EMBO reports

Manuscript EMBOR-2014-39403

Conserved Abundance and Topological Features in Chromatin Remodeling Protein Interaction Networks

Mihaela E. Sardiu, Joshua M. Gilmore, Brad D. Groppe, Damir Herman, Sreenivasa R. Ramisetty, Yong Cai, Jingji Jin, Ronald C. Conaway, Joan W. Conaway, Laurence Florens and Michael P. Washburn

Corresponding author: Michael P. Washburn, Stowers Institute for Medical Research

Review timeline:

Transfer date:	01 August 2014
Editorial Decision:	04 August 2014
Revision received:	07 October 2014
Accepted:	27 October 2014

Editor: Barbara Pauly

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.)

Transfer Note:

Please note that this manuscript was originally submitted to Molecular Systems Biology, where it was peer-reviewed. It was then transferred to EMBO reports with the original referees' comments attached. (Please see below)

Original referees' comments – Molecular Systems Biology

Reviewer #1:

Summary and General Remarks

The present study, titled 'Conserved Abundance and Topological Features in Chromatin Remodeling Protein Interaction Networks,' presents a compelling story regarding the conservation of abundance among orthologous proteins involved in chromatin remodeling complexes between yeast and human. Through a quantitative proteomics and analytical framework, the authors suggest that yeast-human orthologs involved in three different chromatin remodeling complexes retain relative

abundances between themselves, but not between members of the other complexes. Further, through analysis of the complex interaction network topology, they show conservation of connections among not only chromatin remodeling, but its relation to transcriptional regulation and RNA processing, among others. They further suggest a method for predicting missing abundances by inferring them from another species, a potentially useful method for the proteomics community.

The conclusions presented in the study are generally well supported. While the studies findings of conservation of orthologous protein abundances has been shown before among many species (as cited by the authors), previous studies have used various previously published datasets and it has thus been difficult to control for variation based on preparation or equipment differences. The authors base their findings on an in-house proteomics dataset, thus controlling for some of these variables. They also present a novel, though not unexpected, conservation of network topology related to chromatin remodeling and other nucleic acid regulatory functions. The study should be relevant to many systems biologists, particularly those in the proteomics and epigenetic fields.

Major points

1. The authors show a high correlation of abundance values between orthologous complex members, while previous studies have already suggested that orthologous proteins in general have highly correlated abundances. It would therefore be quite interesting to show that the correlation of orthologs in complexes is even higher than that of the general protein abundance conservation between orthologs. Related to this, do the authors anticipate that significant bias is being introduced in the abundance measurements by using baited proteins as opposed to bait free proteomics?

2. It is not clear how the opposite species values are used to infer missing values for the current species. SVDimpute requires complete matrices to start, and several methods exist for assigning initial values to missing points. Are the authors using these? Or is it simply using data from one species to initialize the missing values in the species being tested? If the latter (or former, for that matter), it is totally unclear that this is the case. Further, there is no explanation of why SVDimpute was chosen over the existing methods, such as row-averaging or K-nearest neighbors. I would thus like to see some evaluation of these other methods and comparison against SVDimpute, and also comparison with SVDimpute using different initial values.

3. The section regarding the discovery of a conserved low-abundance subnetwork is somewhat unclear. While interesting, there is no mention of abundance, low or high, anywhere apart from the section heading and the final concluding sentence, and so it is unclear what exactly is meant by 'low-abundance' or why that feature would be relevant or novel. Additionally, Table E5 is mentioned as containing the final microarray filtered list of ortholog pairs, however only the 214 pairs from before filtering are apparent. It is also unclear what groups of proteins are being overlapped in the hypergeometric analysis. Clarification of this should be included in the text, Fig. 7 legend, and methods.

Minor points

1. There are discrepancies throughout the manuscript as to the number of orthologs studied as baits. The abstract mentions 18, the beginning of the results sections mentions 15, and Figure 1 and elsewhere in the text mention 16. Further, the first paragraph of the results mentions "three orthologous proteins between the human TIP60 and yeast NuA4...", while there are actually four labelled in Figure 1.

2. It should be mentioned in the main text regarding use of the Spearman correlation that Spearman is a correlation of ranks, and not of actual values. Perhaps by specifically referring to it as such (i.e. "we computed the Spearman rank correlation") or just mentioning it in the explanation given for what correlation is. Spearman is probably the correct metric here, but would just like to see that detail mentioned in the main text. On a related note, the section describing what correlation values are and what they mean is probably unnecessary for the main text.

3. In Figure 5A, for the gene ACTR5, a k-value of 2 is reported. The text mentions that only k

values from 3-8 were tested. This discrepancy should be clarified.

4. There are several typos and grammatical errors involving sentence structure, etc. throughout. A very thorough reading of the manuscript for grammatical and spelling correctness is suggested.

Reviewer #2:

Summary
Sardiu et al have employed affinity purification with subsequent mass spectrometry based identification of proteins to three different INO80 family chromatin remodeling complexes that are conserved between yeast and humans. Using various computational analysis techniques, the authors argue that protein abundances of the members of this largely overlapping interaction network are conserved between yeast and humans. In the 2nd part of their paper the authors have validated 3 interactors of low abundances in the primary data set by reciprocal pull-downs.

