successes within each evidence category. However, this approach could defeat the purpose of the method. The number of validations within one category and the validation rate within each category do not always have a clear and consistent relationship to the truth of the finding in question, and we propose that at this point they should not be folded into the rubric because of added ambiguity.

- Minor Essential Revisions
4. [page 7] “Hemoglobin S and Malaria Resistance” should be retitled, as it is nearly identical to the previous subheading.

Thank you for pointing this out. We have adjusted the general section title to “Genetic Resistance to Severe Malaria”

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:
I declare that I have no competing interests
Reviewer's report
Title: Diverse Convergent Evidence in the Genetic Analysis of Complex Disease: Coordinating Omic, Informatic, and Laboratory based evidence to prioritize findings for further study
Version:1 Date:11 November 2013
Reviewer: Martin Daumer

Reviewer's report:
I have some concerns with the manuscript at this stage:

- A better overview over the literature should be given. Some key papers are definitely missing (e.g. http://www.ncbi.nlm.nih.gov/pubmed/15705441, http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124) and need to be integrated in the introduction and will have an impact on discussion and conclusion.

We thank Dr. Daumer for highlighting more of Dr Ionannidis' work. We originally cited his work on replication validity in genetic association (Ioannidis et al 2001) but have added new citations (Ioannidis 2005 and 2005) to better set up some key issues that DiCE addresses (see below). These papers allow us to expand the discussion of the advantages of DiCE when the use of standard validation methods alone results in false positive conclusions (see below).

(Page 3, paragraph 1) This conventional confirmation process can help to minimize false positive findings, and in doing so provides fairly compelling evidence for the existence of true associations. Although in recent years it has become evident that chance, limited power, publication bias and a variety of other factors can make this evidence less compelling than it otherwise would be. [3, 4]


(Page 11, Paragraph 2) Widespread application of DiCE also has the potential to increase the credibility of biomedical research by appropriately conveying uncertainty to all audiences and increasing likelihood that highly publicized findings will have biological relevance. Reviewers and editors may still require a specific level of statistical evidence (e.g. p < 5 X 10^-8), but with the addition of a DiCE score both significant and non-significant p-values can be better contextualized in terms of their likelihood of having biological relevance in the pathophysiology of interest. Published findings will be as accessible as they were before DiCE, but bold interpretation, publicity, and translation attempts will be hard to defend in the context of a low DiCE score. A DiCE score can allow readers to quickly gauge the corroborating evidence from beyond the paper they are reading, and a low DiCE score can encourage the lay press to include appropriate caveats in their reports or to wait until the evidence is stronger before reporting.
- I am not convinced about the proposed relative importance of missing true findings as compared to the risk of publishing false positive findings. We agree that the risks of publishing false positive findings are real and DiCE can help to better address these risks (see above). However, we are also quite concerned that many errors are false negatives as evidenced by the lack of heritability explained. We argue that our approach, although not definitive, will help to reduce some of these. DiCE provides adjunctive information and does not require a change to current statistical validation standards. Beyond this we do not want to entertain the full philosophical argument that seeks to achieve the appropriate Type 1 vs Type 2 error balance within a single analysis. We recognize that reviewers and editors will still require whatever validation protocols and p values they find acceptable but with the addition of a DiCE score both low and very low p-values can be better contextualized in terms of their likelihood of having biological relevance in the pathophysiology (see above from Page 11, Paragraph 2)

- The arguments in favor of the new method are more a collection of case studies than an attempt to generate empirical evidence (how have the cases be selected - was there a pre-defined plan? etc.). We agree that our examples are simple case studies chosen because they demonstrate that standard analytic approaches can fail to detect well established biologically relevant factors. Our intent was not to generalize to all possible situations but illustrate how current thinking can be expanded to better utilize more data to solve complex problems.

In order to know if DiCE can reduce type 2 error we needed to study situations where type 2 error is known to have occurred (i.e. the standard analytic approaches did not detect a known signal). In these proof of principle cases (Hemoglobin S and \textit{PPARγ}) DiCE was able to find strong evidence for these factors when conventional validation alone did not. The two exploratory cases (\textit{ATP2B4} and \textit{MARVELD3}) were simply the 2 new hits from Timmann et al (GWAS from the Hemoglobin S-Malaria analysis) where biological relevance was not yet clear.

- Our own experience has been summarized in http://www.biomedcentral.com/1471-2288/8/18/. and may contain some helpful information - although we had at that stage not the additional problem of dealing with "true omics" data. Thank you for this comment. We have now cited this paper (page 4 paragraph 2). Good protocols within one data type (observational data) should reduce false positive conclusions. Furthermore, the consideration of diverse convergent evidence can complement these protocols by integrating additional information that is unlikely to share common biases with the observational data.

Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:
See my forms published e.g. https://peerj.com/articles/82/
Reviewer's report
Title: Diverse Convergent Evidence in the Genetic Analysis of Complex Disease: Coordinating Omic, Informatic, and Laboratory based evidence to prioritize findings for further study
Version:1 Date:16 November 2013
Reviewer: Rita Cantor

Reviewer's report:
The comments below reflect compulsory revisions.

1. Is the question posed by the authors new and well defined?
The authors do not really pose a question. They make a suggestion and then devote the manuscript to supporting the suggestion. The suggestion is that we use a more general method to evaluate SNPs than just GWAS and replication. They present an approach they refer to as DICE. The question is not novel, but their scoring system to evaluate SNPs has not been published before.

Our initial approach emphasized DICE as a potential fix to a problem, but the reviewer’s comments suggested to us that DICE may instead be best presented as a quest to do a better job to find true etiological factors. Therefore, we have now reframed the introduction using the question: How can we increase the amount of knowledge gained from high throughput genetic data? Can we learn more from GWAS data if we better integrate corroborating evidence from non-GWAS sources? DICE is a metric which could facilitate this integration, providing this benefit without having to relax the stringent criteria of current validation.

(Page 11 paragraph 3) Logistics: DICE scores can be quickly added and easily incorporated into any GWAS report

DICE is designed to provide information that complements standard statistical validation methods. Thus DICE can be used to systematically characterize GWAS significant hits to assess for the likelihood of false positive conclusions and suggest future research directions. It can also be used to characterize a small number of sub-threshold statistical associations (e.g. those with the 10 smallest sub-threshold p-values) to assess for the likelihood of false negative conclusions. The utility of DICE may be expanded with the development of semi-automated procedures for calculating DICE scores. With semi-automated implementation protocols DICE could be applied to all nominally significant GWAS findings to detect possible false negative findings in this larger group.

2. Are the methods appropriate and well described, and are sufficient details provided to replicate the work?
The DICE method illustrated with 4 examples, 3 from malaria resistance and one from Type 2 diabetes. They clearly describe how they provided the scores. While I understand what they did, it is not clear that they and I would agree on scores in a novel example. That is, they might interpret something as evidence and give it a certain score, while I might not consider it as evidence. Thus, the concept of DICE is well explained, but the implementation is dependent on the interpretation of the individual.
We agree the DiCE score has components of subjectivity, but the score, if implemented well, will be based on searches that are comprehensive enough to obtain much of the relevant corroborating evidence. This should improve our assessment of potential genetic risk factors. DiCE’s utility does not depend on uniform or perfect implementation.

3. Are the data sound and well controlled?
No experiments have been conducted. Findings have been scored.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
The answer is the same as the one for question 3.

5. Are the discussion and conclusions well balanced and adequately supported by the data?
This question gets at the essence of the manuscript. The text lacks precision that would better explain the points the authors are trying to make. Words are used, but they need to be defined, so the reader can either agree or not agree with the points being made. I will list those words and the section from which they are drawn.

Thank you for these comments about our language. We have now added a glossary of terms to make things clearer (see page 15).

Abstract: masks many true findings (but they are not found)
Agreed. We have changed this to: “masks many true associations”

Cofactors (do you mean covariates)
Cofactor: meaning causal cofactor or component cause (e.g. biological factors that physically interact to generate pathogenic mechanisms). In one place we left the term co-factor to assure we gave this connotation but elsewhere it has been changed to “covariates”. Both terms are defined in the glossary on page 15 (see below).

Cofactor: A component cause or causal cofactor (e.g. biological factors that physically interact to generate pathogenic mechanisms). Component causes are factors that are insufficient to cause disease by themselves but can help cause disease when they occur with other component causes. For more details see [46].

Covariate: A variable that may impact the estimated association between the variable of primary interest and the outcome (via confounding, interaction, and etc.) A covariate may have this impact through causal or non-causal (correlational) relationships. If not properly considered in the analysis covariates may generate bias in the estimated association between the variable of primary interest and the outcome. Cofactors are covariates that may influence estimated associations through causal mechanisms.
False negative findings (I am not sure a false negative is really a finding)
Changed to: “false negative conclusions”

What does ‘confirm’ mean?
We agree that in fact DiCE cannot in theory “confirm”; therefore, we moved away from this term and tried to be more explicit in the text: “. . . identified strong evidence for associations . . .” (see page 2)

I believe that strong and weak evidence are subjective concepts that we would not necessarily agree upon.
We agree that these are subjective concepts but as discussed above (comment #2). Additionally, disagreements among investigators about the evidence can be fruitful in providing useful information. For the purposes of our discussion, by strong we mean more supported, and by weak less supported.

Introduction: define ‘replication’ as it is used in the literature.
Define ‘validation’ as it is used in the literature.
We now define validation as opposed to replication in the introduction (see excerpt below).

