molecular systems biology

Deep proteome and transcriptome mapping of a human cancer cell line

Nagarjuna Nagaraj, Jacek Wisniewski, Tamar Geiger, Jürgen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, Matthias Mann

Corresponding author: Matthias Mann, Max-Planck Institute for Biochemistry

Review timeline:

Submission date:	14 August 2011
Editorial Decision:	16 September 2011
Revision received:	21 September 2011
Accepted:	29 September 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 16 September 2011

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the two referees who accepted to evaluate the study. As you will see, the referees find the topic of your study of potential interest and are supportive. They raise however a series of concerns and make suggestions for modifications, which we would ask you to carefully address in a revision of the present work.

Please deposit your raw data in one of the major public databases and include the respective hashcodes/accessions in a "data availability" sub-section within the Materials and Methods section of the main paper.

*** 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 .
Thank you for submitting this paper to Molecular Systems Biology.

Yours sincerely,

Editor
Molecular Systems Biology

Referee reports

Reviewer #1 (Remarks to the Author):

Review for Nagaraj et al. - Deep proteome and transcriptome mapping of a human cancer cell line.

This is one of two manuscripts submitted back-to-back to Molecular Systems Biology using an established human cancer cell line model to obtain an in-depth proteome description.

Both manuscripts are of high quality and impressively demonstrate the capabilities of modern shotgun proteomics to basically identify, and to a certain degree quantify, the proteome of a mammalian cell. Both manuscripts arrive at similar results and conclusion, despite using slightly different mining strategies. Although, not geared towards specific biology, both manuscripts provide the first true high-level cell/systems level insight into the nearly complete proteome of a simple model system. These data will provide the cornerstone for similar proteome projects in the future.

The data quality of both papers is very high and the high-level analysis of this large amount of data is useful and reasonable. Asides from "real" biology, which is clearly not the focus of these current manuscripts there are only a few minor comments that should be addressed editorially.

An interesting feature of this current paper is the direct comparison to RNA-seq, which nicely demonstrates that both methods are highly comparable. An additional very important observation was the identification of peptides from unannotated exons.

1) Raw data should be deposited to Tranche. This will enable further mining of these data by others.
2) It was not completely clear how the iBAQ quantification works. How would this compare to an independent method, such as SRM, which relies on internal spike-in experiments to arrive at quantification and copy numbers. Peptide signal in MS modes relies on a variety of peptide specific physical properties. How is this taken into account?
3) Page 7: "This unbiased view of gene expression at the transcriptome and proteome levels will be invaluable in biological and biomedical research". This statement might be a bit overly optimistic. In biomedical research clinical samples will have massive variability and limited amount. On top several dozens samples need to be analyzed to arrive at statistically useful results. It is unlikely that a single sample can be analyzed for 12 days. This caveat should be mentioned, or the statement removed.
4) Figures 1B/C are not useful and uninformative. They are just random scans of a peptide. We all know what Orbitrap spectra look like.

In summary, both papers are highly suitable for publication in MSB and should be of high interest to the systems biology readership.

Reviewer #2 (Remarks to the Author):

Deep proteome and transcriptome mapping of a human cancer cell line
In a tour de force Nagaraj et al. performed a very comprehensive proteome analysis of unsynchronized HeLa cells. The authors identified 10, 255 different human proteins encoded by 9207 genes. In addition to sequencing the proteome of this cell line the authors also performed a deep sequencing experiment to validate the presence of these proteins from the transcriptome data. Based on their results that nearly all the proteins identified have an associated transcript that was sequenced the authors suggest that they have achieved a deep coverage of the functional transcriptome and proteome of this cell type. In addition, they applied iBAQ-based quantification of the detected proteins and correlated these determined quantities with transcriptomics data obtained by RNA-seq on an Ilumina platform. The dataset will be very useful for the community and thus should eventually be published despite a lack of follow-up studies.