General remarks
In their introduction the authors nicely work out the fact that interaction proteomics needs to become quantitative and I fully agree with this. Still, I am not very impressed by the biological significance of this paper, primarily because the sample size is so small. Obviously, the major conclusion that protein abundance within this network is largely conserved cannot be generalized to the entire proteome. It would be much more interesting to know on a proteome-wide scale, which functional categories are more and which less conserved. Since proteome-wide interaction and abundance data sets exists for yeast and humans, and this paper is largely computationally focused, I don't understand why those were not analyzed. As is, I feel this manuscript is not suitable for the broader readership of MSB.

Major points
- What does "several replicates on selected yeast and human baits" mean? At least three biological replicates for each bait would be state of the art.
- Regarding Figure 2: The respective paragraph on page 6 is rather confusing because it first talks about "26 orthologous bait proteins" and a few lines below about "26 prey proteins". I believe that baits are meant and that their abundances are compared across affinity purifications (absolute quantification). How was a comparable protein extraction/affinity purification/peptide yield ensured throughout all conditions? How do the values shown in Figure 2 compare to absolute protein abundances that were previously determined in both yeast and humans (e.g. J. Weissmann and M. Mann labs)?
- The authors also argue that they can predict missing values in the human data set based on the yeast data set (using leave out benchmarking). Since they first show that both data sets are highly similar, is this not trivial?
- The paper often concludes sections with essentially saying that the analysis shown here if benchmarked against previous knowledge, nicely demonstrates the capacity of a particular algorithm to classify the data. For example, the final conclusion of the 1st paragraph on page 9 is: "TDA demonstrates that the topology of the yeast and human datasets presented here are highly similar. These results demonstrate that TDA is a powerful query free tool for analyzing protein interaction networks datasets." I believe that this is more appropriate for a Journal dedicated to computational analysis of OMICS data.

Minor points
-Why was MudPIT used for such low complexity samples as affinity purifications?

Reviewer #3:

The authors investigate the evolutionary conservation of protein abundances for three chromatin remodeling complexes, Ino80, NuA4 and Swr1. Using quantitative proteomics they determine protein abundance levels for subunits of both yeast and human versions of the complexes and investigate different aspects of conservation, in particular protein abundances within and between the three chromatin complexes and the overall topology.

General remarks

Although the authors provide more detailed insights with regard to quantitative protein abundances for three chromatin complexes in yeast and human and the conclusions they make seem solid, I do think the conceptual advances provided are limited.

As the authors also indicate in their introduction a lot of effort has already been put in investigating protein complex compositions and protein interaction networks (particularly in yeast), albeit mostly using qualitative data and to some degree also using quantitative data. The concept of interactions being conserved within complexes and to a lesser degree between functionally related complexes (whether at qualitative or quantitative levels) is therefore not that novel. The manuscript should therefore mostly provides more detailed insights into the evolutionary conservation of three chromatin remodeling complexes.

The manuscript also has many analysis and textual descriptions that seem to describe the same data but using slightly different ways of presenting the same data. For instance fig. 2 and 3 all show very similar data, that is protein abundances within complexes being very conserved and therefore leading to clearly distinguishable protein complexes. This can be presented much more concise, for instance by moving figures to supplemental and avoiding repetitive conclusions/remarks in the text (more details below).
Overall, the manuscript text also needs some thorough revisions to make it more readable. Besides many long and grammatically incorrect sentences (making it hard to read). Some sections are also quite technical, making it difficult for a wider audience to read. I particularly found the topological data analysis (TDA) section confusing and way too technical. It is full with very technical terms/sentences that are not explained. Examples include: "To do this, we applied topological data analysis, which incorporates geometric approaches for the shape recognition within the data", "values of the geometric lens" and "principal metric SVD". It is also presented as novel in the abstract, whereas it clearly is a technique developed before as indicated by the references in the text.

Major points (in order of appearance)
- Page 7,8. The section dealling with figure 3 seems mostly a recap of figure 2, using a slightly different slice or viewpoint on the same data. For instance, the clusters shown in fig. 3a will of course look very similar to the ones provided in fig. 2 since it is a slightly different slice of the same data. The TDA analysis is also based on the same data and mostly seems needed to show that the TDA method in principle can work. I would suggest making fig. 3A supplemental to fig. 2 and merge 3b with fig. 4 to introduce the TDA method.
- Page 7,8. The whole TDA section is very technical (as indicated above) and needs to be much more clearly written. Judging the comparison of the topology between the yeast and human complexes is also difficult, since besides very beautifully looking pictures, the individual protein names can not be traced back. The authors only provide GO term enrichments as proof of the topology being the same between the two species (besides the Y shape), but since these are quite general, the underlying proteins could still be very different. Having access to the individual protein names will make it easier to judge the exact properties/degree of conservation. Panel C of fig. 4 is not needed for the comparison between the two yeast species, just showing A and B is sufficient.
- Page 11, figure 5. Panels D and E show the exact same information as Panel B and C, presented in a different way. Either show the correlation plot with the fitted regression line and corresponding p-value (B,C) or the residuals from the regression fit (D,E), but not both. I think most people are more familiar with correlation plots of the underlying data and not Q-Q plots, suggesting to keep B,C.
- Page 11, the conclusion "The resulting R2 values were high with significant p-values, indicating a similarity in protein abundance across two species in these two baits." doesn't fit with the paragraph. Besides the fact that this has already been concluded three times before (fig2,3a, 3b), the paragraph deals with the ability to predict protein abundances from one species to another.
- Page 12, the following sentence again draws a similar conclusion from similar data: "... indicating similar abundance levels between the yeast and human pulled-down proteins, thus confirming our previous results." Since the data used is highly overlapping with previous data, I don't see how this really confirms the data, it is not independent.
- Page 14, discussion, repetitive sentences saying almost the same thing: "We found that the abundance of orthologous protein pairs between yeast and human are highly correlated for all three complexes... " and "Within stable complexes, even the protein abundances of different members strongly correlate with each other..."
- Page 16, "We demonstrated that not only are members of chromatin remodeling complexes conserved among species, ..." This information is already obtained from the ortholog mapping, but

