molecular systems biology

Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling

Yuval Tabach, Tamar Golan, Abrahan Hernández-Hernández, Arielle R. Messer, Tomoyuki Fukuda, Anna Kouznetsova, Jian-Guo Liu, Ingrid Lilienthal, Carmit Levy, Gary Ruvkun

Corresponding author: Gary Ruvkun, Massachusetts General Hospital

Review timeline:

Submission date:	27 February 2013
Editorial Decision:	17 April 2013
Re-submission:	21 June 2013
Editorial Decision:	28 June 2013
Author's response:	02 July 2013
Editorial Decision:	07 August 2013
Revision received:	24 August 2013
Accepted:	29 August 2013

Editor: Thomas Lemberger

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 17 April 2013

Thank you again for submitting your work to Molecular Systems Biology and apologies for the delay in getting back to you. We have now heard back from the three referees whom we asked to evaluate your manuscript. As you will see from the reports below, the opinions of the reviewers are not unanimous. Thus, while reviewer #2 and #3 provide supportive recommendations, reviewer #1 raises significant concerns and is much less supportive. The major concerns refer to the limited novelty of the approach as compared to previous works (for example, Tabach et al 2013) and to the disconnect between the global analysis of the phylogenetic profiles and the follow up validation.

In view of the divergent opinions expressed by the reviewers, I have read the paper again and I have to agree with reviewer #1 that, from a methodological point of view, the conceptual advance remains limited as compared to Tabach et al 2013. Having said that, I also agree with reviewer #2 and #3 that the performing 'continuous' phylogenetic profiling across 86 species to map human disease genes is a potentially interesting application. It is also appreciated that two examples of follow up investigations are provided (MITF/SuH and ccdc105). However, I also share the opinion of reviewer #1 that these examples appear rather 'anecdotal'. While it is interesting that 4-5% of the 3513 HPO groups include genes with correlated profiles, a more global characterization of the forward predictive value of the human phylogenetic clusters should be performed. This could also entail a deeper integration with other resources to characterize the power of continuous phylogenetic profiling for "human disease locus discovery", as mentioned in the title.

Given these considerations, we agree with reviewer #1 that the analysis, as it stands, remains

© European Molecular Biology Organization

incomplete and the conceptual advance in terms of methodology remains too limited for publication in Molecular Systems Biology. As such, I am afraid I see no choice but to return the manuscript with the message that we cannot publish it.

Nevertheless, the application of phylogenetic profiling to study systematically human disease genes could be a potentially interesting topic. We would thus not be opposed to consider a new submission on the topic provided the analysis can be significantly extended to provide a much more systematic validation of the forward predictive value of phylogenetic clusters to identify new disease-related loci. It is also possible that using large-scale genetic perturbation/expression profiles panels such as the new LINCS CMap (lincscloud.org) could provide a means to evaluate more systematically the biological significance of your phylogenetic analysis (eg do knockdown/overexpression of (disease) genes belonging to the same phylogenetic cluster tend to produce overlapping signatures?) or to improve its predictive power using an integrative approach.

We recognise that this may involve further experimentation and analysis, and we can give no guarantee about its eventual acceptability. However, if you do decide to follow this course then it would be helpful to enclose with your re-submission an account of how the work has been altered in response to the points raised in the present review. This would have a new number and receipt date.

I am sorry that the review of your work did not result in a more favourable outcome on this occasion, but I hope that you will not be discouraged from sending your work to Molecular Systems Biology in the future.

Thank you for the opportunity to examine this work.

Referee reports:

Reviewer #1:

In this work, Tabach et al. present an approach for finding genes that are associated with particular diseases. They employ Tabach's previous analysis of phylogenetic profiling in 86 genomes and show that the method detects clusters of genes of which a significant amount are linked to the same disease. For two cases there are some specific additional experiments. MITF is involved in melanoma and one gene with a similar phylogenetic profile was selected for further analysis. It is shown to be functionally related. In the second validation, gene ccdc105 was linked in meiosis and spermatogenesis.

