Author's response to reviews

Title: Integrated Genome-wide Association, Coexpression Network, and Expression Single Nucleotide Polymorphism Analysis Identifies Novel Pathway in Allergic Rhinitis

Authors:

Supinda Bunyavanich (supinda.bunyavanich@mssm.edu)
Eric E. Schadt (eric.schadt@mssm.edu)
Blanca E. Himes (blanca.himes@channing.harvard.edu)
Jessica Lasky-Su (jessica.a.su@gmail.com)
Weiliang Qiu (stwxg@channing.harvard.edu)
Ross Lazarus (ross.lazarus@channing.harvard.edu)
John P. Ziniti (rejpz@channing.harvard.edu)
Ariella Cohain (ariella.cohain@mssm.edu)
Michael Linderman (michael.linderman@mssm.edu)
Dara G. Torgerson (dara.torgerson@ucsf.edu)
Celeste S. Eng (celeste.eng@ucsf.edu)
Maria Pino-Yanes (mariadelmar.delpinoyanes@ucsf.edu)
Badri Padhukasahasram (BPADHUK1@hfhs.org)
James J. Yang (JYANG2@hfhs.org)
Rasika A. Mathias (rmathias@jhmi.edu)
Terri H. Beaty (tbeaty@jhsph.edu)
Xingnan Li (xinli@wakehealth.edu)
Penelope Graves (pgraves@u.arizona.edu)
Isabelle Romieu (romieu@iarc.fr)
Blanca del Rio Navarro (blancadelrionavarro@gmail.com)
M. Towhid Salam (msalam@usc.edu)
Hita Vora (hvora@usc.edu)
Dan L. Nicolae (nicolae@galton.uchicago.edu)
Carole Ober (c-ober@bsd.uchicago.edu)
Fernando D. Martinez (fdmartin@email.arizona.edu)
Eugene R. Bleecker (ebilecker@wfubmc.edu)
Deborah A. Meyers (dmeyers@wakehealth.edu)
W. James Gauderman (jimg@usc.edu)
Frank Gilliland (gillilan@usc.edu)
Estaban G. Burchard (Esteban.Burchard@ucsf.edu)
Kathleen C. Barnes (kbarnes@jhmi.edu)
L. Keoki Williams (KWILLIA5@hfhs.org)
Stephanie J. London (london2@niehs.nih.gov)
Bin Zhang (bin.zhang@mssm.edu)
Benjamin A. Raby (rebar@channing.harvard.edu)
Scott T. Weiss (scott.weiss@channing.harvard.edu)

Version: 4
Date: 27 May 2014
Author's response to reviews: see over
May 27, 2014

Dr. Struan Grant
Section Editor
BMC Medical Genomics

Dear Dr. Grant:

Enclosed please find marked-up and clean versions of our manuscript, “Integrated Genome-wide Association, Coexpression Network, and Expression Single Nucleotide Polymorphism Analysis Identifies Novel Pathway in Allergic Rhinitis,” as well as a detailed point-by-point response to the concerns of each reviewer. We thank you and the reviewers for your helpful comments, which have significantly strengthened our manuscript.

RESPONSE TO REVIEWERS

Reviewer 1

Comment (C)1: “They authors only used weighted gene co-expression network analysis to generate the co-expression network. It would be useful to construct several co-expression networks using both linear and non-linear models. Such a comparison among several networks would allow identify “a real world network”.”

Response (R)2: The aims of this manuscript were to perform genome-wide association studies (GWAS) of allergic rhinitis in ethnically diverse populations, and to link these GWAS results to function through coexpression network, eSNP , and pathway approaches. We accomplished these aims using well-recognized methods and focused our manuscript on these findings. Although we agree that building several different coexpression networks would be very interesting, this would distract from the central aim and findings of the manuscript. We have used a previously applied, well-established, well-recognized, and validated method to construct the coexpression network (Zhang and Horvath, Stat Appl Genet Mol Biol 2005, cited 440 times). This method has yielded networks that have informed on many disease areas [14, 28, 33-40]. A comparative study of coexpression networks would yield compelling content for a paper in itself, which we will consider.

