

Author's response to reviews

Title: Integrative network analysis reveals different pathophysiological mechanisms of insulin resistance among Caucasians and African Americans

Authors:

Swapan K Das (sdas@wakehealth.edu)
Neeraj K Sharma (nsharma@wakehealth.edu)
Bin Zhang (bin.zhang@mssm.edu)

Version: 2 Date: 17 December 2014

Author's response to reviews: see over
Responses to Reviewers

We thank the Associate Editor Prof Alexander Statnikov and two reviewers for appreciating the value of our research and providing helpful suggestions. We have revised our manuscript as suggested by the reviewers and we provide a point-by-point response to each suggestion below.

Responses to all of the reviewers’ comments are provided below with our responses given in blue.

**Reviewer 1 (Bisakha Ray):**

1) It is also not clear why a population of only non-diabetics was considered and a case-controlled study was not done. Moreover, the number of AAs is much smaller compared to the number of CAs.

*Reply:* Our strategy is based on the premise that modulation of glucose homeostasis traits (including changes in insulin sensitivity, $S_I$) precede and are more prevalent than type 2 diabetes (T2D), and that gene expression changes causally related to these intermediate traits are masked by additional factors triggered at the onset of T2D (hyperglycemia and dyslipidemia) and altered by T2D complications, duration of the disease, and T2D therapy. The primary focus of this study is on insulin sensitivity ($S_I$) as the primary physiological outcome. We sought to identify ethnic-specific correlations of expression profiles in adipose with $S_I$ and related pre-diabetic metabolic phenotypes. Thus, we included only non-diabetic individuals in this study. We agree that our study cohort includes far more Caucasian subjects (N= 99) than African Americans (N= 37). Unequal racial distribution in our cohort was influenced by the population demographics of Arkansas (recruitment site) and volunteer participation bias. However, to our knowledge, this is the largest single cohort of adipose tissue gene expression related to glucose homeostasis that includes phenotypically well-characterized subjects of both races. While sample size is an important factor for all analyses (e.g. differential expression, gene-trait correlation and network construction), the integrative network analysis was developed in this study to minimize the impact of a relatively small sample size, to obtain a more comprehensive understanding of differentially expressed genes and the network context under which individual genes operate. The highly significant overlap between the results from the two ethnic groups studied here and an independent but much larger study (the deCODE study) in Caucasian subjects (*included in the revised manuscript*) suggests that the integrative network analysis approach was appropriate in balancing for the impact of the small sample size.

We address this issue in Page 19 of the current revision.

2) The heuristic for identifying key regulators using NHNN needs a reference.

*Reply:* We have now added a relevant reference for this method.

3) In Table 1, the proportion of males to females is different in CAs and AAs. The authors might want to consider this as gender and adipose tissue distribution are players in insulin resistance.
Reply: The Caucasian group in our study includes 42 males and 57 females, and the African American group includes 21 males and 16 females. The difference in gender distribution between two groups is not statistically significant (two-sided Fisher’s exact test p value = 0.176). Moreover, our all analyses were appropriately adjusted for age and gender.

Reviewer 2 (Petter Storm):

General Comment: The paper is very well written and easy to follow. The flow is perfectly logical and overall I am quite optimistic about this paper.

Response: We thank you for appreciating the quality of our manuscript and helpful suggestions

1) No validation of the results are performed (save for technical validation using qPCR) using alternative data sets/methods. This makes the paper descriptive and lowers my enthusiasm for the results. If alternative methods (for example metabolomics) has been performed on the data set and can be related or, alternatively, the gene sets identified in this paper can be tested for association in separate data sets that would greatly strengthen the paper.

Reply: We agree that validating our results in an independent dataset would be valuable. Our primary goal was to identify transcript co-expression modules correlated with insulin sensitivity and other related metabolic phenotypes in each ethnic group. However, to our knowledge currently no publicly available data set includes an adipose tissue expression profile of metabolically characterized (by FSIVGT) non-diabetic subjects, of both Caucasian and African American ethnicity. Published gene expression studies are mostly in Caucasian subjects. In an ongoing study, our laboratory is currently generating adipose and muscle gene expression data on African American subjects.