seems to be presented here as novel. The focus should be on the conservation of protein abundances.
- The above sentence is followed by another sentence, concluding the same thing "Not only are subunits of chromatin remodeling complexes conserved, but the relative abundance of components of these complexes is also conserved."

Minor points
- Page 7, first paragraph. according to the authors, spectral counts and total number of peptides provide nearly identical results. This is hard to judge exactly from figure 2 as you have to compare the individual heatmaps. Providing a correlation plot will make this much easier to gauge for the individual reader.
- page 8, Figure 2B should be Figure 3B I think.
- page 12, I could not find the list of conserved interactions in table E5.
- Page 12, three proteins passed the criteria and selected for experimental validation. Out of how many proteins? This is important to know in order to judge whether this is only a fraction of all proteins that could have been selected or not.
- Page 14, Figure E3 is used to indicate that the biological processes are similar. Don't know how to get this information from this figure.
- many textual errors, examples include (but not limited to): "... but also the abundance of those proteins is also conserved between...", "... a node can contains ...", "... the most important features TDA.", "For example, proteins that were located at the center of the data were members of the three complexes and closely associated proteins involved in chromatin machinery, were always located at the end of the Y shape as colored in blue", "... not only are members of chromatin remodeling complexes are conserved among ...".
- Please ensure that the proteomics data is deposited in the appropriate public repositories.

1st Editorial Decision after transfer to EMBO reports 04 August 2014

Many thanks for transferring your manuscript to EMBO reports. I think it could be a very good fit for our journal and I would thus invite you to revise it based on the referee reports from Molecular Systems Biology.

Specifically, referee 1 raised some points that I would like you to address (for example, it would be quite interesting if you could show that protein abundance in complexes is higher than that of individual orthologous proteins). Other points raised by this reviewer (e.g. points 2 and 3) and his/her two colleagues would just need further clarifications that should be rather straight forward and I would refer you to their respective reports for details.

We would, of course, not require you to extend your analysis to the entire proteome as suggested by referee 2.

Formally, the manuscript is slightly too long for our format, but referee 3 makes good suggestions on how to condense it. I would recommend shortening the text to about 35,000 characters if possible (it is now about 55,000 including spaces). Also, it would be good if you could move one of the currently seven figures to the supplementary section or combine it with one of the other main figures so that in the end, there are not more than six main figures.

Once you have modified your study accordingly, please submit the final version through our website.

I look forward to seeing a revised form of your manuscript when it is ready.

Response to Reviewer's Comments

To begin, we would like to thank all three reviewers for their careful evaluation of our manuscript and for their positive and constructive comments. Based on all three reviewer's comments significant changes were made to the manuscript. By focusing and clarifying our work on the key topics, we believe our manuscript is significantly improved and more accessible to researchers in systems biology, proteomics, and chromatin remodeling.

Response to Reviewer #1.

Summary and General Remarks

The present study, titled 'Conserved Abundance and Topological Features in Chromatin Remodeling Protein Interaction Networks,' presents a compelling story regarding the conservation of abundance among orthologous proteins involved in chromatin remodeling complexes between yeast and human. Through a quantitative proteomics and analytical framework, the authors suggest that yeast-human orthologs involved in three different chromatin remodeling complexes retain relative abundances between themselves, but not between members of the other complexes. Further, through analysis of the complex interaction network topology, they show conservation of connections among not only chromatin remodeling, but its relation to transcriptional regulation and RNA processing, among others. They further suggest a method for predicting missing abundances by inferring them from another species, a potentially useful method for the proteomics community.

The conclusions presented in the study are generally well supported. While the studies findings of conservation of orthologous protein abundances has been shown before among many species (as cited by the authors), previous studies have used various previously published datasets and it has thus been difficult to control for variation based on preparation or equipment differences. The authors base their findings on an in-house proteomics dataset, thus controlling for some of these variables. They also present a novel, though not unexpected, conservation of network topology related to chromatin remodeling and other nucleic acid regulatory functions. The study should be relevant to many systems biologists, particularly those in the proteomics and epigenetic fields.

