molecular systems biology

Protein localization as a principal feature of the etiology and comorbidity of genetic diseases

Solip Park, Jae-Seong Yang, Young-Eun Shin, Juyong Park, Sung Key Jang, Sanguk Kim

Corresponding author: Sanguk Kim, Pohang University of Science and Technology

Review timeline:

Submission date:	10 November 2010
Editorial Decision:	23 December 2010
Revision received:	01 March 2011
Accepted:	19 April 2011

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 23 December 2010

Thank you again for submitting your work to Molecular Systems Biology. 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 referees find the topic of your study of potential interest. However, they raise some concerns on your work, which should be convincingly addressed in a revision of the present manuscript.

The major concerns raised by the reviewers refer to the following issues:

- the need to extend the study using HGMD database and associations on non-mendelian diseases

- while the ICD9 mapping used here has been published before, it appears that it would nevertheless be crucial to verify the quality of this mapping, possibly by comparison with other mapping tools.

We would also kindly ask you to include in supplementary information the key datasets that result from your computational analysis (for example the DPL matrix, list of co-morbid disease pairs) so that others can easily reproduce and build upon your work.

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors. If you have any questions about this initiative, please contact the editorial office msb@embo.org.

Yours sincerely,

Editor

Molecular Systems Biology

REFEREE REPORTS

Reviewer #1 (Remarks to the Author):

Kim et al systematically constructs the subcellular localization profiles for diverse human diseases for the first time and demonstrates the subcellular-localization specificity for different diseases. Moreover, the authors show that the proteins involved in the same diseases or phenotypically similar diseases tend to be connected by the common subcellular localization. More interestingly, the authors demonstrate that subcellular localization could explain additional comorbidity of disease pairs for diseases sharing no common known disease proteins or with no known protein-protein interactions. All this highlights the importance of subcellular localization for understanding disease mechanisms. The methods are sound and the results are well demonstrated. The manuscript should be understandable by a broad audience.

Mimor: Given that genes implicated in the same or related diseases tend to be co-localized, the authors shoul be able to demonstrate or discuss how subcellular localization can be used to improve our understanding the mechanisms of specific diseases or to help identify novel disease genes. Examples would be useful.

Reviewer #2 (Remarks to the Author):

The authors present a computational analysis of the phenotypic similarity and co morbidity of diseases sharing the same subcellular localization. The authors apply a broad range of computational analyses to this question. The breadth of the analyses presented here is impressing, the methods seem overall robust, and the results are interesting. However, there is no experimental follow up on any novel discoveries, and no clear application besides a descriptive analysis. This perhaps makes the paper more suitable for a computationally oriented journal such as PLoS computational biology, where I think this paper would be a better fit.

Major points.

1) There are many examples of sentences that are hard to understand, which significantly reduces my ability to follow the points raised by the authors throughout the text. For instance the sentence in the abstract "The spatial constraints from subcellular localization significantly strengthened the disease associations among the proteins that share subcellular localizations", is an example of a sentence which does not read well. Another example on page 8: "Next, to investigate the coverage requirement of disease-subcellular localization associations", is not very clear.

2) In OMIM, there are many MIM records that point to the same disease. E.g., Fanconi Anemia is mentioned in tens of different MIM files. Does this have an effect on the DPL matrix? The authors mention that they combine disease subtypes into single diseases and refer to the Goh et al PNAS 2008 paper, but I think some details about this procedure needs to be described in the main text or methods so the readers can understand this critical point, and how this could potentially confound their analysis.

3) Many of the subcellular localization annotations in Swiss Prot database are inferred from orthologous genes in other organisms. I think there needs to be some discussion on the reliability of these annotations, and how it affects the analyses presented here. The authors could consider doing some sort of test of the annotations. For example one could ask how many of the mitochondrial proteins (Pagliarini, Calvo et al., Cell 2008) have mitochondrial annotations in Swiss Prot? Although this might be a little circular because the annotations could be influenced by the publication it would be a nice sanity check. If something like this has already been done in previous work it should be reiterated briefly in this paper.

4) Is there some sort of independent validation of the scores in the DPL matrix? It is not obvious that these scores truly capture which diseases are caused by molecular defects in which organelles, and some independent validation of the association score reliability is needed in my opinion.

Minor points:

1) Figures should be numbered.