Overall the result that genes involved in the same disease is not surprising; they should be functionally related and identifying such relationships is the idea of phylogenetic profiling. Making specific applications to diseases is also not novel. The authors themselves acknowledge (in the Results section) that "Phylogenetic profile analysis has been a powerful tool for the study of two types human disease, Bardet-Biedl syndrome (Mykytyn et al., 2004) and mitochondrial diseases (Pagliarini et al., 2008)."

The presentation of this work is extremely disjointed with two distinct parts that do not work well together. The first is the analysis of phylogenetic profiles. This is the same kind of analysis as that in the Nature paper (though the methods are provided again for some reason). The analyses are very poorly described or not even described at all. For example, Figure 2 has six sections but only one is described. There are some actually good controls there - how could the authors completely ignore them in the text?

The second part itself has two parts each completely disjointed as well. The reason why RBP-Jk was selected in order to query for a relationship with MITF is not clear. To make a real prediction, one would need to make a list of closely related profiles and then provide real rationale as to why one was singled out for detailed analyses. Finally the ccdc105 analysis suffers from the same lack of relationship to phylogenetic profiles and in particular to diseases. It seems completely tacked on. For example, for the identification of the MITF-associated factor there was no evidence provided that RBP-Jk is also involved in cancer as much as MITF. Collectively, the work falls short of both a

© European Molecular Biology Organization

general analysis of phylogenetic profiles in diseases and genes involved in melanoma.

Reviewer #2 :

The authors present an extension oh phylogenetic profiling to human proteins. By aligning human protein to 85 other eukaryotic proteomes, they are able to construct analog phylogenetic profiles, and show that clusters of these often overlap genes sets found in MSigDB. They also validate two of the functional inferences that they make using additional experimental data. Overall I found this to be a strong study that convincingly demonstrates the utility of phylogenetic profiles to assign functions to human proteins. My primary concern had to do with the difficulty of inferring profiles in organisms that have poorly annotated genomes. How can one be sure that the absence of a strong homolog is due to the fact that it is truly lacking from a genome, rather than incorrectly missing from the existing annotation. I assume this could be a significant issue for organisms with poorly annotated genomes. On the other hand, if a protein is missing across a set of related organisms then this is less likely to be due to misannotation. The authors should address this issue and discuss why it is not likely to have a large impact on their results. It would also be useful to see additional benchmarking to convince the reader that the algorithm choices that were made (i.e. nature of value within profiles, and metric for profile comparisons) are indeed optimal or close to optimal.

Reviewer #3:

The manuscript by Tabach and colleagues describes a phylogenetic profiling algorithm to discover functionally co-evolved genes and their relationship with human diseases. By clustering continuous conservation scores across 86 organisms, the authors found enrichment of functionally related groups of genes in the phylogenetic profiling tree, as well as those that are not fully expected to be co-evolved, such as those sharing similar promoter binding sites and miRNA targeting motifs. The authors further demonstrate with convincing experimental data on RBPjk and MITF interaction, as well as potential roles of ccdc105.

Overall, this is a very well done piece of work that will be of great interest to scientists in many fields.

I have one concern, which may have already been addressed. The concern is associated with the interesting yet unexpected finding that genes with common regulatory motifs, such as many miRNA targets, are co-evolved. I wonder whether this finding could be an artifact of the algorithm. Specifically, if we consider a family of proteins that evolves from a single protein in a more primitive organism, this family will have a highly correlated phylogenetic profile. This leads to the question whether the enrichment, of co-evolved genes in miRNA targets or those share common promoter motifs, was driven by a single or a small number of protein families. If so, this provides a possible explanation for common regulatory motifs in such gene sets. If not, the authors' suggested model of independent acquisition of regulatory motif may be possible. I noticed that in the method, the authors effectively eliminated contribution to correlation scores by highly homologous genes in human, but I am not sure whether this elimination is "clean" with regard to the issue discussed above.

Minor:
1. There are two figure 2e.
2. There is no page number in the manuscript.
3. The reference to supplemental figure 6a, 6b is incorrect-should be sup fig 4?
4. The reference to figure 2e in the manuscript is somewhat odd-did the authors mean Fig 2f, with red and brown bars?

