A gene-centered C. elegans protein-DNA interaction network provides a backbone for functional predictions

Juan Fuxman Bass, Carles Pons, Lucie Kozlowski, John Reece-Hoyes, Shaleen Shrestha, Amy Holdorf, Akihiro Mori, Chad Myers and Albertha Walhout

Corresponding author: Albertha Walhout, University of Massachusetts Medical School

Review timeline:
Submission date: 13 June 2016
Editorial Decision: 14 July 2016
Revision received: 12 September 2016
Editorial Decision: 26 September 2016
Revision received: 27 September 2016
Accepted: 28 September 2016

Editor: Maria Polychronidou

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 14 July 2016

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from two of the three referees who agreed to evaluate your study. Reviewer #2 has informed us that s/he is not going to be able to provide comments in a timely manner, so in the interest of time, I prefer to make a decision based on the two available reports. As you will see below, the reviewers acknowledge that the study presents a potentially useful resource for the C. elegans community. However, they raise a number of concerns, which should be carefully addressed in a revision of the manuscript. The recommendations of the referees are clear, so there is no need to repeat the points listed below.

REFEEEREE REPORTS

Reviewer #1:

Bass et al analyze C. elegans protein-DNA interactions using a heterologous yeast one hybrid system and a large collection of transcription factors and promoters. A total of approx 10% of all such interactions are mapped in the study, representing the largest single effort in this area. The authors proceed to validate their dataset bioinformatically, making many interesting predictions of protein function. They also validate some predictions experimentally, perhaps most interestingly they find that many C. elegans nuclear hormone receptors are required for efficient xenobiotic response.

The manuscript is clearly written and informative, and the conclusions are justified by the evidence.
The bioinformatic methods used are well applied, and the biological findings are interesting, albeit somewhat preliminary. The main strength of the work is the provided resource that complements work based on other methods, and allows systems level analyses of gene regulation in C. elegans.

Specific points
- Despite the large dataset, the authors do not make any effort to analyze sequence-function relationships. Such analyses should be included, even if the analyses fail to differentiate between Y1H active elements that have ChiPseq peaks or motifs and active elements that do not etc.  
- Authors should indicate effect sizes in addition to the p-values in the text. Also, the focus is on predictions based on aggregate data, and it is not clear what is the confidence of individual edges. This could be estimated and discussed.
- p4 top authors discuss that presence of binding site is poor predictor of binding in vivo. However, binding is also poor predictor of transcriptional activity. Does binding in Y1H equal activity? To what extent does Y1H help to predict which binding events are activating? The authors should comment on this.
- p4 bottom authors should discuss the area more broadly, including the work in sea urchin to map regulatory interactions, and new massively parallel reporter assays etc
- p8 many promoters were removed as they conferred high background. The authors should analyze the sequences, and test whether these contain binding sites for yeast factors etc. Also, can the method not be modified to detect repressors?
- p14 authors validate findings using RT-PCR. It is unclear whether the target genes tested were included in the original discovery set.
- p17 clarify the concepts dauer constitutive and dauer defective
- p20-21 authors argue that factors with different targets are essential because of lack of redundancy. But their example is confusing and not consistent with their earlier idea that nuclear hormone receptors respond to environment (and thus allow different inputs to regulate the same genes). Both inputs and outputs need to be equivalent in order for something to be redundant.
- p24 last sentence of first paragraph is incorrect, bifunctional factors can be detected with less sensitivity
- some figure panels (1a, 1c, 3a, 7b) are uninformative, or contain excess data that is not central for understanding of the point.

Reviewer #3:

Summary and general assessment:

This report from the Walhout group focuses on an extensive analysis of transcription factor-promoter interactions in C. elegans, leading to the identification of several thousand of such interactions via the yeast one-hybrid system. Although this is an in vitro assay, the authors validate the relevancy of their findings by comparison with independent large scale datasets such as ChiPseq and gene expression. They build a network of these interactions, and assign activator or repressor functions to a subset, using the yeast one-hybrid data in combination with other existing datasets. The authors then test whether target genes share functions, whether related TFs show more similar patterns, etc. Overall, this analysis provides a substantial new resource of information for others to mine for potential regulatory interactions and greatly expands our first-level understanding of TF binding to candidate promoter sequences. The analysis is generally well done (see concerns below), and the data is presented in such a way as to facilitate others utilizing it.