2) On page 5: "We constructed for the first time the matrix of disease-associated proteins and their subcellular localizations." This should be "a matrix of diseases-associated.....".

3) Methods section: How was low confidence interactions in the ppi data filtered out?

4) In Box 1 the cellular localization names are halfway hidden and difficult to read.

5) What is on the x-axis of supplementary figure 4 A?

Reviewer #3 (Remarks to the Author):

The authors present a systematic analysis of disease-associated genes incorporating sub-cellular localization information for associated proteins to evaluate the relationship between subcellular localization and characteristics disease genes and their associated pathologies. Although the molecular pathology of many diseases have been characterized as having pathophysiological phenomena linked to specific aspects of subcellular localization, it appears that the work presented in this manuscript represents the first effort to perform a systematic evaluation of these relationships across a compendia of diseases. The results and conclusions should provide valuable insight to a broad audience of biomedical researchers, and the data and methods described would likely be valuable in future, large-scale, integrative studies of disease mechanisms. Integration of comorbidity data from Medicare patients works favorably to elevate the translational value of the findings.

Overall the manuscript is well-written, and the length seems appropriate to convey the necessary information regarding the study and results. Nonetheless, I do have some concerns about specific aspects of the manuscript that I suggest the authors consider to help improve the clarity and impact of their findings.

- The current title of the manuscript seems a bit bland ambiguous to me, and does not seem to reflect the overall message conveyed by the findings as well as it could. Specifically, it's not clear how this manuscript relates to disease "profiling", which, in the context of molecular biology, seems to evoke a sense of measurement (e.g. microarray analysis) which is not a principal aspect of this manuscript. Perhaps something like, "Protein subcellular localization is a principle feature of etiology and comorbidity of genetic disease". The present title only alludes to the actual content of the manuscript.

- I do not believe that the authors are correct in stating that, "... OMIM is the most complete database available for disease-gene associations ...". It is the experience of this reviewer that the Human Gene Mutation Database HGMD contains much more complete and rich information regarding disease-gene association as compared to OMIM. It's not clear why HGMD was not considered.

- OMIM is restricted to Mendelian disease mutations, which are important, but it's not clear why the authors chose to regard disease genes associated with common, complex diseases such as Type 2 diabetes. Many of these associations are available through dbGAP, HGMD, and the NHGRI GWAS catalog. It would be interesting to know if the subcellular localization enrichments also hold true for non-Mendelian disease mutations.

- The integration and comparison of PPI data is a very nice addition to the manuscript, but another important functional association between genes is co-expression (see Wolfe et al. BMC Bioinformatics 6 p227 (2005 Sep 14)). There is a tremendous amount of gene expression data available in the public domain, so it would be interesting to see how information from subcellular localization compared to co-expression. It is reasonable to believe that disease genes that co-locate subcellularly might also exhibit significant co-expression. Since microarray analysis is now so abundant and inexpensive, I would like to see how much more information is gained from the addition of subcellular localization characteristics.

- The authors state that the ICD-9-CM codes from the Medicare claims data were mapped to OMIM disease IDs manually, however the process is not described in detail, and the nature of this process is likely to significantly affect the enrichment analysis. Such mapping is non-trivial and needs careful consideration for the nature of how the codes are applied. For example, it is well known that ICD-9-CM codes are usually applied to patient records in a manner that optimizes financial reimbursement rather than accurate clinical annotation of the disease phenotype. It's not clear why the authors chose to ignore the available tools and data sources in the freely-available Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/) that would enable them to map ICD-9-CM codes to OMIM diseases in a systematic, unbiased manner.

- I think this manuscript would benefit from a figure showing a schematic workflow of the overall informatics approach taken that gives a high level overview of the relationships between data and methods.

- Fig. 3A could benefit from some additional visual elements to show how the patient population ties into the Relative Risk score.

- Fig. 3B might look better using simple points with error bars rather than a barplot, which adds clutter and makes it difficult to discern the trend.

- It seems that Fig. 3C-E could be merged into a single figure. 3C-D on their own don't seem to convey much information relative to the amount of visual space they consume

- It would be great to see some statistical significance annotations between the groups shown in Fig. 3F-G

- "Nucleus" is misspelled in Fig. 2B

If the above concerns are addressed satisfactorily, I would highly recommend publication of the manuscript in Molecular Systems Biology as the impact, nature, and scope of the article are commensurate with those published by MSB and similar caliber journals.