To clarify that we used a well-established, validated method to construct our coexpression network, we have added additional references and now state in lines 483-484 (of the marked manuscript): “We used a previously applied, well-established, well-recognized, and validated method to construct the coexpression network [14, 28, 31, 33-40].”

C2: “Authors did not generate any “random” network. It would be ideal to re-shuffle the data and generate random networks using identical parameters for the significance analyses of modules/pathways identification.”

R2: In response to this comment, we generated multiple random networks and carried them forward into the integrated analyses. Please see Supplementary Results 2 and Figure S7: Plots of scale-free topology for coexpression networks constructed from randomized CD4+ gene expression data
In Methods, lines 492-495, we now state, “To provide support for the specificity of our coexpression network, we generated multiple random coexpression networks where gene assignments were randomized (Random Networks 1-3), as well as random networks where the gene expression levels were randomized (Random Networks 4-6). We carried these random networks forward into the integrated analysis as well.”

In Results, lines 308-309, we have added: “Results from randomized networks did not yield meaningful results (Supplementary Results 2, Figure S7).”

We have added Supplementary Results 2, where we state: “Randomized coexpression networks
We generated multiple random coexpression networks where gene assignments were randomized (Random Networks 1-3), as well as random networks where the gene expression levels were randomized (Random Networks 4-6). The randomization of gene names preserved the correlation structure, resulting in the same network topology as we observed with the original, non-randomized data (Figure S7). However, enrichment of these modules for eSNPs followed a random distribution, in contrast to the strong eSNP enrichment patterns seen with the modules built with the non-randomized data (Table 3). In Random Networks 4-6, we randomized the entire matrix, thus destroying the correlation structure. The network topology was therefore changed (Figure S7), and in fact was no longer able to well fit a power law distribution. Distance structures became empty and no modules could be detected in any of these networks. This is consistent with previous reports that meaningful scale-free networks cannot be obtained when correlation structures are destroyed [24].

C3: “Is it possible to integrate large-scale publically protein-protein interactions datasets onto their GWAS/transcriptomics datasets?”

Our group’s previous work demonstrates the value of integrating multiple dimensions of data to build networks. For example, we have previously integrated metabolite concentration, RNA expression, DNA variation, DNA-protein binding, protein-metabolite interaction, and protein-protein interaction data to construct probabilistic causal networks that elucidated the complexity of cell regulation in a segregating yeast population (Zhu et al., Plos Biology 2012). However, in this case, we were in the fortunate situation of having genotype and gene expression data for the same individuals, which we deemed as distinctly favorable and sufficient for what we were aiming to show. Data used to support the majority of prior GWAS have been gene expression data, and most of these gene expression data were from individuals different from those who were genotyped (lines 134-135). A distinct strength of the results we have presented is that we use gene expression data generated from disease-relevant tissue collected from the same individuals who were genotyped to enhance power over standard GWAS to identify disease-relevant loci (lines 143-145). The use of multiscale data generated from overlapping subjects minimizes potential bias due to differences in subjects who were genotyped versus subjects with transcription data (line 134-147). While integration of publically available protein data could be considered, these data would not be from the same individuals and could contribute bias. As presented, our study provides a rich context for our allergic rhinitis GWAS results that integrates coexpression network, eSNP, and pathway approaches. The data used in this study will be publically accessible, and therefore available for further integration with publically available datasets by any group.
C4: “Authors may also like to explore the co-regulation aspect of the network. It is suggested to identify common cis-regulatory elements in the modules and investigate the common transcription factor(s).”