Major points:

Why were so few regulatory interactions predicted from the analysis described in Figure 3b - 44/153? This seems low and is never explained by the authors. They simply say it's more than random expectation. It is not surprising that they were able to achieve greater than random, but that is not sufficient to say this was a robust analysis. Is there a reason why so many factors did not yield predictive interactions from the y1h? The data provided do not allow for a good judgment regarding the reliability and robustness of the network. Moreover, the authors frequently use percentages throughout the subsequent analysis of this PDI network, but by the time 44 factors are parsed out, the actual N is often very low. This also leads to some worry that the interaction data is less robust than one would hope.
The effects on target gene regulation are underwhelming for activators and repressors, and it is not really clear why this is so. The authors list some possibilities in the discussion, but no obvious explanation why so few genes change expression - perhaps it is the stage at which the RNA was collected or the fact that it was collected from whole animals, missing or masking the regulatory event. Moreover, the methods describe different stages of development selected for different RNA analyses without really explaining why such stages were chosen. For example, when looking at daf-19 targets, they looked in embryos when the phenotype is a larval dauer phenotype. This all should be made clearer.

Minor points:

Intro line 2: comma after "stresses" (necessary because you are listing phrases, not just words)

Intro p5: The sentences that equate compact genome with "few distal enhancers" should be re-written. A compact genome does not "therefore" rule out the possibility of distal enhancers. No one has actually systematically, experimentally, looked for distal enhancers in C. elegans, so it is difficult to say whether they are rare, non-existent or frequent.

Results p7: Although previously published, perhaps a sentence about the genes represented in the promoterome would be useful to the reader - do they tend to encode particular functions? chromosome locations?

Results p10: Reference(s) needed for the "compendium of co-expression data from 123...datasets"

Results p12: The analysis in which activators and repressors are shown to have more similar expression patterns with their targets than non-targets (Fig 3F) seems problematic, as there are so few targets (usually around 10 or so) than non-targets (~3000). Therefore the comparison between targets and nontargets is extremely lopsided, and there could be many reasons why the non-targets have less correlated expression.

Results p 21, and again in Discussion, p 24: "identify TF pairs that bind overlapping sets of targets in eY1H assays and that, therefore, likely bind similar DNA sequences". It is not clear why factors that bind at the same genes likely bind similar consensus elements. It's entirely possible that they bind unrelated sequences to coordinately regulate target gene expression. This statement seems conflated.

I do not understand the methods for the coexpression network. I read this as starting with a set of 123 global gene expression experiments, some of which were removed for various reasons. Why would such a gene expression experiment only have 5 genes in it, or need at least 4 experiments? I don't follow at all what is meant here.

Why use 10 promoter binding events as the cutoff for inclusion of a TF in the network? What happens if this cutoff is raised or lowered? The authors need to provide justification for this choice in boundary.

The terms "out degree" or "in degree" in figure 1c are unnecessarily confusing. The legend provides an explanation but it is not easy to figure out why these phrases are used to denote the number of promoters bound by a TF, and the number of TFs that bind a given promoter. This seems like excess jargon.

We thank the reviewers and Editor for their helpful comments and suggestions that helped us to further improve our manuscript.

Reviewer #1:
Bass et al analyze C. elegans protein-DNA interactions using a heterologous yeast one hybrid system and a large collection of transcription factors and promoters. A total of approx 10% of all such interactions are mapped in the study, representing the largest single effort in this area. The authors proceed to validate their dataset bioinformatically, making many interesting predictions of protein function. They also validate some predictions experimentally, perhaps most interestingly they find that many C. elegans nuclear hormone receptors are required for efficient xenobiotic response.