(Page 3, Paragraph 1) Testing for replication involves assessing consistency by trying to repeat results in an independent sample from the original population with the same analytic approach. [2] However, many large genetic epidemiology studies and meta-analyses do not use samples from one source population, and therefore, do not attempt replication per se, but validation. [2]

We also define these terms in the glossary on page 15 (see below)

Replication: An attempt to assess the consistency of association by trying to repeat the results in an independent sample from the original population with the same analytic approach. [2]

Validation: An attempt to assess the consistency and generalizability of association by trying to repeat the results in an independent sample from a different population using either the same analytic method or a different approach. [2]

Provide a paragraph describing the rules you are arguing against.
We are arguing against doing things only one way, and nothing more (i.e. without explicitly considering relevant non-omic evidence). This view suggests that we defer drawing strong conclusions from statistical analyses until diverse corroborating evidence is strong. We are not strongly arguing against particular rules but rather we are arguing for additional validation guidelines that can complement the standard statistical evidence. We describe how this perspective can be added to current validation procedures without requiring any change to standard validation criteria (more details can be found on page 11 in the “Logistics: DiCE scores can be quickly added and easily incorporated into any GWAS report” section).
We argue that the conventional procedures for risk factor validation could be enhanced with the addition of a supplementary method that systematically assesses diverse independent lines of evidence. This type of multifaceted strategy could provide useful information in the presence of causal heterogeneity, unrecognized bias, imperfect study designs and other settings where traditional omic validation may yield erroneous conclusions. In this approach researchers actively gather multiple distinct sources of evidence to assess a given factor (e.g., variant, gene, exposure, or pathway) in the pathophysiology of interest. Then multiple findings from various research fields can be combined to gauge whether a critical mass of evidence implicates a given factor. In this process the weaknesses of one methodology can be addressed by the complementary strengths of others; for example, evidence from knockout animal models can support information from genetic epidemiology, and findings from experimental toxicology can strengthen information from environmental epidemiology.

We also further contextualize why this additional validation is needed/useful on page 14 (paragraph 3):

We argue that DiCE, when properly implemented, should leverage multidisciplinary information to reduce rates of both false positive and false negative conclusions. Standard validation protocols implicitly assume that there is one truth (i.e. a marginal finding) and it will be discoverable no matter what the contextual background (covariates, biases, confounding). Furthermore, these validation procedures, when used in isolation, can lead to incorrect conclusions when there is a consistent bias in the observational studies. Therefore, many causal factors will go unnoticed and some meaningless “hits” may be over-interpreted without the development of additional validation approaches, such as DiCE.

Clarify “aim to optimize the differentiation between true signals and noise.”

We are emphasizing that both type 1 and type 2 error should be reduced. Our method can facilitate this by not ignoring potential noise until it looks like noise in at least 2 corroborating ways, and not accepting signal as signal until it looks like signal in at least 2 corroborating ways. We do agree that the point is not optimization but improvement. Therefore, we have clarified this in the text (see below and elsewhere in the paper).

Thus, even strict significance thresholds cannot always separate true positive from false positive findings, and more evidence will generally be needed to determine which associations are worthy of follow-up.

Given that the efficacy and efficiency of research depends on reducing both false positive and false negative conclusions, validation approaches should be developed that can better prevent both types of erroneous conclusions.
P-values in this situation are used to infer association and not causality.

We agree. The sentence now reads:

*Overall, we know that p-values have a variety of weaknesses when being used in scientific reasoning* [11, 12], *and we should recognize these limitations by reinforcing our frameworks for discovery and validation.*

Please include a discussion of why this method will not be arbitrary. Perhaps it’s okay to be arbitrary, but then it seems that a single score would not be appropriate.

DiCE has subjective components, but is based on a rationale that is not arbitrary (revised text below).

*Although the scores themselves are arbitrary, they convey ordinal information about the available diverse evidence, and there is a strong rationale for the relationship between the component scores and the chosen threshold. No single category of evidence is necessary or sufficient to achieve a score of 6. This threshold requires convergent evidence from at least two types of data, but protects the conclusion from being deleteriously affected if one category of evidence (out of the three) is missing or flawed. Overall the DiCE process yields a semi-formal dynamic heuristic that is based in logic and empiricism.*

Most individuals include the additional information in their discussions of observed associations and prioritize SNPs based on them. You are suggesting that replication or a stringent p-value not be required and these more subjective criteria be used as an alternative to replication. It’s an interesting point of view that could help us find some of the ‘missing heritability’.

We thank the reviewer for these comments that have captured our intent.

6. Do the title and abstract accurately convey what has been found?
The title accurately conveys the message of the manuscript.

7. Is the writing acceptable?
The writing is acceptable once the precision in 5. is addressed.

Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:
No computing interests.