R4: Thank you for this suggestion to explore cis-regulatory elements. In response to this comment, we performed cis-regulatory region enrichment analysis using T-lymphocytes in different states. We found a 2-fold enrichment for eSNPs that fell in cis regulatory regions (P = 0.001). These eSNPs falling in cis regulatory regions were 5.5 fold enriched for falling in the blue module (P value 1.0 x 10^{-5}), one of the largest modules with 2059 genes. The blue module was 1.6-fold enriched for the GO term ATP binding (FDR-adjusted P value 1.6 x 10^{-9}). The above findings were not dependent on allergic rhinitis status, however, which was likely due to small numbers limiting intersections when we tried to enforce reasonable significance along the GWAS and eSNP dimensions. Given these intriguing preliminary findings, we agree that further investigation of co-regulation and common transcription factors would be interesting to pursue with larger numbers.

Reviewer 2

C1: “Asthma is a potential confounder in this study. Could the authors please indicate the percentage of asthma among their Allergic rhinitis cases and controls? and explain why they didn't include asthma status as a covariate in their association analysis?”

R1: We have added information on percentage of asthma among allergic rhinitis cases and controls as requested. We did not include asthma status as a covariate in our association analyses because none of the previous GWAS of allergic rhinitis adjusted for asthma (Andiappan et al. 2011; Ramasamy et al. 2011; Hinds et al. 2013) and we wanted to maximize comparability of our findings with previous work. However, recognizing that individuals with allergic rhinitis frequently have comorbid asthma, and allergic rhinitis with and without asthma may be subphenotypes, we did perform GWAS of allergic rhinitis stratified by asthma status.

In Results lines 206-212, we now state:
“Individuals with allergic rhinitis frequently have comorbid asthma [1, 32]. Indeed, we observed that 2051 (76%) of those with allergic rhinitis had asthma, and 1195 (41%) of those without allergic rhinitis had asthma. As subphenotypes of AR based on asthma status are possible, we also performed secondary GWAS stratified by asthma status. These results are shown in the supplementary file (Supplementary Results 1, Table S4, and Figure S6) and similarly showed ethnicity-specific findings. In Table S5, we show the sample composition of the stratified analyses according to asthma status.”

In Methods lines 458-459, we now state:
“To allow for comparability with previous GWAS of allergic rhinitis [4,8,9], we did not include asthma status as a covariate.”

C2. “For the genome wide significant SNPs shown in Figure S2, some do not have enough supporting evidence from neighboring SNPs in LD with them. As this is one of the important parts of their findings and the basis for other parts presented in the manuscript, further evidence is needed. A good validation method would be Taqman SNP genotyping assay. If this is not feasible to do, the authors please show the genome studio cluster plots of these SNPs in Figure S2.”
**R2:** The SNPs shown in Figure S2 do not have many SNPs in LD with them (please see new Figure S4: Regional LD plots for genome-wide significant loci (P value ≤ 5 X 10^{-8}) in the GWAS of allergic rhinitis among Latinos), so the absence of support from neighboring SNPs reflects absence of SNPs in LD rather than lack of association with such SNPs. We agree with this reviewer that fine mapping of top regions would be helpful to identify which, if any, SNPs lead to biologically important functional changes. Unfortunately, as there are multiple collaborative studies occurring within the EVE consortium (http://eve.uchicago.edu), there is limited shared DNA, and Taqman SNP genotyping for the SNPs in Figure S2 across centers is not feasible at this time. We would like to emphasize that the SNPs reported in Figure S2 were directly genotyped in 2 of the 7 populations, and the imputation scores for these SNPs demonstrated good confidence (please see new Table S2: Mean Rsq values for genome-wide significant loci in the GWAS of allergic rhinitis among Latinos). The platform used in these 2 populations was the Affymetrix 6.0 array (i.e. not Illumina), so we do not have Genome Studio cluster plots to provide. We would like to reassure readers that each center used very conservative metrics to filter SNPs as detailed in Torgerson et al., Nature Genetics 2011 (cited in Methods, lines 443-445), so it is unlikely that SNPs reported do not represent true genotypes.

To clarify the above, we have added the following to Results, lines 195-198:

“For context, each of these SNPs was directly genotyped in 2 of the 7 populations, imputation was performed using very conservative metrics [25], and the imputation scores for these SNPs demonstrated good confidence (Table S2). The regional LD plots for these loci (Figure S4) show that there were limited SNPs in LD with these genome-wide significant loci.”