Major Points

1. The authors show a high correlation of abundance values between orthologous complex members, while previous studies have already suggested that orthologous proteins in general have highly correlated abundances. It would therefore be quite interesting to show that the correlation of orthologs in complexes is even higher than that of the general protein abundance conservation between orthologs. Related to this, do the authors anticipate that significant bias is being introduced in the abundance measurements by using baited proteins as opposed to bait free proteomics?

Response: Reviewer #1 raises an interesting question regarding conservation between complex members and general protein abundance conservation. In fact, the general protein abundance conservation between orthologs has been studied previously [1]. Weiss *et al.* combined the spectral counts together with the Spearman rank correlation to examine the conservation of abundance of all orthologous proteins between five different organisms using proteomics [1].
They reported a Spearman rank correlation of 0.64 between all yeast and human orthologous proteins. As described in our study we observed higher correlation (_ >0.9) for stable complexes such as INO80, however for complexes with shared proteins we observed a correlation with an average of 0.7. Thus, we can say that unique protein complexes show a higher abundance correlation than that of general protein abundance conservation between orthologos. We have revised the manuscript on page 16 to include this information.

Most biological processes are performed by protein complexes. For several years our work has been focused on chromatin machinery interrogation and thus we aimed to provide further information on these complexes rather than protein identities alone. There are many ways to biochemically investigate these complexes, each method with its individual advantage and drawbacks. We used an affinity purification approach to enrich for these proteins and their interactions. The use of affinity purification coupled with quantitative proteomics is a widely used approach for the study of protein complexes and protein networks [2-6]. Because of this approach, proteins of these complexes and their associated interactions will have an abundance significantly higher than if we were to simply identify these proteins from a whole cell lysate or nuclei preparation, for example. However, without purifying the protein complexes in advance, we would not have the critical connectivity of association that is important in distinguishing protein complexes from each other. That being said, this question raised here by Reviwer #1 regarding conservation of abundance of orthologs in complexes versus orthologs in general is an interesting question that requires further investigation by the field.

2. It is not clear how the opposite species values are used to infer missing values for the current species. SVDimpute requires complete matrices to start, and several methods exist for assigning initial values to missing points. Are the authors using these? Or is it simply using data from one species to initialize the missing values in the species being tested? If the latter (or former, for that matter), it is totally unclear that this is the case. Further, there is no explanation of why SVDimpute was chosen over the existing methods, such as row-averaging or K-nearest neighbors. I would thus like to see some evaluation of these other methods and comparison against SVDimpute, and also comparison with SVDimpute using different initial values.

Response: This was another insightful comment from Reviewer #1. We have endeavored to clarify our description of the use of SVDimpute in the manuscript on page 7 in the results and discussion and pages 14-15 in the methods.

In order to predict missing values in human data we indeed used the yeast data. The rationale of choosing SVDimpute over other existing methods in now provided on page 7. We started our evaluation with SVDimpute since it was shown by Troyanskaya *et al* that the row average approach yielded drastically lower accuracy than either KNN- or SVD-based estimation [7]. In addition, KNN approach replaces NaNs (i.e. missing values) in data with the corresponding value from the nearest-neighbor column or replaces NaNs in data with a weighted mean of the k nearest-neighbor columns [7], which we felt that generally is more simplistic than the SVD method.

However, in response to Reviewer #1's comment, we directly tested and compared the KNN and SVD impute methods, and SVDimpute outperformed KNN (See Table 1 below). In six of the seven cases where a true spectral count (SpC) of a prey was available, SVDImpute provided a closer approximation of the true value than KNN based on the % of True Value column (See Table 1 below). We implemented two different functions for KNN algorithm in R (i.e. impute.knn() and knnImputation()) and both methods give similar results. Therefore, for this study we believe that SVDImpute is the better missing value estimation method. However, we believe that Reviewer #1 is correct and missing value estimation, like the approach shown in this manuscript, is an area in need of further study by the field.

Table 1. Model prediction of proteomics data across species using svdimpute and impute.knn

Baits from INO80 complex	Prey Protein	Total SpC in the bait	True SpC value of the prey	Estimated SpC of the prey using SVDimpute method	% of True Value	k	Estimated spectra value using Impute.KNN method	% of True Value	Spearman correlation
INO80B_2	INO80B	9744	1472	1428.051	3	6	1214.33	17	1
INO80B_1	INO80	10138.31	515	582.1604	13	6	1485.38	188	1
ACTR5	RUVBL2	777	81	89.62841	10	3	55	32	1
INO80C_3	ACTR5	241	49	49.72644	1	3	24.33	50	1
ACTR8	ACTL6A	961	29	31.47442	9	3	24.66	15	1
INO80C_3	ACTL6A	241	22	18.53566	15	3	20	9	1
ACTR5	ACTL6A	777	20	25.48	27	2	12.5	38	1
INO80*	INO80C	89	0	3.851589	Nd	3	6.7	Nd	NA
INO80C_2*	INO80B	126	0	10.52477	Nd	3	10.5	Nd	NA

3. The section regarding the discovery of a conserved low-abundance subnetwork is somewhat unclear. While interesting, there is no mention of abundance, low or high, anywhere apart from the section heading and the final concluding sentence, and so it is unclear what exactly is meant by 'low-abundance' or why that feature would be relevant or novel.