1st Revision - authors' response 01 March 2011

We would like to express our sincere thanks to you for your kind consideration of our manuscript. We are also deeply grateful to the Reviewers for their remarks and constructive suggestions. We now have a detailed response to the Reviewers' comments, and a summary of updates to the manuscript prompted by their suggestions. We believe that they have helped us improve our manuscript greatly, and hope that you find our manuscript now ready for publication.

Reply to Reviewers' Comments

Editor's comments

We would kindly ask you to include in supplementary information the key datasets that result from your computational analysis (for example the DPL matrix, list of co-morbid disease pairs) so that others can easily reproduce and build upon your work.

This is an excellent suggestion, and we now added the datasets (subcellular localization information of disease-associated genes, DPL matrix, non-mendelian disease-associated genes, and list of co-morbid disease pairs) to Supplementary Files 1-6.

Reviewer 1

Kim et al systematically constructs the subcellular localization profiles for diverse human diseases for the first time and demonstrates the subcellular-localization specificity for different diseases. [...] All this highlights the importance of subcellular localization for understanding disease mechanisms. The methods are sound and the results are well demonstrated. The manuscript should be understandable by a broad audience.

We are grateful to the Reviewer for their positive assessment of our work. Below we present our responses to the Reviewer's comments and updates to the manuscript, which we hope the Reviewer finds satisfactory.

1) Given that genes implicated in the same or related diseases tend to be co-localized, the authors should be able to demonstrate or discuss how subcellular localization can be used to improve our understanding the mechanisms of specific diseases or to help identify novel disease genes. Examples would be useful.

We fully agree with the Reviewer that demonstrations and examples would be very helpful. Therefore, we now illustrate several cases of newly found disease-associated genes that share the same localization in the Discussion Section (pages 14-15). Also, in Supplementary Figure 8, we show disease modules representing the clusters of interacting proteins that are connected via subcellular localizations and share disease annotations: one example therein is the disease module of Cerebral Degeneration comprising eight mitochondrial proteins among which five were previously known to be involved in the disease. We expect that the newly found proteins are potentially associated with Cerebral Degeneration, since they are connected via the same localization, and interact with proteins associated with the same disease.

Reviewer 2

The authors present a computational analysis of the phenotypic similarity and co morbidity of diseases sharing the same subcellular localization. The authors apply a broad range of computational analyses to this question. The breadth of the analyses presented here is impressing, the methods seem overall robust, and the results are interesting.

We are grateful to the Reviewer for their positive assessment of our work. Below we present our responses to the Reviewer's comments and updates to the manuscript, which we hope the Reviewer finds satisfactory.

Major points:

1) There are many examples of sentences that are hard to understand, which significantly reduces my ability to follow the points raised by the authors throughout the text.

Prompted by the critism, we have made corrections where necessary to the best of our abilities, and we hope that the Reviewer finds the revised manuscript more comprehensible.

2) In OMIM, there are many MIM records that point to the same disease. E.g., Fanconi Anemia is mentioned in tens of different MIM files. Does this have an effect on the DPL matrix? The authors mention that they combine disease subtypes into single diseases and refer to the Goh et al PNAS 2008 paper, but I think some details about this procedure needs to be described in the main text or methods so the readers can understand this critical point, and how this could potentially confound their analysis.

We agree with the Reviewer that a detailed explanation of the procedures would be extremely helpful. Therefore, we have added the details about our procedures in the Result and Methods section (page 6 and 19).

Prompted by the suggestion, we have also analyzed the effect of combining disease subtypes into single diseases on the DPL matrix: We found that the disease subtypes are also enriched in the same subcellular localizations on the DPL matrix as we have observed previously from the analysis of single diseases (Supplementary Figure 10), suggesting that disease subtypes tend to share their subcellular localizations. For instance, Fanconi anemia subtypes are mostly enriched in the nucleus, whereas Complement deficiency subtypes are enriched in the extracellular region. This finding has been added to the Discussions Section with further details (page 17).

3) Many of the subcellular localization annotations in Swiss Prot database are inferred from orthologous genes in other organisms. I think there needs to be some discussion on the reliability of these annotations, and how it affects the analyses presented here. The authors could consider doing some sort of test of the annotations. For example one could ask how many of the mitochondrial proteins (Pagliarini, Calvo et al., Cell 2008) have mitochondrial annotations in Swiss Prot? If something like this has already been done in previous work it should be reiterated briefly in this paper.

