

Author's response to reviews

Title: Multifactor Dimensionality Reduction Analysis identifies specific nucleotide patterns promoting genetic polymorphisms

Authors:

Eric Arehart (eric.j.arehart@dartmouth.edu)
Scott Gleim (scott.r.gleim@dartmouth.edu)
Bill White (bill.white@dartmouth.edu)
John Hwa (john.hwa@dartmouth.edu)
Jason Moore (jason.moore@dartmouth.edu)

Version: 2 **Date:** 26 October 2008

Author's response to reviews: see over

Introduction to Letter of Response

We are grateful for the insightful comments from the Reviewers. The questions we received have helped us to improve our manuscript. The Reviewers' references to our work as "interesting" and being of "outstanding merit and interest" have been particularly encouraging. Additions within the document have been bolded and underlined to facilitate your review.

Reviewer: B. McKinney

1. "How does knowing the position relative to a mutation site help you predict the future occurrence of a mutation?"

Leveraging the MDR approach, we have built an association between the occurrence of specific polymorphisms and patterns of nucleotide distributions surrounding that location. Unfortunately, due to a lack of directionality for the identified polymorphisms (i.e., handling an A→G change as equivalent to a G→A change), does not allow definitive mutational prediction. Nevertheless, identification of which surrounding sites obtain significant association with polymorphism occurrence and the distribution of nucleotide patterns surrounding that position provide the highest resolution of polymorphism associated sequence patterns currently available.

2. "Motif in bioinformatics usually means a sequence pattern."

Under your advisement, we have refrained from using the term 'motif' throughout the manuscript, with the hope that our current description of nucleotide distribution patterns improves clarity for the reader.

3. "A few more words about the permutation testing procedure would be beneficial for the general reader."

We have added a brief definition of permutation testing to facilitate comprehension by unfamiliar readers.

4. "Can you comment on the independence of the NCBI and Broad data sets?"

The NCBI dataset includes the data utilized in our Broad Institute pilot set, as the NCBI dataset incorporated the Broad database following our initial pilot analysis.

5. "An optional suggestion for the coordinate system... [-10, 10] with 0 being the mutation site."

We agree that labeling the positions relative to the mutation site aids in clarifying the positions in the discussion, and have adopted this improved system throughout the manuscript.

6. "... in Fig. 5... it might help the reader by highlighting the sites... involved in each model..."

To clarify which sites were found to be significant, we have faded the surrounding nucleotide distributions, giving the foreground sites the emphasis.

Reviewer: R. Casadio

1. **“... thorough explanation of the different models, their occurrence in the human genome and their relation to already known genetic diseases would add to the paper.”**

We have added a reference to the involvement of this type of alteration with known genetic diseases. As the input data lacks mutation directionality, we are not yet able to specify which changes are directly responsible for known genetic diseases; however, we are eagerly seeking ways to expand these findings to address these questions.

2. **“...it is not clear to the reader whether one model over the other is more likely to find hypothetical SNPs... would it be possible to associate a scoring index to the different models?”**

It would be ideal to be able to refer to our identified models and select which will have more predictive power. The lack of directional mutation information prevents these conclusions. We are currently evaluating ways to biologically confirm the influence of certain distribution patterns, as well as direct those findings toward a predictive methodology.

Reviewer: S. Williams

1. **“...important that the authors detail somewhere how the direction of the mutations was determined.”**

With the integration of the Broad Institute database into the NCBI dataset, information regarding polymorphism directionality was removed. From our understanding this information was determined prior to database entry at the Broad Institute, an adequate description of the methodology has not been found. Accordingly, we have removed references to directionality of mutations, losing the primary benefits of our initial pilot study. It is likely that changes occur from most prevalent nucleotide to least prevalent nucleotide but such predictions may not always be the case. It may be possible that bidirectional changes can occur at these positions. This would need confirmation using in vivo models.

2. **“...the choice of controls needs to be adequately justified...”**

We have added a brief explanation for the generation and employment of the control sequences for the NCBI dataset, and clarified the description of using the MDR impute functionality for the Broad Institute dataset. Following our initial use of the Broad Institute data, we decided that directly utilizing sequence data lacking polymorphisms would be preferable, as these sequences were biologically relevant. We feel this approach avoids the pitfalls of random number generation and as such allows for a more “biological” approach to sequence analysis.

3. **“... evolutionary literature that directly asks questions about mutation and sequence milieu...”**

Despite numerous investigations into the relationship between sequence and mutation, to our knowledge current approaches to sequence analysis have not afforded the identification of detailed sequence patterns associated with SNPs. Previous work by others has added to the literature in various ways. Zhao et.al. have addressed long range assessment of individual nucleotides in the context of all general SNPs; however, such methods were not focused on resolving nucleotide interactions relative to individual SNP types. Golding and Glickman identified a tendency toward co-occurrences of multiple simultaneous mutations through phylogenic analysis of alpha-interferon genes. Our work adds to the growing literature by identifying specific flanking region base-pair interactions immediately adjacent to specific SNP models and may in future aid in the identification of genome regions with a high predisposition for SNP mutagenesis.