Response: We have revised the section on page 10 that describes the abundance of these proteins in comparison to the core proteins in the three complexes. The key sentence now states "The three low abundant proteins in yeast were TMA19, YAP1802, and DHH1 which were 38, 21, and 16 fold lower in abundance than core proteins in the three complexes, respectively (Tables E1-E2)." We determined this by averaging the dNSAF values of all the proteins in the core complex, and then comparing these results to these three proteins. In addition we added a new Figure E6 that shows these proteins statistical significance in the proteomics samples and microarray mutants.

Regarding the relevance of this analysis, we chose to examine proteins with lower abundance for several purposes. First, our lab has demonstrated the usefulness in identifying low abundant proteins that interact with related protein complexes [8, 9]. Second, these proteins are less studied (less than 50 interactions in BIOGRID for the human proteins) and as shown by OMIM (OMIM: 600763, OMIM: 603025 and OMIM: 600326) are involved in multiple diseases.
Therefore to gain more insight in their potential function, we examined also their candidate interactions which linked them also to chromatin machinery pathway.

Additionally, Table E5 is mentioned as containing the final microarray filtered list of ortholog pairs, however only the 214 pairs from before filtering are apparent.

Response: In Table E6 under the "expression profiles" page we have reported all the genes in 11 deletion mutants lacking chromatin machinery components of the three complexes studied in our work from the dataset of Leenstra *et al.* [10]. We have now listed all the proteins that are passing the criteria under the spreadsheet number 5 and added this detail in the manuscript on pages 9-10.

It is also unclear what groups of proteins are being overlapped in the hypergeometric analysis. Clarification of this should be included in the text, Fig. 7 legend, and methods.

Response: We have added more details to the Figure 7 (now Figure 5) legend and to the methods on page 15. Basically, the groups that are being overlapped here are the chromatin remodeling complexes found in the purifications. They are listed in Figure 5 next to or above the individual proteins in the figure. The protein complexes listed include SAGA/STAGA, INO80, SIN3, SWR/SRCAP, and SWI/SNF, for example.

Minor Points

1. There are discrepancies throughout the manuscript as to the number of orthologs studied as baits. The abstract mentions 18, the beginning of the results sections mentions 15, and Figure 1 and elsewhere in the text mention 16. Further, the first paragraph of the results mentions "three orthologous proteins between the human TIP60 and yeast NuA4...", while there are actually four labelled in Figure 1.

Response: We used 15 orthologous baits from three chromatin complexes to begin our analysis.
To this, three additional orthologous baits corresponding to the newly identified proteins (i.e. TMA19/TPT1, YAP1802/PICALM and DHH1/DDX6) were then added to a total of 18 orthologous baits. We have carefully checked the manuscript to clarify this. Finally, there are indeed four pairs labelled in Figure 1, however one of the pair, YL1/VPS72, is shared between SRCAP and TIP60 complexes.

2. It should be mentioned in the main text regarding use of the Spearman correlation that Spearman is a correlation of ranks, and not of actual values. Perhaps by specifically referring to it as such (i.e. "we computed the Spearman rank correlation") or just mentioning it in the explanation given for what correlation is. Spearman is probably the correct metric here, but would just like to see that detail mentioned in the main text. On a related note, the section describing what correlation values are and what they mean is probably unnecessary for the main text.

Response: Based on the reviewer's comment we have revised and simplified the results and discussion section on page 5 and the methods section on page 13. We have now specified that the correlation is based on protein's abundance rank, and we removed the background information as the reviewer suggested.

3. In Figure 5A, for the gene ACTR5, a k-value of 2 is reported. The text mentions that only k values from 3-8 were tested. This discrepancy should be clarified.

Response: We have corrected this in the manuscript on page 8.
4. There are several typos and grammatical errors involving sentence structure, etc. throughout. A very thorough reading of the manuscript for grammatical and spelling correctness is Suggested.

Response: We have carefully revised, corrected, shortened, and focused the manuscript to alleviate this issue.

Response to Reviewer #2

Summary
Sardiu et al have employed affinity purification with subsequent mass spectrometry based identification of proteins to three different INO80 family chromatin remodeling complexes that are conserved between yeast and humans. Using various computational analysis techniques, the authors argue that protein abundances of the members of this largely overlapping interaction network are conserved between yeast and humans. In the 2nd part of their paper the authors have validated 3 interactors of low abundances in the primary data set by reciprocal pull-downs.

General remarks
In their introduction the authors nicely work out the fact that interaction proteomics needs to become quantitative and I fully agree with this. Still, I am not very impressed by the biological significance of this paper, primarily because the sample size is so small. Obviously, the major conclusion that protein abundance within this network is largely conserved cannot be generalized to the entire proteome. It would be much more interesting to know on a proteomewide scale, which functional categories are more and which less conserved. Since proteomewide interaction and abundance data sets exists for yeast and humans, and this paper is largely computationally focused, I don't understand why those were not analyzed. As is, I feel this manuscript is not suitable for the broader readership of MSB.

Response: While Reviewer #2 is correct that it would be interesting to conduct a study like ours on a large scale, we respectfully disagree that our study is somehow diminished by covering the entire human and yeast protein interaction networks. While data exists, it is fragmented, not universally quantitative, and there is no single quantitative protein interaction network covering either *S. cerevisiae* or *H. sapiens*. We chose to conduct a focused study on matched protein complexes in chromatin remodeling using the same quantitative proteomics approach, which will dramatically alleviate the issues that arise when attempting to merge datasets generated in different labs using different approaches.