However, prior to accepting this publication, we suggest the following revisions and expanded discussions:
The abstract should be revised and reworded; e.g. "... 166,420 sequence unique peptides ..."; "... are

within a factor 60 of the median." It is not clear what this means until one reads the article. Thus we suggest adding e.g. "...90 % of them cover less than 4 orders of magnitude and are within a factor of median."
Page 3/1st Results and Discussion Paragraph: in a rather cryptic and incomplete way the authors refer to Experiment 1, which does not really add to the manuscript. Thus, I strongly suggest significantly shortening this paragraph and reducing it to the bottom line that the method was optimized to provide "very deep coverage" with 'only' 12 days of instrument time.
On page 4, the authors indicate that ~500 proteins are 1-peptide identifications. It would be helpful to include a short discussion as to how these proteins were reliably quantified.
Page 4/description of the RNA-seq data:
a. How many of the 49 000 transcripts were unique?
b. Why was FPKM used for normalization of the data as opposed to RPKM?
c. Is a subset of these transcripts splice variants of the same genes?
d. Were some of these transcripts ambiguous as to which homolog they may belong to (gene families, etc)?
e. How did the authors deal with homologous genes?
f. Given the data in the following references (Mingyao Li et al. Science 330(6012) (2010) PMID 21596952) Did any mass spectra match to any "non-coding regions" thus identifying new open reading frames?

Page 5: Given the importance of iBAQ for the manuscript at hand, it would be very beneficial for the readers to have a short description/discussion about this method.
Page 5: How were the 37 proteins selected for the determination of the absolute copy numbers? Do the selected proteins show a similar functional/localization distribution as the entire protein set? If not, how do the authors ensure that the determined absolute copy numbers are transferrable to other protein families?
Page 5/last line: the authors mention that some protein complexes showed lower coverage. How low? For the comparison purposes it would be helpful to have more definite information.
Page 6: The authors claim that "..., judged against the coverage achieved by deep-sequencing transcriptomics, the proteomics data was more than 90% complete (...)." Depending on the calculation, we determined values between 73 and 77 % (9207/(11936+598) or 9207/11936). Please clarify!
Page 6/last line: "...integral membrane proteins account for 25% of the genome but contribute much less to the transcriptome and the proteome (6% of identified proteins)." What percentage of these genes are found in the transcriptome data? We would like this number to make a proper comparison.

1st Revision - authors' response 21 September 2011

Monday, September 19, 2011

Point by point response to reviewers of "Nagaraj et al. - Deep proteome and transcriptome mapping" **(MSB-11-3166).**

We thank the reviewers for their constructive and positive comments and we have endeavored to answer all of them below and in the revised manuscript.

Reviewer #1 (Remarks to the Author):

Review for Nagaraj et al. - Deep proteome and transcriptome mapping of a human cancer cell line.

This is one of two manuscripts submitted back-to-back to Molecular Systems Biology using an established human cancer cell line model to obtain an in-depth proteome description.

Both manuscripts are of high quality and impressively demonstrate the capabilities of modern shotgun proteomics to basically identify, and to a certain degree quantify, the proteome of a mammalian cell. Both manuscripts arrive at similar results and conclusion, despite using slightly different mining strategies. Although, not geared towards specific biology, both manuscripts provide the first true high-level cell/systems level insight into the nearly complete proteome of a simple model system. These data will provide the cornerstone for similar proteome projects in the future.

The data quality of both papers is very high and the high-level analysis of this large amount of data is useful and reasonable. Asides from "real" biology, which is clearly not the focus of these current manuscripts there are only a few minor comments that should be addressed editorially.

An interesting feature of this current paper is the direct comparison to RNA-seq, which nicely demonstrates that both methods are highly comparable. An additional very important observation was the identification of peptides from unannotated exons.

We thank the reviewer for his or her positive comments.

1) Raw data should be deposited to Tranche. This will enable further mining of these data by others.

We completely agree with the reviewer and we are in the process of uploading the data to TRANCHE. However due to technical difficulties with TRANCHE, hash codes for only parts of the data are provided

now in the methods section. We will provide the hash codes for all the remaining files before final submission and likely by the end of this week.

2) It was not completely clear how the iBAQ quantification works. How would this compare to an independent method, such as SRM, which relies on internal spike-in experiments to arrive at quantification and copy numbers. Peptide signal in MS modes relies on a variety of peptide specific physical properties. How is this taken into account?

Both reviewers inquired about the iBAQ method and we therefore expanded the explanations in the text. The iBAQ method has been described by Selbach and colleagues in their recent Nature paper (Schwanhausser et al, 2011), which we cite in our paper. Basically it normalizes the summed peptide intensities to the number of theoretically observable peptides of the protein. This calculation provides only a rough estimate of protein abundance and it does not provide the accurate protein copy number for any given protein of interest. Furthermore, the accuracy depends on the number of identified peptides and on their intensities. Therefore, quantification accuracy of proteins that were identified with single peptides is even lower. This was already noted in the legend to supplementary Table S7.