The manuscript is clearly written and informative, and the conclusions are justified by the evidence. The bioinformatic methods used are well applied, and the biological findings are interesting, albeit somewhat preliminary. The main strength of the work is the provided resource that complements work based on other methods, and allows systems level analyses of gene regulation in C. elegans.

Specific points

- Despite the large dataset, the authors do not make any effort to analyze sequence-function relationships. Such analyses should be included, even if the analyses fail to differentiate between Y1H active elements that have ChIPseq peaks or motifs and active elements that do not etc.

To further show the quality of our PDI network, we evaluated motif enrichments in the promoter sequences of eY1H positive targets using CisFinder (PMID: 19740934). For each TF, the 500 bp proximal promoter sequences of eY1H targets was used to identify motifs enriched compared to the sequences for eY1H non-targets (promoters not bound by that TF, but present in the overall network). Elementary motifs were obtained using an FDR of 0.05 and a motif enrichment greater than 2. Motifs were determined for the 77 TFs that have 50 or more eY1H targets in the PDI network to avoid obtaining low-quality motifs. These motifs are included in Dataset EV2. For the 25 TFs for which PBM data was available we compared the motif logo derived from eY1H data with the one derived from PBMs. Statistical significance between the eY1H- and PBM-derived motifs was determined using the TOMTOM software version 4.11.2 (PMID: 17324271). Twenty-one of 25 motifs were statistically significant showing that eY1H can be used for de novo motif discovery (shown in Figure EV1). These results are discussed in the revised manuscript (page 9 bottom, methods, Figure EV1 and Dataset EV2).

Comparison of PBM-derived motif enrichment between eY1H and ChIP data has been recently reported (PMID: 25905672). This study, which was performed using the same eY1H dataset included here, found that motifs are more frequently enriched in eY1H data than in ChIP-seq data, further supporting the quality of our dataset.

Comparison of de novo motif predictions with ChIP data was more challenging since there was only one TF (LSY-2) with enough eY1H targets to perform the enrichment analyses, and with PBM data with which to compare our predictions. Overall, the motifs predicted based on eY1H positive sequences (both positive or negative for ChIP interactions) was similar to the motif derived from PBM data (see Rebuttal Figure 1). Motif prediction using ChIP positive and eY1H negative interactions did not match the reported motif. This further supports the high quality of the eY1H data. However, given that these results are based on only one TF, we did not include them in the revised manuscript.
Authors should indicate effect sizes in addition to the p-values in the text. Also, the focus is on predictions based on aggregate data, and it is not clear what is the confidence of individual edges. This could be estimated and discussed.

P-values are indicated in the figures and referenced in figure legends. Effect sizes can be estimated in each of the figures. Although it would be of great use, quantitative confidence on individual edges cannot be estimated with the current knowledge in the field as it depends on multiple variables whose relative contribution have not been determined yet.

- p4 top authors discuss that presence of binding site is poor predictor of binding in vivo. However, binding is also poor predictor of transcriptional activity. Does binding in Y1H equal activity? To what extent does Y1H help to predict which binding events are activating? The authors should comment on this.

The reviewer is right to point out that Y1H binding does not necessarily imply transcriptional activity. In a previous work, we found that several eY1H interactions are regulatory but others are not, at least in the one tissue / condition and stage tested (PMID: 26430702). We added a sentence in the introduction (page 5, bottom): “Further, we have found that eY1H interactions could be validated in animals harboring transcriptional fusion reporter constructs fed with bacteria expressing RNAi clones against different TFs”. Further, in the present manuscript we found that 30-60% of interactions tested in TF mutant animals (9-34 targets per TF tested in seven TF mutant animals) are regulatory, even when a single stage and condition was evaluated. This is discussed in Figures 3G, 5B and 8B.