© European Molecular Biology Organization

Re-submission 21 June 2013

We have revised our paper " Manuscript MSB-13-4411, "Human disease locus discovery through phylogenetic profiling". The changes are substantial actually, for example, adding a global mapping of all human genes into phylogenetic clusters and characterizing how each of these clusters intersects with HPO disease loci gene clusters and MSigDB molecular pathway clusters. We show that phylogenetic clusters are significantly enriched with genes that are associated with the same disease and biological function, emphasizing the forward predictive value of our method. Our current analysis were significantly extended and provides a much more systematic validation of the forward predictive value of phylogenetic clusters to identify new disease and pathway-related loci. In addition by mapping all human genes to phylogenetic profile clusters, we found that the genes in previously unrelated diseases and pathways are actually coevolved. The evolutionary connection suggests functional association between these diseases and pathways and can explain similar phenotype between diseases and crosstalk between pathways.
We also polished the informatic presentation so that the paper is readable by the general reader but scholarly for bioinformatians, mainly in the Supplementary Figures and Tables.
Below is a point by point response to the reviews.

We hope that the paper is now above the bar for MSB.

Response to reviews
Nearly all of the comments in the reviews were addressed in the revised version, and we believe that the changes improved the paper dramatically. While in the previous version we only demonstrated that many disease and functional gene groups are significantly co-evolved, In the current version we took our analysis to the next level and integrated the human disease database with the information obtained from the 6600 functional gene groups like KEGG, HPO, GO, co-expressed genes in cancer, TF and miRNA targets. We mapped all human genes to clusters of coevolved genes and found that the genes in these clusters have a particular biological function. Our analysis revealed many molecular pathways that map to the same phylogenetic clusters as genes associated with specific human diseases. Furthermore, we show that many genes that thought to map to different diseases are actually coevolved together and they mapped into the same phylogenetic clusters. Our analysis thus establishes a connectivity between different diseases and pathways, linking diseases phenotypes and functional gene groups.

Detailed replies to the editor and the referees' comments are below. The original comments are reproduced in italics

The editor comments:
1. "I also share the opinion of reviewer #1 that these examples appear rather 'anecdotal" ... "While it is interesting that 4-5% of the 3513 HPO groups include genes with correlated profiles, a more global characterization of the forward predictive value of the human phylogenetic clusters should be performed"... "the application of phylogenetic profiling to study systematically human disease genes could be a potentially interesting topic. We would thus not be opposed to consider a new submission on the topic provided the analysis can be significantly extended to provide a much more systematic validation of the forward predictive value of phylogenetic clusters to identify new disease-related loci."

Our response: In the revised manuscript we map the human genes into coevolved clusters established a more global characterization of the forward predictive value of the human phylogenetic. These phylogenetic clusters are significantly enriched with genes that are associated with the same disease and biological function, emphasizing the forward predictive value of our method. In the paper we demonstrated the power of our method to identify the connection between co-expressed genes, genes that are associated with diseases and pathways in addition to transcription and miRNA regulation. Our examples cover different biological topics including: cancer, mitochondrial and heme diseases, transcriptional regulation in order to emphasize the generality of our findings and their importance to the biomedical science community.

© European Molecular Biology Organization

2. "It is also possible that using large-scale genetic perturbation/expression profiles panels such as the new LINCS CMap (lincscloud.org) could provide a means to evaluate more systematically the biological significance of your phylogenetic analysis or to improve its predictive power using an integrative approach.

Our response: *Following private communication with two members of the LINCS group, we realized that the LINCS database currently suffers from several major technical and experimental problems that expected to be solved only in several months. Since the LINCS is still on it beta version and following the recommendation of member of the LINCS team, we decided not to integrate the LINCS database in our paper. Although LINCS expected to be useful in the future we believe that the extensive data analysis and integration (see paper) we generate in the current version is much more reliable at this point.*

Referee #1:
3. "Overall the result that genes involved in the same disease is not surprising; they should be functionally related and identifying such relationships is the idea of phylogenetic profiling. Making specific applications to diseases is also not novel. The authors themselves acknowledge (in the Results section) that "Phylogenetic profile analysis has been a powerful tool for the study of two types human disease, Bardet-Biedl syndrome (Mykytyn et al., 2004) and mitochondrial diseases (Pagliarini et al., 2008)."