Major points
- What does "several replicates on selected yeast and human baits" mean? At least three biological replicates for each bait would be state of the art.

Response: We have used at least three replicates for the majority of the baits. However, some of the baits in our human dataset have fewer replicates. We intentionally included these in the analysis for the development of an approach for the prediction of the missing values (pages 7-9)

- Regarding Figure 2: The respective paragraph on page 6 is rather confusing because it first talks about "26 orthologous bait proteins" and a few lines below about "26 prey proteins". I believe that baits are meant and that their abundances are compared across affinity purifications (absolute quantification). How was a comparable protein extraction/affinity purification/peptide yield ensured throughout all conditions? How do the values shown in Figure 2 compare to absolute protein abundances that were previously determined in both yeast and humans (e.g. J. Weissmann and M. Mann labs)?

Response: We have carefully revised the manuscript to properly use 'baits' and 'preys' where appropriate throughout the manuscript. In particular, the section regarding Figure 2 has been revised on page 5. Next, Reviewer #2 is correct that ensuring a comparable protein extraction/affinity purification/peptide yield ensured throughout all conditions is important. A major reason that we use MudPIT for the analysis of protein complexes is to ensure this. For example, we have written specifically on the reproducibility of MudPIT for protein complex analysis [9]. Figure 2 is an excellent demonstration of the value of our methods, not only are the results within an organism comparable and insightful, but also the results between yeast and humans are comparable and insightful. Finally, we are not presenting nor claiming absolute protein abundance values in this study. Also we present a focused study on chromatin remodeling complexes rather than entire proteomes. While a detailed analysis and comparison of our results to the absolute protein abundances may be interesting, it is unclear what new insights this may provide, and it is beyond the scope of the current manuscript.

- The authors also argue that they can predict missing values in the human data set based on the yeast data set (using leave out benchmarking). Since they first show that both data sets are highly similar, is this not trivial?

Response: Based on our analysis of the literature we are unaware of prediction methods for proteomics data with missing values. We believe that is not a trivial problem. It is important to note that Reviewer #1 found this important and an area in need of further development (see page 2 point #2). In our response to Reviewer #1's question regarding missing value estimation we present a comparison of methods to show why we chose to use SVDImpute. This demonstrates the computational challenge of missing value prediction in quantitative proteomics datasets.

- The paper often concludes sections with essentially saying that the analysis shown here if benchmarked against previous knowledge, nicely demonstrates the capacity of a particular algorithm to classify the data. For example, the final conclusion of the 1st paragraph on page 9 is: "TDA demonstrates that the topology of the yeast and human datasets presented here are highly similar. These results demonstrate that TDA is a powerful query free tool for analyzing protein interaction networks datasets." I believe that this is more appropriate for a Journal dedicated to computational analysis of OMICS data.

Response: In this revision, we have focused and clarified the manuscript on the key aspects of our research that are relevant to the systems biology, proteomics and chromatin remodeling communities. We present a quantitative proteomic analysis of conserved yeast and human chromatin remodeling complexes and we show that the protein complexes from purifying these complexes shows a higher abundance correlation than that of general protein abundance conservation between orthologs. We are among the first to employ a topological analysis for quantitative proteomics data and showed that yeast and human interactions of chromatin machinery have similar network topology. We demonstrated the value of the missing estimation method using cross-species prediction. From the yeast-human conservation data we could pin point new associated protein interactions with these complexes. While there is certainly a dominant computational component to our manuscript,

we strongly believe that our research presented here will be of high interest to multiple fields of study.

Minor points
-Why was MudPIT used for such low complexity samples as affinity purifications?

Response: We use MudPIT for protein complex analysis to obtain detailed analyses of protein complexes and protein interaction networks. It has been a highly successful approach that has led to many important biological discoveries, please see [11-16] for a few recent examples. We believe that while protein complexes have lower complexity than whole proteomes and cell extracts, it is still critically important to use a powerful chromatographic approach to obtain deep and detailed information on proteomics samples of all types.

Response to Reviewer #3:

The authors investigate the evolutionary conservation of protein abundances for three chromatin remodeling complexes, Ino80, NuA4 and Swr1. Using quantitative proteomics they determine protein abundance levels for subunits of both yeast and human versions of the complexes and investigate different aspects of conservation, in particular protein abundances within and between the three chromatin complexes and the overall topology

General remarks

Although the authors provide more detailed insights with regard to quantitative protein abundances for three chromatin complexes in yeast and human and the conclusions they make seem solid, I do think the conceptual advances provided are limited.

As the authors also indicate in their introduction a lot of effort has already been put in investigating protein complex compositions and protein interaction networks (particularly in yeast), albeit mostly using qualitative data and to some degree also using quantitative data. The concept of interactions being conserved within complexes and to a lesser degree between functionally related complexes (whether at qualitative or quantitative levels) is therefore not that novel. The manuscript should therefore mostly provides more detailed insights into the evolutionary conservation of three chromatin remodeling complexes.