We added the following to the main text (p5): **"To calculate the approximate abundance of each protein we used the iBAQ algorithm (Schwanhausser et al, 2011), which normalizes the summed peptide intensities by the number of theoretically observable peptides of the protein. These normalized protein intensities was translated to protein copy number estimates based on the overall protein amount in the analyzed sample."**

3) Page 7: "This unbiased view of gene expression at the transcriptome and proteome levels will be invaluable in biological and biomedical research". This statement might be a bit overly optimistic. In biomedical research clinical samples will have massive variability and limited amount. On top several dozens samples need to be analyzed to arrive at statistically useful results. It is unlikely that a single sample can be analyzed for 12 days. This caveat should be mentioned, or the statement removed.

We agree with the reviewer and have removed the sentence.

4) Figures 1B/C are not useful and uninformative. They are just random scans of a peptide. We all know what Orbitrap spectra look like.

We provided the spectra in Figures 1B/C to emphasize the high-resolution MS analysis. This is because in our experience many readers still have trouble distinguishing 2D gels from mass spectrometry approaches and even more difficulties distinguishing low resolution MS approaches from high accuracy and high resolution data. However we have removed these panels according to the reviewer's suggestion.

In summary, both papers are highly suitable for publication in MSB and should be of high interest to the

systems biology readership.

Reviewer #2 (Remarks to the Author):

Deep proteome and transcriptome mapping of a human cancer cell line
In a tour de force Nagaraj et al. performed a very comprehensive proteome analysis of unsynchronized HeLa cells. The authors identified 10, 255 different human proteins encoded by 9207 genes. In addition to sequencing the proteome of this cell line the authors also performed a deep sequencing experiment to validate the presence of these proteins from the transcriptome data. Based on their results that nearly all the proteins identified have an associated transcript that was sequenced the authors suggest that they have achieved a deep coverage of the functional transcriptome and proteome of this cell type. In addition, they applied iBAQ-based quantification of the detected proteins and correlated these determined quantities with transcriptomics data obtained by RNA-seq on an Ilumina platform. The dataset will be very useful for the community and thus should eventually be published despite a lack of follow-up studies.

We thank the reviewer for his or her positive comments.

However, prior to accepting this publication, we suggest the following revisions and expanded discussions:
The abstract should be revised and reworded; e.g. "... 166,420 sequence unique peptides ..."; "... are within a factor 60 of the median." It is not clear what this means until one reads the article. Thus we suggest adding e.g. "...90 % of them cover less than 4 orders of magnitude and are within a factor of median."

We agree and have changed the abstract to make this clearer. In particular we changed the term “sequence unique peptides” to “peptides with unique amino acid sequence” and we now explicitly refer to “median protein expression level” which was not clear before.

Page 3/1st Results and Discussion Paragraph: in a rather cryptic and incomplete way the authors refer to Experiment 1, which does not really add to the manuscript. Thus, I strongly suggest significantly shortening this paragraph and reducing it to the bottom line that the method was optimized to provide "very deep coverage" with 'only' 12 days of instrument time.

We agree that the description may have been cryptic and confusing and we have now shortened and streamlined it. However, the parts of the paragraph describing the experimental workflow used in both Experiment 1 and Experiment 2 were kept so that it becomes clear to the reader what was involved in obtaining the deep proteomic coverage.

On page 4, the authors indicate that ~500 proteins are 1-peptide identifications. It would be helpful to include a short discussion as to how these proteins were reliably quantified.

In our view, it is difficult to obtain a reliable quantification on 1-peptide identifications. We describe the iBAQ approach in more detail in response to point 2 of reviewer 1 above and already point out the limitations of the abundance estimation for extremely low abundance proteins in the legend to supplementary Table S7.

Page 4/description of the RNA-seq data:
a. How many of the 49 000 transcripts were unique?

The 49,000 transcripts referred to in the manuscript are the transcript set annotated in the Ensembl database version 59. Many genes generate multiple transcripts. Each of these transcripts has a unique transcript structure and thus a unique identifier in Ensembl. In order to clarify it, we now mention in the main text as 49,000 "**unique**" transcripts.

b. Why was FPKM used for normalization of the data as opposed to RPKM?