- p4 bottom authors should discuss the area more broadly, including the work in sea urchin to map regulatory interactions, and new massively parallel reporter assays etc

A discussion on regulatory interaction mapping in different organisms, and on MPRAs have been added to the introduction (page 3):

“Gene regulatory networks have been extensively studied at a small or medium scale in different metazoan organisms including nematodes, sea urchins, fruit flies, and mammals. For instance, GRNs regulating development in the sea urchin Stronglylocentrotus purpuratus embryo have been delineated based on spatiotemporal expression patterns of TFs and signaling molecules, ChIP-seq data and gene knockdown (reviewed in 26962438). Similarly, developmental GRNs have been studied in Drosophila melanogaster using genetic perturbations, expression profiling, transgenic reporters, ChIP-seq data, and mathematical modeling (reviewed in 19874814). More recently, massively parallel reporter assays have been used for an in depth study of the function of regulatory sequences in cell lines, mouse tissues, yeast and bacteria (reviewed in 26072432).”

- p8 many promoters were removed as they conferred high background. The authors should analyze the sequences, and test whether these contain binding sites for yeast factors etc. Also, can the method not be modified to detect repressors?
eY1H assays can detect both activators and repressors; it solely detects binding. The assay is based on binding of a TF that is fused to the strong yeast Gal4p transcriptional activation domain. Thus, even repressors will activate in yeast in this context. We agree that high background may be caused by yeast TFs binding and activating some promoters. However, a comprehensive analysis of yeast TF binding sites would require extensive effort (especially the follow-up testing – which would be needed to added meaning to the analysis) is beyond the scope of the paper.

- p14 authors validate findings using RT-PCR. It is unclear whether the target genes tested were included in the original discovery set.

All the targets tested by RT-qPCR were identified in the eY1H dataset. This is mentioned in the text in each example. For instance: “We evaluated the expression of ten CEH-17 eY1H targets...”

- p17 clarify the concepts dauer constitutive and dauer defective

Dauer constitutive: A mutant phenotype where animals form dauer larvae under conditions that would not induce dauer formation in wild type animals.

Dauer defective: A mutant phenotype where animals are unable to form dauer larva under conditions that would induce dauer formation in wild type animals.

Given that these phenotypes are well known in the C. elegans community, we have not included the definitions in the manuscript.

- p20-21 authors argue that factors with different targets are essential because of lack of redundancy. But their example is confusing and not consistent with their earlier idea that nuclear hormone receptors respond to environment (and thus allow different inputs to regulate the same genes). Both inputs and outputs need to be equivalent in order for something to be redundant.

We agree with the reviewer that redundancy in binding does not necessarily imply redundancy in function (regulation by different inputs, in different tissues/stages or different regulatory activity influence redundancy in function). As such, each edge in the association network should not be interpreted as redundancy between the TFs involved. However, when analyzed globally we detected an enrichment for non-essential TF among those that are connected to 2+ other TFs validating our overall conclusions based on the aggregated data. The limitations of this analysis are discussed in page 26:

“In addition, many TF-pairs connected in the association network could have unrelated functions if they are expressed in different tissues/conditions/stages, and/or respond to different environmental cues as we have observed for NHR mutant animals exposed to different xenobiotics. Overall, these observations highlight the complexity in C. elegans gene regulation and the importance of integrating multiple functional datasets.”

- p24 last sentence of first paragraph is incorrect, bifunctional factors can be detected with less sensitivity

The analyses performed to predict regulatory functions assign as activator to TFs that have higher coexpression with targets than with non-targets and vice versa for repressors. Given this binary assignment we would miss bifunctional TFs. We modified this sentence to clarify this point:

“Finally, given that predictions are based on the comparison between TF coexpression with targets and non-targets (higher or lower), we cannot identify bifunctional TFs using this approach. Therefore, TFs that can function as both activators and repressors would have been missed in this analysis.”

- some figure panels (1a, 1c, 3a, 7b) are uninformative, or contain excess data that is not central for understanding of the point.

We respectfully disagree and do consider these figures to be informative and relevant for the paper.
1A: Although some of these numbers are referenced in the text, the flowchart helps understand where the data is coming from.

1C: This figure condenses information on the distribution of the interactions detected as well as TF and promoter connectivity. We believe that presenting network structure will be useful for the systems biology community studying biological networks.