Additionally, although some regional association signals do rely on a single SNP as shown in Figure S2, the coexpression network analysis that we performed to provide functional context for the GWAS results did not dependent on individual SNPs. Coexpression modules tagged by GWAS loci were based on genes within 250kb of GWAS loci with P < 1x10^{-6}, so the neighborhoods of all loci in Table 2 (i.e. not just those with P< 5x10^{-8}) were considered. Thus, even if some single SNP results were not true, the overall results from the coexpression would not change significantly.

**C3:** “The coexpression network was constructed based on gene expression data from the CAMP cohort, the samples of which were 100% from European Americans. However the GWAS loci (p<1x10^{-6}) were mainly from Latino population. It is known that gene expression varies between different ethnic groups. It is a limitation of the study. The authors please further expand discussion on this point.”

**R3:** We have expanded our discussion of this limitation as recommended.

In lines 414-418, we now state:

“We had CD4+ lymphocytes from European-American CAMP subjects only, and it is possible that coexpression results would have differed had we additionally had expression profiles from subjects of other ethnic backgrounds, as gene expression can vary by ethnicity. While expression differences can change with ethnicity, the connectivity structure is expected to be much more highly conserved, however, as is even seen across species [57].”

**C4:** “Did the authors take into account of imputation uncertainty when testing for association?”

**R4:** Yes, SNPs with imputation quality scores below a threshold (Rsq < 0.3) were removed from the analysis.
We referenced quality control protocols in lines 442-443, but to emphasize, this important point, we have added this statement to lines 443-444:
“Of note, SNPs with imputation quality scores below a threshold (Rsq < 0.3) were removed from the analysis.”

C5: “Please show the manhattan plot separately for study in each ethnic group and the meta-analysis for better viewing and comparison.”

R5: We have added these plots as Figure S1: Manhattan plots for the European American, Latino, and African-American genome-wide association and meta-analysis results for allergic rhinitis.

In lines 183-184, we now state:
“For additional views, Figure S1 shows the Manhattan plots separately for each ethnic group and for the meta-analysis.”

C6: What is the genomic inflation factor of each individual GWAS?

R6: The genomic inflation factors for each individual GWAS are shown in the new Figure S1: Manhattan plots for the European American, Latino, and African-American genome-wide association and meta-analysis results for allergic rhinitis

C7: “The authors please check the meta-analysis p-value for SNP rs868688 presented in Table 2, as the p-value for each ethnic group is 0.49, 0.89 and 7.3E-07 and the meta-analysis p-value is 7.3E-07, which seems that two ethnic specific studies have no weight in the met-analysis. This seems problematic and is not consistent with results of other SNPs presented in Table S2.”

R7: This has been corrected in Table 1. We apologize for this typo. Downstream results were not affected.

C8: “In Table S1, please indicate clearly what allele is presented?--risk allele or reference allele?”

R8: We have added the following note to Table S1.
“*Reference allele assigned as the allele coded 0 in the HapMap release 21 phased consensus haplotypes during genotype imputation in MACH”

C9: “In Figure 3B, please show the gradation of color with the corresponding p-value range.”

R9: We have added a key for the color gradation as recommended (please see revised Figure 3B)

C10: “Please indicate the sample composition (the number of Allergic Rhinitis cases and controls in each ethnic specific study) in the stratified analyses according to asthma status.”

R10: We have added this information as Table S5: Sample composition of GWAS stratified by asthma status

In line 211-212, we now state:
“In Table S5, we show the sample composition of the stratified analyses according to asthma status.”
Thank you very much for your consideration of our work for publication in *BMC Medical Genomics*.

Sincerely yours,

Supinda Bunyavanich, MD, MPH  
Assistant Professor  
Department of Genetics and Genomic Sciences  
Department of Pediatrics- Allergy and Immunology  
Icahn School of Medicine at Mount Sinai