The manuscript also has many analysis and textual descriptions that seem to describe the same data but using slightly different ways of presenting the same data. For instance fig. 2 and 3 all show very similar data, that is protein abundances within complexes being very conserved and therefore leading to clearly distinguishable protein complexes. This can be presented much more concise, for instance by moving figures to supplemental and avoiding repetitive conclusions/remarks in the text (more details below).

Overall, the manuscript text also needs some thorough revisions to make it more readable. Besides many long and grammatically incorrect sentences (making it hard to read). Some sections are also quite technical, making it difficult for a wider audience to read. I particularly found the topological data analysis (TDA) section confusing and way too technical. It is full with very technical terms/sentences that are not explained. Examples include: "To do this, we applied topological data analysis, which incorporates geometric approaches for the shape recognition within the data", "values of the geometric lens" and "principal metric SVD". It is also presented as novel in the abstract, whereas it clearly is a technique developed before as indicated by the references in the text.

Response: We used Reviewer #3's suggestions extensively during the revision of our manuscript. The first version of our manuscript was 29 pages with 7 figures in the main body of the text. Here we present a drastically revised manuscript that is 20 and 1/3rd pages with 5 figures in the main body of the text. For example, the new figure 3 combines parts of prior figures 3 and 4 to have only one figure on topological data analysis. The topological data

analysis section of the manuscript on pages 6-7 and in the methods on pages 13-14 have been rewritten to be more focused, clear, and concise.

Major points (in order of appearance)

- Page 7,8. The section dealing with figure 3 seems mostly a recap of figure 2, using a slightly different slice or viewpoint on the same data. For instance, the clusters shown in fig. 3a will of course look very similar to the ones provided in fig. 2 since it is a slightly different slice of the same data. The TDA analysis is also based on the same data and mostly seems needed to show that the TDA method in principle can work. I would suggest making fig. 3A supplemental to fig. 2 and merge 3b with fig. 4 to introduce the TDA method.

- Page 7,8. The whole TDA section is very technical (as indicated above) and needs to be much more clearly written. Judging the comparison of the topology between the yeast and human complexes is also difficult, since besides very beautifully looking pictures, the individual protein names cannot be traced back. The authors only provide GO term enrichments as proof of the topology being the same between the two species (besides the Y shape), but since these are quite general, the underlying proteins could still be very different. Having access to the individual protein names will make it easier to judge the exact properties/degree of conservation. Panel C of fig. 4 is not needed for the comparison between the two yeast species, just showing A and B is sufficient.

Response: Reviewer #3 is correct and we have done exactly what was suggested and rearranged figures according to these suggestions. Figure 3A is now Figure E3 and Figure 2 is a merge of what was Figure 3B with Figure 4A and 4B and prior Figure 4C is now Figure E2. This has resulted in a rewriting and clarification of 'Conservation of Topology Between Species' section on pages 6-7 in the results and discussion section, the 'Topological Data Analysis' section on pages 13-14 in the methods section, and the legend to Figure 3 on pages 19-20 of the manuscript.
Finally, as requested, we have added Table E5 to the manuscript that contains the protein names of individual proteins in each of the flares in both the yeast and human topological networks. As a result of Reviewer #3's careful consideration of our original manuscript, we believe that this section is much more concise, clear, and insightful.

- Page 11, figure 5. Panels D and E show the exact same information as Panel B and C, presented in a different way. Either show the correlation plot with the fitted regression line and corresponding p-value (B,C) or the residuals from the regression fit (D,E), but not both. I think most people are more familiar with correlation plots of the underlying data and not Q-Q plots, suggesting to keep B,C.

Response: The Q-Q plots have been removed from what is now Figure 4 as suggested.

- Page 11, the conclusion "The resulting R2 values were high with significant p-values, indicating a similarity in protein abundance across two species in these two baits." doesn't fit with the paragraph. Besides the fact that this has already been concluded three times before (fig2,3a, 3b), the paragraph deals with the ability to predict protein abundances from one species to another.

Response: We have revised this section of the manuscript on page 8 in accordance with the reviewer's comment.

- Page 12, the following sentence again draws a similar conclusion from similar data: "... indicating similar abundance levels between the yeast and human pulled-down proteins, thus confirming our previous results." Since the data used is highly overlapping with previous data, I don't see how this really confirms the data, it is not independent.

Response: We have '…thus confirming our previous results' from page 9 of the revised manuscript.

- Page 14, discussion, repetitive sentences saying almost the same thing: "We found that the abundance of orthologous protein pairs between yeast and human are highly correlated for all three complexes... " and "Within stable complexes, even the protein abundances of different members strongly correlate with each other..."

Response: As a result of revising the manuscript for EMBO Reports and combining the results and discussion section, these sentences have been removed.

- Page 16, "We demonstrated that not only are members of chromatin remodeling complexes conserved among species, ..." This information is already obtained from the ortholog mapping, but seems to be presented here as novel. The focus should be on the conservation of protein abundances.

- The above sentence is followed by another sentence, concluding the same thing "Not only are subunits of chromatin remodeling complexes conserved, but the relative abundance of components of these complexes is also conserved."

Response: The final paragraph of the combined results and discussion section on page 11 of the revised manuscript has been rewritten to focus on the conservation of protein abundance.