Our response: *There is a debate in the field and several papers point out for the limited predictive power of the binary phylogenetic profile analysis in eukaryotes (Jothi et al., 2007; Loganantharaj and Atwi, 2007; Ruano-Rubio et al., 2009; Singh and Wall, 2008; Snitkin et al., 2006). To demonstrate that Bardet-Biedl syndrome and mitochondrial diseases are not the exception, we systematically analyzed more than 10,000 disease phenotypes and biological functional groups. Our finding that a wide variety of disease descriptors associated with a gene tend to cluster in a phylogenetic profile with other non-homologous genes also bearing the same descriptor will lead to better understanding of many diseases and acceleration in identification new genes implicated the same disease. We also believe that the current version of the paper demonstrates the importance of the phylogenetic analysis in revealing an additional level of interactions between genes, diseases, and biological pathways.*

4. This is the same kind of analysis as that in the Nature paper (though the methods are provided again for some reason). The analyses are very poorly described or not even described at all. For example, Figure 2 has six sections but only one is described. There are some actually good controls there - how could the authors completely ignore them in the text?
Our response: We removed the redundant methods part with the Nature paper and improved the description of our analysis in the results and the methods sections. Specifically we added more detailed explanation in figure 2.

5. The presentation of this work is extremely disjointed with two distinct parts that do not work well together.
Our response: The paper is indeed contains two parts and actually in the current version an additional global analysis module which is very short actually was added. The first two parts introduce a systematic phylogenetic profile analysis of diseases and pathways and the evolutionary crosstalk between them. In these parts we demonstrate the wide application of the phylogenetic analysis to different biological questions. Specifically we want our work to be observed outside the bioinformatics community, by biologists and medical researchers. We expect that the spreadsheets that highlight the 20% of MSigDB entries that have significant phylogenetic clustering and highlights the actual gene names in that overlap will influence prioritization of genes for analysis, especially as genome sequences emerge. But to demonstrate that the phylogenetic profiling can be used on a gene by gene basis, we show two examples of how it has already proven useful in the last section of the results. We have now made the transition to gene by gene analysis less jarring.

4. The reason why RBP-Jk was selected in order to query for a relationship with MITF is not clear. To make a real prediction, one would need to make a list of closely related profiles and then provide real rationale as to why one was singled out for detailed analyses. …., for the identification of the MITF-associated factor there was no evidence provided that RBP-Jk is also involved in cancer as much as MITF.

© European Molecular Biology Organization

Our response: We agree with the referee and actually RBPJ was chosen form the top 50 list after several filters. Obviously in the previous version, although we wrote all the details we massed and the story was less understandable. In the current version we better explain the steps that led us to chose RBPJ over the other candidates.

5. Finally the ccdc105 analysis suffers from the same lack of relationship to phylogenetic profiles and in particular to diseases.

Our response: *The ccdc105 example shows how phylogenetic profile analysis allowed a cofactor for meiotic gene function to be easily spotted. And this hypothesis was then tested easily, as shown in the paper.*

Reviewer #2:

1. The authors present an extension oh phylogenetic profiling to human proteins. By aligning human protein to 85 other eukaryotic proteomes, they are able to construct analog phylogenetic profiles, and show that clusters of these often overlap genes sets found in MSigDB. They also validate two of the functional inferences that they make using additional experimental data. Overall I found this to be a strong study that convincingly demonstrates the utility of phylogenetic profiles to assign functions to human proteins. My primary concern had to do with the difficulty of inferring profiles in organisms that have poorly annotated genomes. How can one be sure that the absence of a strong homolog is due to the fact that it is truly lacking from a genome, rather than incorrectly missing from the existing annotation. I assume this could be a significant issue for organisms with poorly annotated genomes. On the other hand, if a protein is missing across a set of related organisms then this is less likely to be due to misannotation. The authors should address this issue and discuss why it is not likely to have a large impact on their results.