As suggested by the Reviewer, we have now added new analyses based on the comprehensive mitochondria database, MitoCarta (Pagliarini, Calvo et al., Cell 2008), to the Discussion Section (page 16) and the Supplementary File 4: First of all, we have confirmed that 67% of the MitoCarta proteins were annotated as mitochondria in Swiss Prot (Supplementary Figure 9A). A small number of the proteins were tagged ìlocalization unknownî or annotated as other subcellular localizations such as cytosol or nucleus.

As the Reviewer points out, we fully acknowledge that proper subcellular localization annotation is a key ingredient of the analysis presented here. In fact, we have in the past developed a consensus localization prediction method called ConLoC (Park S, JPR, 2009), which we have applied to our current study. ConLoc predicts protein subcellular localization based on optimization of the prediction results from thirteen well-known localization prediction programs, and achieves the highest prediction accuracy of 0.96 and Matthew's correlation coefficient (MCC) of 0.86 on the localization prediction of human proteins. Thus we reported that ConLoc outperforms all the individual predictor and shows the highest sensitivity on the independent test set of 345 mitochondrial proteins. Moreover, ConLoc archives the equivalent accuracy on the prediction of multi-localized proteins to that of single-localized proteins. Now this point is further described in detail in the manuscript (page 20).

Further prompted by the Reviewer's comments, we performed a test of mitochondrial localization by using three different subcellular localization annotation sets which include Swiss Prot annotation, ConLoc, and comprehensive localization annotations from MitoCarta. We observed that MitoCarta covered more associated diseases and showed higher correlations (PCC) between subcellular localization similarity and comorbidity tendency (Supplementary Figure 9B). While MitoCarta gave a somewhat higher correlation (PCC = 0.86), the present method of applying ConLoc showed a comparable coverage of associated diseases and correlation (PCC = 0.83). As we see here, given that there exist several compatible yet slightly distinct schemes and databases it is important to note that not one can be (if ever) declared absolute best; rather, they present opportunities for comparison and cross-checking one's research, which we believe helps up to demonstrate the robustness of our study.

4) Is there some sort of independent validation of the scores in the DPL matrix? It is not obvious that these scores truly capture which diseases are caused by molecular defects in which organelles, and some independent validation of the association score reliability is needed in my opinion.

We again agree with the Reviewer. Therefore, we have now added a new analysis for the validation of the association score reliability to the Results Section (page 7). There we present the Z-values of the subcellular localization-disease association scores (Supplementary Figure 1A). The Z-values represent the significance of the subcellular localization enrichment of diseases. We have also added the details of the procedure in the Method section (page 21). We observed that Z-value and subcellular localization-disease association score are highly correlated (R-square = 0.97), and an association score 0.05 was considered to be statistically significant (P < 0.01). Furthermore, diseases caused by molecular defects in specific organelles showed significant association scores (association score 0.2, Z-value > 10, P < 1.00 x 10-10) (Supplementary Figure 1B). For example Mitochondrial Complex I-III deficiency, a mitochondrial disease, showed statistically significant enrichment in the mitochondria (Z-value = 10.6, P < 1.00 x 10-10).

Minor points:

1) Figures should be numbered.

Now figures are correctly numbered.

2) On page 5: "We constructed for the first time the matrix of disease-associated proteins and their subcellular localizations." This should be "a matrix of diseases-associated.....".

The sentence has been corrected.

3) Methods section: How was low confidence interactions in the ppi data filtered out?

Protein interactions were excluded from high-throughput methods, orthologous interactions from lower organisms than human, or as predicted by in silico methods. We now describe in the Methods section (page 23) the process of filtering out low confidence interactions from PPI data.

4) In Box 1 the cellular localization names are halfway hidden and difficult to read.

The issue has been corrected.

5) What is on the x-axis of supplementary figure 4 A?

The x-axis now clearly shows ëSubcellular localization PCC'.

Reviewer 3

The authors present a systematic analysis of disease-associated genes incorporating sub-cellular localization information for associated proteins to evaluate the relationship between subcellular localization and characteristics disease genes and their associated pathologies. [Ö] Overall the manuscript is well-written, and the length seems appropriate to convey the necessary information regarding the study and results.