RPKM refers to a method to normalize by the number of Reads per Kilobase of exon and Million reads whereas FPKM (Fragments per Kiliobase of exon and Million fragments) is an updated measure that takes into account the fact that more than one sequencing read can be made from the same RNA molecule. The first publication discussing RNA-Seq (Mortazavi et al, 2008) analyzed only single read RNA-Seq data and therefore used RPKM. When paired end sequencing was later introduced this measure was updated to FPKM (Trapnell et al, 2010) to take into account the fact that multiple reads can originate from the same cDNA fragment. Though we analyzed single read RNA-Seq data, we use the more general and recent terminology.

c. Is a subset of these transcripts splice variants of the same genes?

Yes, the Ensembl transcript set described in (a) may contain multiple transcript variants of the same gene.

d. Were some of these transcripts ambiguous as to which homolog they may belong to (gene families, etc)?
e. How did the authors deal with homologous genes?

We answer the comments (d) and (e) together.

Read mapping to highly similar genomic loci is handled in the following way: Due to the length of the sequenced RNA fragments, the genomic origin of any read is ambiguous for only ~10% of sequence reads, and less than 1% could originate from more than 6 positions in the genome. Sequence reads were aligned to the human reference genome using TopHat v1.0.13 (Trapnell et al, 2009) which allows up to 40 equally good mappings of a read. In cases where a read can be mapped to multiple transcripts,

each transcript is assigned 1/#mappings in the quantification step. If more than 40 potential mapping locations are identified the read is not considered for quantification.

We added the following to the methods section "**This method allows up to 40 equally good mappings of a read. In cases where a read can be mapped to multiple transcripts, each transcript is assigned 1 per number of mappings in the quantification step. If more than 40 potential mapping locations are identified the read is not considered for quantification.**"

f. Given the data in the following references (Mingyao Li et al. Science 330(6012) (2010) PMID 21596952) Did any mass spectra match to any "non-coding regions" thus identifying new open reading frames?

The indicated reference addresses mis-matches of non-coding RNA and DNA sequences. In our mass spectrometric data we found peptides matching to exons with no ENSEMBL identifiers, but predicted according to GENSCAN, and these are given in supplementary Table S3. Finding completely novel genes in a statistically and biologically robust way based on mis-matches using the proteomic data would require much validation and is beyond the scope of the current work.

Page 5: Given the importance of iBAQ for the manuscript at hand, it would be very beneficial for the readers to have a short description/discussion about this method.

We thank the reviewer for the comment, which was also raised by reviewer 1. In the revised manuscript we have added more iBAQ explanations to the main text (see answer to point 2 of reviewer 1).

Page 5: How were the 37 proteins selected for the determination of the absolute copy numbers? Do the selected proteins show a similar functional/localization distribution as the entire protein set? If not, how do the authors ensure that the determined absolute copy numbers are transferrable to other protein families?

These protein fragments were selected to cover as much of the dynamic range as possible and with no particular preference to any class of proteins or protein families. They originate from quantitative comparison to expressed protein fragments of the PrEST library of M. Uhlen and co-workers. This is now explained in the legend to Table S5.

Page 5/last line: the authors mention that some protein complexes showed lower coverage. How low? For the comparison purposes it would be helpful to have more definite information.

The three molecular complexes that have lower coverage due to cell type specificity and are mentioned in the text have coverage of 20%, 40% and 50%, respectively. We added this information to the manuscript.

Page 6: The authors claim that "..., judged against the coverage achieved by deep-sequencing transcriptomics, the proteomics data was more than 90% complete (...)." Depending on the calculation, we determined values between 73 and 77 % (9207/(11936+598) or 9207/11936). Please clarify!

We thank the reviewer for pointing out this ambiguity. We refer to the coverage of pathways by RNA-Seq vs by proteomics. This is now clarified in the manuscript ("**judged against the coverage of pathways achieved by deep-sequencing transcriptomics**). Reflecting the ambiguity of the gene expression numbers at the transcript and protein basis, we also changed the following sentence to "**Together the transcriptome and proteome data suggest that between at least 10,000 and to 12,000 genes are expressed in HeLa cells**".

Page 6/last line: "...integral membrane proteins account for 25% of the genome but contribute much less to the transcriptome and the proteome (6% of identified proteins)." What percentage of these genes are found in the transcriptome data? We would like this number to make a proper comparison.

We noted a mistake with the labeling of the graph in Figure 3C and which we corrected in the revised version of Figure 3. In terms of category counting, proteins contain 17.55% integral to membrane proteins and transcripts contain 20.43%.