Our response: We have added a paragraph discussing this issue (see method). In general we tried to maximize the coverage of species across the eukaryotic tree of life without reducing the quality by choosing poorly annotated genomes. For that we used several filters in order to remove poorly annotated genomes. From the available eukaryotic genomes, we filtered out the low coverage genomes and used highly annotated and high coverage genomes obtained from Ensembl. Ensembl includes both automatic annotation, i.e. genome-wide determination of transcripts, and manual curated, i.e. reviewed determination of transcripts on a case-by-case basis. Since Ensembl has a limited number of fungi and protists, 33 additional high quality genomes (with coverage of X8) form NCBI genome database were added to supplement the analysis. As another quality control, we calculated the correlation of the protein conservation between species. We removed several species that showed low correlation with their closely related species like Nasonia vitripennis or Equus caballus since low correlation might reflect problem in the genome assembly. In addition following referee #2 comment, we revised our data and found two species (Dasypus novemcinctus and Felis catus) that show slightly lower correlation to their closely related species. To test the impact of this species on our analysis we ran the MSigBD analysis, looking at the p-value distribution after removing each these two species. While the effect of removing these species on the p-values was minor, it significantly decreased the performance of the method (in general the MSigBD groups got worse p-values and lower Co-10 scores excluding these two species). These emphasizes two things: 1. The effect of each organism is low. 2. Although Dasypus novemcinctus and Felis catus are probably have the poorest annotation of the species in our database, they still improve the performance of the phylogenetic profiles. In addition since we used close to 90 species, an error in one or two samples should not a have major effect on the coevolution correlation between genes. Finally signal clustered significantly better than noise and as such there is low chance of having false positive i.e. more noise decreases the chance to find coevolved clusters and that those clusters will be over represented with functional groups.

2. It would also be useful to see additional benchmarking to convince the reader that the algorithm choices that were made (i.e. nature of value within profiles, and metric for profile comparisons) are indeed optimal or close to optimal.

Our response: While it not clears to us how to define the optimality of the method in this revision; we expanded the discussion of the poor performance of the binary method compared to the

© European Molecular Biology Organization

normalized phylogenetic profile. In addition we show the importance of using different numbers of genomes to the prediction power of the methods. In addition while the normalized phylogenetic profiles identifies a large set of the gene functional groups to be significantly coevolved (i.e. more then 50% of the KEGG pathways), most of the previous methods used over the years shows limited success in eukaryotes (Jothi et al., 2007; Loganantharaj and Atwi, 2007; Ruano-Rubio et al., 2009; Singh and Wall, 2008; Snitkin et al., 2006). In prokaryotes, normalizing the phylogentic profile data can dramatically improve the performance (Enault et al., 2003) . Importantly we believe that extensive computational study could be performed to systematically examine a wide variety of phylogenetic methods in eukaryotes (Date and Peregrin-Alvarez, 2008; Kensche et al., 2008).

Reviewer #3
I have one concern, which may have already been addressed. The concern is associated with the interesting yet unexpected finding that genes with common regulatory motifs, such as many miRNA targets, are co-evolved. I wonder whether this finding could be an artifact of the algorithm. Specifically, if we consider a family of proteins that evolves from a single protein in a more primitive organism, this family will have a highly correlated phylogenetic profile. This leads to the question whether the enrichment, of co-evolved genes in miRNA targets or those share common promoter motifs, was driven by a single or a small number of protein families. If so, this provides a possible explanation for common regulatory motifs in such gene sets. If not, the authors' suggested model of independent acquisition of regulatory motif may be possible. I noticed that in the method, the authors effectively eliminated contribution to correlation scores by highly homologous genes in human, but I am not
sure whether this elimination is "clean" with regard to the issue discussed above.

Our response: miRNA complementary sites as TF binding sites are rarely conserved outside of mammalians and as such it is less likely that the motif enrichment found in coevolved gene results from duplication in a protein primordial ancestor. Saying that, in cases of very recent gene duplication the bias suggested by the referee might happen. In order to eliminated the homology effect we calculated the Co10 after removing paralogous, possibly duplicated, genes. Finally in our current version we calculate the hyper-geometric p-values between coevolved clusters and the MSigDB groups using homologous gene groups instated of a single genes (e.g. A group of 10 proteins, 5 of which have sequence similarity, would be considered to have a size of 6). Similarly, if the overlap between the functional gene group (HPO or MSigDB) and coevolved protein cluster contains homologous proteins, we considered only one protein per homologue family. Proteins are considered to be in the same family if they have blast score >100 or sequence identity larger than 10%. Our results suggest that promoter and UTR motifs present a simple mechanism by which an organism can fine tune sets of genes that are coevolved together across millions of years.