We are grateful to the Reviewer for their positive assessment of our work. Below we present our responses to the Reviewer's comments and updates to the manuscript which we hope the Reviewer finds satisfactory.

1) The current title of the manuscript seems a bit bland ambiguous to me, and does not seem to reflect the overall message conveyed by the findings as well as it could. The present title only alludes to the actual content of the manuscript.

We fully agree with the Reviewer, and the manuscript now sports a new title ìProtein localizations as a principal feature of the etiology and comorbidity of genetic diseases,î which we believe better captures the intents of the contents of our work.

2) I do not believe that the authors are correct in stating that, "... OMIM is the most complete database available for disease-gene associations ...". It is the experience of this reviewer that the Human Gene Mutation Database HGMD contains much more complete and rich information regarding disease-gene association as compared to OMIM. It's not clear why HGMD was not considered. OMIM is restricted to Mendelian disease mutations, which are important, but it's not clear why the authors chose to regard disease genes associated with common, complex diseases such as Type 2 diabetes. Many of these associations are available through dbGAP, HGMD, and the NHGRI GWAS catalog. It would be interesting to know if the subcellular localization enrichments also hold true for non-Mendelian disease mutations.

We fully agree with the Reviewer that, in light of the existence of multiple databases focusing on distinct groups of important human diseases, applying any statistical methodology to separate databases would be very important. Following the Reviewer's suggestion we have deleted the sentence pointed out and added an analysis of non-mendelian diseases. Although we have tried our best to access the HGMD, unfortunately it was not available for academic users to large-scale analysis as we have done for our original manuscript at the moment. While we note that the HGMD is definitely worth analyzing in the near future, at this point in order to understanding non-mendelian complex diseases we used the Gene Association Database (GAD, Becker KG, Nat Genet 36, 2004): based on 427 GAD diseases, we reconstructed the matrix of disease-associated proteins and their subcelluar localizations. From the matrix, we observed that proteins associated with non-mendelian diseases from GAD also show subcellular localization enrichments as we observed from mendelian diseases from OMIM (Supplementary Figure 11A). For example, the proteins associated with Bipolar Disorder, a complex disease, are enriched in the cytosol, whereas the proteins associated with Type-2 Diabetes are mostly enriched in the plasma membrane (Supplementary Figure 11B). Interestingly, some proteins associated with Type-2 Diabetes are also enriched in the cytosol compartment which is functionally related to the plasma membrane. We present a detailed analysis in the Discussion Section (page 17) and the Methods section (page 24). We also provided the disease-gene association from GAD in Supplementary File 5.

3) The integration and comparison of PPI data is a very nice addition to the manuscript, but another important functional association between genes is co-expression (see Wolfe et al. BMC Bioinformatics 6 p227 (2005 Sep 14)). It is reasonable to believe that disease genes that co-locate subcellularly might also exhibit significant co-expression. I would like to see how much more information is gained from the addition of subcellular localization characteristics.

We fully agree with the Reviewer that co-expression is an important factor in disease comorbidity. We indeed confirm that when subcellular localization information is combined with co-expression, comorbidity tendency increases (Supplementary Figure 7). The detailed analyses are presented in the Results section on page 14 and the Method section on page 23.

4) It's not clear why the authors chose to ignore the available tools and data sources in the freely-available Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/) that would enable them to map ICD-9-CM codes to OMIM diseases in a systematic, unbiased manner.

We would like to thank the Reviewer for bringing UMLS to our attention, and we agree that UMLS provides us with another opportunity for a systematic mapping. Prompted by the Reviewer's remarks we have performed our analysis using a mapping created from UMLS, which we present in the revised manuscript and present as a part of our reply to this point.