Minor points

- Page 7, first paragraph. according to the authors, spectral counts and total number of peptides provide nearly identical results. This is hard to judge exactly from figure 2 as you have to compare the individual heatmaps. Providing a correlation plot will make this much easier to gauge for the individual reader.

Response: All the correlations values used to generate Figure 2 are provided in supplementary Table E4.

- page 8, Figure 2B should be Figure 3B I think.

Response: As described earlier, the section regarding topological data analysis on pages 6-7 have been rewritten and the figures reorganized as suggested by Reviewer #3.

- page 12, I could not find the list of conserved interactions in table E5.

Response: The conserved interactions together with their quantitative values can be found in the supplementary Table E6 in the “orthologs” worksheet.

- Page 12, three proteins passed the criteria and selected for experimental validation. Out of how many proteins? This is important to know in order to judge whether this is only a fraction of all proteins that could have been selected or not.

Response: We have added this information to the manuscript on the top of page 10.

- Page 14, Figure E3 is used to indicate that the biological processes are similar. Don't know how to get this information from this figure.

Response: We have now provided all the proteins located on each of the main network flared in the supplementary Table E5. Also, prior Figure E3 is now Figure E7 in the revised manuscript.

- many textual errors, examples include (but not limited to): "... but also the abundance of those proteins is also conserved between...", "... a node can contains ...", "... the most important features TDA.", "For example, proteins that were located at the center of the data were members of the three complexes and closely associated proteins involved in chromatin machinery, were always located at the end of the Y shape as colored in blue", "... not only are members of chromatin remodeling complexes are conserved among ...".

Response: We have endeavored to remove all textual errors during the process of revising our manuscript.

- *Please ensure that the proteomics data is deposited in the appropriate public repositories.*

Response: This information is provided on page 13 of the manuscript. Data from this publication is available via the PeptideAtlas database (http://www.peptideatlas.org/), assigned the identifier PASS00491 (password JM6934n) at ftp://PASS00491:JM6934n@ftp.peptideatlas.org/.

References Used in Response to Reviewer's Comments

1. Weiss M, Schrimpf S, Hengartner MO, Lercher MJ, von Mering C (2010) Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **10:** 1297-1306
2. Sowa ME, Bennett EJ, Gygi SP, Harper JW (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138:** 389-403
3. Hauri S, Wepf A, van Drogen A, Varjosalo M, Tapon N, Aebersold R, Gstaiger M (2013) Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. *Mol Syst Biol* **9:** 713
4. Joshi P, Greco TM, Guise AJ, Luo Y, Yu F, Nesvizhskii AI, Cristea IM (2013) The functional interactome landscape of the human histone deacetylase family. *Mol Syst Biol* **9:** 672
5. Sardiu ME, Cai Y, Jin J, Swanson SK, Conaway RC, Conaway JW, Florens L, Washburn MP (2008) Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc Natl Acad Sci U S A* **105:** 1454-1459
6. Breitkreutz A *et al* (2010) A global protein kinase and phosphatase interaction network in yeast. *Science* **328:** 1043-1046
7. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17:** 520-525
8. Mosley AL, Pattenden SG, Carey M, Venkatesh S, Gilmore JM, Florens L, Workman JL, Washburn MP (2009) Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Mol Cell* **34:** 168-178
9. Mosley AL, Sardiu ME, Pattenden SG, Workman JL, Florens L, Washburn MP (2011) Highly reproducible label free quantitative proteomic analysis of RNA polymerase complexes. *Mol Cell Proteomics* **10:** M110 000687
10. Lenstra TL *et al* (2011) The specificity and topology of chromatin interaction pathways in yeast. *Mol Cell* **42:** 536-549
11. Sardiu ME *et al* (2014) SAHA Induced Dynamics of a Human Histone Deacetylase Protein Interaction Network. *Mol Cell Proteomics*, 10.1074/mcp.M113.037127
12. Herz HM *et al* (2014) Histone H3 lysine-to-methionine mutants as a paradigm to study chromatin signaling. *Science* **345:** 1065-1070
13. Banks CA *et al* (2014) Controlling for gene expression changes in transcription factor protein networks. *Mol Cell Proteomics* **13:** 1510-1522
14. Rossi M, Duan S, Jeong YT, Horn M, Saraf A, Florens L, Washburn MP, Antebi A, Pagano M (2013) Regulation of the CRL4(Cdt2) ubiquitin ligase and cell-cycle exit by the SCF(Fbxo11) ubiquitin ligase. *Mol Cell* **49:** 1159-1166
15. Hu D *et al* (2013) The little elongation complex functions at initiation and elongation phases of snRNA gene transcription. *Mol Cell* **51:** 493-505
16. D'Angiolella V *et al* (2012) Cyclin F-mediated degradation of ribonucleotide reductase M2 controls genome integrity and DNA repair. *Cell* **149:** 1023-1034

2nd Editorial Decision 27 October 2014

Thank you for the submission of your revised study to our editorial office. I am very pleased to accept your manuscript for publication in the next available issue of EMBO reports.

Thank you for your contribution to EMBO reports and congratulations on a successful publication. Please consider us again in the future for your most exciting work.