Minor:
1. There are two figure 2e.
Fixed
2. There is no page number in the manuscript.
Changes made.
3. The reference to supplemental figure 6a, 6b is incorrect-should be sup fig 4?
Yes, change made.
4. The reference to figure 2e in the manuscript is somewhat odd-did the authors mean Fig 2f, with red and brown bars?
Yes. Change made.

2nd Editorial Decision 28 June 2013

Thank you for having submitted a revised version of the manuscript entitled "Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling" for consideration for publication in Molecular Systems Biology.

Your paper has now been seen by Editors of the Journal, and we have decided to return it to you without sending it for extensive peer review.

© European Molecular Biology Organization

We do appreciate that you included a further analysis of phyolgenetic clusters and describe now, for example, that 54 out of 1076 clusters are enriched in genes associated to an HPO group ('PhyloDisease' clusters), 48 of which are also enriched in genes from the MSigDB gene sets. We recognize that this analysis is complementary to your previous results (ie significant co-evolution score for 340 out of 6600 MSigDB gene signatures and 156 out of 3,413 HPO classifications). However, we are not convinced that the new results would fundamentally extend and strengthen the major conclusions or the impact of the study. We nevertheless also asked our Senior Editors for further advice on this study and I am afraid that the recommendation was not to send the manuscript for a new round of review. I am very sorry not to be able to bring better news on this occasion, but I hope that this early decision will allow you to submit your work elsewhere without undue delay.

Authors' response 02 July 2013

I am writing to disagree with your handling of our resubmission, "Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling". We had read the April 17 reviews of our paper as really very positive for two of the three reviews and were therefore shocked when you decided to not send it back to the reviewers. Yes one of the reviews was negative, but two out of three reviews were positive, which you mentioned in your communiqué of April 17 (*"Thus, while reviewer #2 and #3 provide supportive recommendations, reviewer #1 raises significant concerns and is much less supportive"*).

Reviewer 2: *"Overall I found this to be a strong study that convincingly demonstrates the utility of phylogenetic profiles to assign functions to human proteins".*

Reviewer 3: *"Overall, this is a very well done piece of work that will be of great interest to scientists in many fields."*

In addition to these positive responses to our work, the negative reviewer gave very few comments. As such we mainly improved our paper based on the editor (Lemberger) comments:

"...a more global characterization of the forward predictive value of the human phylogenetic clusters should be performed"

" ... deeper integration with other resources to characterize the power of continuous phylogenetic profiling for human disease locus discovery"

"...We would thus not be opposed to consider a new submission... provide a much more systematic validation of the forward predictive value of phylogenetic clusters to identify new disease-related loci."

We followed the editor comments and generated and validated the forward predictive value of phylogenetic clusters. As such the editor decision to reject our resubmission without rereview is inconsistent with the reviews from the first round and from the fact that we followed the editor suggestion and improved the manuscript exactly as requested.

3rd Editorial Decision 07 August 2013

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from referee #1 whom we asked to evaluate your revised study. As you will see, this referee is satisfied with the changes made and is now supportive. We are pleased to inform you that we will be able to accept your study for publication pending the following minor points:

1. Supplementary information: please prepare supplementary information according to our instructions at http://www.nature.com/msb/authors/index.html#a3.4.6 (essentially, supplementary information should be provided as a single PDF files starting with a Table of Content and large tables should be submitted as Excel or csv or tab-delimited files).

© European Molecular Biology Organization

Thank you for submitting this paper to Molecular Systems Biology.

Referee reports:

Reviewer #1:

In this revision, Tabach et al answer all of my comments. I am swayed by their argument that this paper performs phylogenetic profiling for the first time for disease pathways. I appreciate that the revision includes descriptions of the figures, of additional analyses, and of explanation of why particular validation experiments were done. I would support publication of this revision.

1st Revision - authors' response 24 August 2013

We would like to thank you for sending our paper for re-review and accepting the paper "Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling"

The supplementary information was prepared according to MSB instructions and uploaded as PDF.

© European Molecular Biology Organization