First of all, though, we believe a somewhat detailed discussion of the origin of the manual mapping we used would be helpful in clarifying the issue: The mapping was originally commissioned by Dr Nicholas A. Christakis, a former collaborator of one of us (J. Park), of Harvard Medical School and School of Public Health to professional coders at his University Hospital. As one of the very first attempts to bridge the field of molecular biology and population-level hospitalization statistics, they determined that a reasonable way to produce the mapping was to hire professional record keepers at a leading institution with many years of experience in working with actively practicing medical doctors, given that the Medicare database itself is a hospitalization record. In the sense that they were clearly one of the most experienced groups in connecting medical diagnoses (made by practicing medical doctors) and ICD-9-CM codes, we believe that they still hold validity. Also, the same mapping was used for several recent publications on disease comorbidity (most notably Lee et al., PNAS 2008 and Park et al., MSB 2009) in well-respected academic journals, demonstrating their utility and robustness. This is, we believe, highlights the challenges and the benefits of the type of our research as ours: As more databases are built and come to light in medical and biological sciences (just as the UMLS was brought to our attention by our kind Reviewer) more opportunities for cross-checking the databases using systematic, sophisticated methodologies will inevitably open up, and we are wholeheartedly open to the challenges to which our current work is a contribution.

In this spirit of improving our research based on the Reviewer's suggestion, therefore, we have performed an identical analyses using a mapping based on UMLS which we discuss here (also added in our manuscript): When we apply the UMLS-based mapping, we again observe that disease pairs connected via subcellular localizations show higher comorbidity than average over all disease pairs (Supplementary Figure 12). Furthermore, we also observe that comorbidity increases when subcellular localization information is combined with shorter network distances. Lastly, we would like to mention that, understandably, there may exist subtle disagreements between the two mappings; For instance, in the case of ìAchondroplasia (MIM ID: 100800)î the human experts of the original mapping chose to utilize 733.9 in ICD-9-CM while the UMLS resulted in it being mapped to 756.4 in ICD-9-CM. Most importantly, though, we observe the aforementioned similarity in the trends of our analyses based on the two mappings, and that we believe that they strongly indicate the robustness of our conclusions. Again, we hope that the Reviewer kindly understands that as we go forward in this type of research we are keenly aware of the issues and research opportunities in dealing with developments in molecular systems research and that to the best of our knowledge that we wish to have made a valuable contribution.

5) I think this manuscript would benefit from a figure showing a schematic workflow of the overall informatics approach taken that gives a high level overview of the relationships between data and methods.

We fully agree with the Reviewer, and we now provide a flowchart explaining the overall informatics approaches as Supplementary Figure 5.

6-1) Fig. 3A could benefit from some additional visual elements to show how the patient population ties into the Relative Risk score. Fig. 3B might look better using simple points with error bars rather than a barplot, which adds clutter and makes it difficult to discern the trend.

As suggested by the Reviewer, we have now updated Figure 3A and 3C with more clear examples of comorbid disease pairs with patient population ties into the Relative Risk score. We have also modified Figure 3B and Supplementary Figure 6 accordingly.

6-2) It seems that Fig. 3C-E could be merged into a single figure. 3C-D on their own don't seem to convey much information relative to the amount of visual space they consume.

We have now merged Figure 3C and 3D into a single figure.

6-3) It would be great to see some statistical significance annotations between the groups shown in Fig. 3F-G

We have now added the statistical significance annotations between the groups in revised Figure 3E and 3F.

6-5) "Nucleus" is misspelled in Fig. 2B

We have corrected the spelling.

This concludes our response to the Reviewers' comments. We believe they have helped us improve our work considerably. We hope that they now find our manuscript acceptable for publication.

Acceptance letter 19 April 2011

Thank you again for sending us your revised manuscript. We are now satisfied with the modifications made and I am pleased to inform you that your paper has been accepted for publication.

NOTE Reviewer #2 is asking for some additional clarification on how diseases were merged. Please address this with the suitable amendment in the Materials and Methods section and send us the revised word file directly by email.

Proofs will be forwarded to you within the next 2-3 weeks.

Thank you very much for submitting your work to Molecular Systems Biology.

Sincerely,

Editor

Molecular Systems Biology

Reviewer #2 (Remarks to the Author):

I believe that the authors have done a good and thorough job of clarifying and extending their analyses to comply with the issues raised by this reviewer. There is only one minor point which I would like further clarify before publication.

1) Although the authors have extended the section on OMIM, and provided more text, they do still not explain how diseases are merged. They mention that eleven Fanconi anemia files are grouped into a single disease id 523, but how is this done? Manual curation, text-mining, some other method?

Reviewer #3 (Remarks to the Author):

I am very satisfied with the changes made by Park et al. to the manuscript in response to reviewer concerns. The paper has greatly improved since the first version, and I believe that it is now highly relevant and well-suited for the MSB readership.