3A: This figure, although it does not convey information, briefly presents the rationale for the method employed in the entire figure.

7B: This figure presents the data from which aggregated analyses in Figures 7C-E are based. In addition, the expanded view of this figure (Figure EV3) show TF-pairs that may be functionally related (redundant, antagonistic, etc.) that could be explored by the C. elegans community for follow up experiments.

Reviewer #3:

Summary and general assessment:

This report from the Walhout group focuses on an extensive analysis of transcription factor-promoter interactions in C. elegans, leading to the identification of several thousand of such interactions via the yeast one-hybrid system. Although this is an in vitro assay, the authors validate the relevancy of their findings by comparison with independent large scale datasets such as ChIPseq and gene expression. They build a network of these interactions, and assign activator or repressor functions to a subset, using the yeast one-hybrid data in combination with other existing datasets. The authors then test whether target genes share functions, whether related TFs show more similar patterns, etc. Overall, this analysis provides a substantial new resource of information for others to mine for potential regulatory interactions and greatly expands our first-level understanding of TF binding to candidate promoter sequences. The analysis is generally well done (see concerns below), and the data is presented in such a way as to facilitate others utilizing it.

Major points:

Why were so few regulatory interactions predicted from the analysis described in Figure 3b-44/153? This seems low and is never explained by the authors. They simply say it's more than random expectation. It is not surprising that they were able to achieve greater than random, but that is not sufficient to say this was a robust analysis. Is there a reason why so many factors did not yield predictive interactions from the y1h? The data provided do not allow for a good judgment regarding the reliability and robustness of the network. Moreover, the authors frequently use percentages throughout the subsequent analysis of this PDI network, but by the time 44 factors are parsed out, the actual N is often very low. This also leads to some worry that the interaction data is less robust than one would hope.

We agree with the reviewer that although the number of predictions is encouraging we were not able to provide predictions for many TFs. We thoroughly discuss this in the Discussion section:

“For instance, by integrating the PDI network with a compendium of expression profiling data we provide predictions for the regulatory functions of 44 TFs, which are largely consistent with published regulatory functions and those predicted based on TF-cofactor interactions. This number of predictions is encouraging, considering that there are several reasons that can explain lack of correlation between TF and target gene expression. First, our predictions are based on the assumption that there is a strong correlation between TF mRNA expression and TF activity. However, this may not be the case for many TFs whose regulatory function is modulated by ligand binding, post-translational modifications or dimerization with other factors (Walhout, 2011). For instance, although we provide regulatory predictions for 29% of the TFs evaluated, we did so only for 15% of the NHRs tested, a family known to be modulated by ligands and that frequently forms heterodimers. Second, predictions for some TFs may be missed due to redundancy as changes in the expression of a single TF may not be sufficient to affect the expression of its target gene. Finally, given that predictions are based on the comparison between TF co-expression with targets and non-
targets (higher or lower), we cannot identify bifunctional TFs using this approach. Therefore, so many TFs that can function as both activators and repressors would have been missed in this analysis."

The main reason for not mentioning this in Results is that the interpretation is supported by results discussed in later figures in the paper.

Given the comments above, the fact that we did not provide functional predictions for many TFs does not stem from a lack of robustness in the dataset but from the complexity of functional relationships. Indeed, the high quality of the dataset is illustrated by the overlap with existing datasets and the multiple functional validations performed in Figures 3, 5, 6 and 8.

The effects on target gene regulation are underwhelming for activators and repressors, and it is not really clear why this is so. The authors list some possibilities in the discussion, but no obvious explanation why so few genes change expression - perhaps it is the stage at which the RNA was collected or the fact that it was collected from whole animals, missing or masking the regulatory event.

We were able to ‘validate’ 30-60% of the interactions tested in TF mutant strains. We consider this number to be encouraging for multiple reasons:

- Functional validations were performed in a single stage and condition and using whole animals. Therefore, testing more conditions/stages or measuring expression changes in tissues (rather than in whole animals) will likely ‘validate’ more of the interactions.
- Redundancy with different TFs may mask the effect of a particular TF in gene regulation.
- Not all TF binding is regulatory in vivo, even for interactions detected by ChiP, as previously reported (PMID: 26430702).