However, as suggested by the reviewer, we now compared our co-expression modules with the coexpression network modules from an independent human adipose gene expression data (called deCODE data) from 640 Caucasian subjects. Twelve modules in the AA network and nineteen modules in the CA network significantly (Bonferroni corrected FET p-value < 0.05) overlap with the network modules from the deCODE data. Notably, 5 of the top 10 Caucasian modules associated with SI identified in our study significantly overlap with the modules from the deCODE data, and the yellow (inflammatory response), pink (mitochondrion), and peru (steroid metabolism) modules are among the top five most conserved modules, with FET p values of 3.6 X 10^{-151} (5.6 fold enrichment), 6.6 X 10^{-58} (17.5 fold enrichment) and 1.4 X 10^{-29} (13 fold enrichment), respectively. On the other hand, 4 of the top 10 African American modules associated with SI significantly overlap with the gene modules from the deCODE data, but to a lesser degree (Bonferroni corrected FET p-values are between 8.2 X 10^{-3} and 2.4 X 10^{-5}). This comparison basically validates our findings about both the African American and Caucasian subjects, although the results from the Caucasian data are more profoundly replicated in another independent study of Caucasian subjects. However, clinical and metabolic phenotype data are not publicly available for deCODE study subjects to validate our other findings, including correlations between module and phenotypes.

The new results are reported on Page 13 of the current submission.
2) The “integrative network approach” is given a lot of focus in this paper. I think the justification for this type of method and its applicability in this paper is lacking.

Reply: A pressing task in analyzing omics data like those generated by this study is to identify and visualize a global landscape of interactomes that contribute to clinical endpoints such as disease onset and progression. Weighted gene coexpression network analysis (WGCNA) has emerged to solve this problem through identification of gene modules comprised of highly interconnected genes over a gene-gene interaction heatmap [17]. WGCNA has been widely used to identify pathways and gene targets for a variety of common human diseases such as cancer, atherosclerosis, Alzheimer’s disease, obesity, and diabetes. A key contribution of the work is the integration of differential expression analysis, gene-trait correlation analysis, and gene network analysis to identify not only the coexpressed gene modules that contribute to insulin sensitivity, but also the key causal regulators.

We justified the use of the integrative network approach on Page 5 in the revised submission.

3) Are the data deposited in a public repository?

Reply: The gene expression data were generated using an institutional intramural fund (non-federal / non NIH fund). Data presented in this paper formed the basis of our several federal and nonfederal research grant applications. We also anticipate additional manuscripts which use a part of the data for analysis to address different hypotheses (other than those in our current paper). Decisions are pending on those grants and manuscripts. Thus, our data are not yet in a public repository. If it is a requirement for the compliance with journal policy, we will upload our gene expression data to the Gene expression omnibus (GEO) database as suggested by the reviewer. However, based on our experience, access to post computed/analyzed data is significantly greater than for raw data. We have included most of our extensive and unselected analysis results as an online supplementary result tables (12 result tables in Excel file format) with this manuscript for our readers.

4) It is surprising to this reviewer that the overlap in transcriptional response in relation to SI is so small comparing AA vs CA (Figure 2). For example only 800 genes are shared whereas 2400 genes are unique to either group. Has the authors excluded every possible alternative explanation?

Reply: Our study indicates that: “We identified significant correlations of 1,625 transcripts with insulin sensitivity (Si) in both CA and AA subjects (812 and 813 transcripts showed positive and negative correlations, respectively, with Si).” Thus a large proportion of Si associated transcripts were common to both ethnicities. We have discussed sample size, statistical power, and methodological issues that may influence this finding in the revised manuscript.

5) A number of hypotheses are tested throughout this paper and it is quite dense in p-values. Could the authors please clarify on how correcting for multiple testing was performed?

Reply: We have added a more detailed description of our analytic approach to the Materials and Methods section (Page 10). In general, we used two different methods to control for multiple testing issues: (1) the highly conservative Bonferroni correction method, to constrain
the study-wise significance level; and (2) an empirical false discovery rate (FDR) method that constrains the overall rate of false positive events. For instance, for the enrichment of functional categories in modules, we corrected for the number of modules, Gene Ontology processes, and other categories tested. For the enrichment of expression of clinical trait correlation signatures in modules, we used a combination of FDR <0.05 and nominal P<0.05. In addition, we have revised description of the various statistical procedures throughout the manuscript to provide much more detail.

6) The data are adjusted for using age and gender. Data driven identification of covariates using e.g. PEER could help.

Reply: A big concern with PEER (Probabilistic Estimation of Expression Residuals) is that we do not know whether the covariates it identifies are biological or non-biological. Therefore, it is dangerous to correct the data for biological co-variates. PEER is more suitable for genetic association study of population-based gene expression (eQTL analysis) since it can remove the unintended non-genetic perturbations. However, for the gene coexpression network analysis like WGCNA used in this study we want to capture the total interactions between genes induced by both genetic and non-genetic perturbations. In addition, our coexpression network analysis shows that there is no obvious batch effect on the network modularity, as reflected by the relatively small size (<2000) of the largest modules in the two coexpression networks and the highly enriched gene ontology categories of the largest modules.

7) There is quite an unbalance in number of AA and CA participants. This not addressed/discussed adequately by the authors.

Reply: Please see response to comment 1 from Reviewer 1. We address this issue in Page 19 of the current revision.