We agree with the reviewer about discussing these points, which are now in the Discussion section (page 24, bottom).

Moreover, the methods describe different stages of development selected for different RNA analyses without really explaining why such stages were chosen. For example, when looking at daf-19 targets, they looked in embryos when the phenotype is a larval dauer phenotype. This all should be made clearer.

Stages were chosen based on when TFs and target genes are expressed. For example, daf-19 is most highly expressed at late stages in embryo development and studies looking at targets of DAF-19 were performed at this stage (PMID: 21740898). Developmental defects occurring at embryonic stages can result in phenotypes that are evidenced at later stages (chemotaxis, dauer, osmotic avoidance). This is now clarified on page 18 of the revised manuscript.

Minor points:

Intro line 2: comma after "stresses" (necessary because you are listing phrases, not just words)

Corrected.

Intro p5: The sentences that equate compact genome with "few distal enhancers" should be re-written. A compact genome does not "therefore" rule out the possibility of distal enhancers. No one has actually systematically, experimentally, looked for distal enhancers in C. elegans, so it is difficult to say whether they are rare, non-existent or frequent.

We removed the reference to distal enhancers. The sentence now reads as: “Therefore, most gene regulation likely occurs through proximal gene promoters.”

Results p7: Although previously published, perhaps a sentence about the genes represented in the promoterome would be useful to the reader - do they tend to encode particular functions? chromosome locations?

The following sentence was added in page 8: “This collection comprises ~5,500 promoter regions of 0.3-2 kb of genes encoding for different molecular functions, and that are distributed across all C.
"Instead of 123 gene expression experiments, we used only 101 gene expression datasets that are considered high-confidence datasets.

We thank the reviewer for pointing this out. Pearson correlation, the metric we used for measuring similarity of gene pairs' expression profiles is not robust when computed over too short a profile (i.e. the number of expression datasets). As some expression datasets have very few samples, we required genes to appear in at least 25 of the 101 expression datasets with 4 or more samples in each, which we defined as high-confidence expression datasets. Genes whose co-expression values were calculated with fewer high-confidence expression datasets were not used in our analysis. We agree that the "5 genes" statement is confusing-- we had a small number of datasets for which most of the genes were filtered for technical reasons. These datasets are of course not informative and did not contain information for any of our genes in the PDI network. We have removed this statement and just indicate that the 101 datasets with a minimum of four experiments and expression measurements for genes in our PDI network were used for the co-expression analysis.

Why use 10 promoter binding events as the cutoff for inclusion of a TF in the network? What happens if this cutoff is raised or lowered? The authors need to provide justification for this choice in boundary.
The choice of boundary is based on sample size to be able to detect statistically significant differences between co-expression with targets and non-targets. In the revised manuscript we included and analysis changing that boundary to 5, 8, 15 and 20 (Figure 3B, bottom), obtaining similar results compared to the boundary of 10 PDIs.

The terms "out degree" or "in degree" in figure 1c are unnecessarily confusing. The legend provides an explanation but it is not easy to figure out why these phrases are used to denote the number of promoters bound by a TF, and the number of TFs that bind a given promoter. This seems like excess jargon.

These are standard terms in the field. Having the explanation in the figure would make it harder to read. The definitions are included the figure legend.

2nd Editorial Decision 26 September 2016

Thank you again for sending us your revised manuscript. We have now heard back from reviewer #1 and as you will see below, s/he thinks that the reviewers' concerns have been addressed. However, s/he requests adding a sentence in the discussion regarding the biological significance of the TF-p promoter interactions (see comments below).

REFEREE REPORTS

Reviewer #1:

The authors have addressed most of my criticisms. The remaining issue is the reliability of the individual predictions, which is also pointed out by the other referee. To address this, I would suggest that the authors add statement to the discussion that more explicitly states that the validation experiments establish that the Y1H network is enriched in biologically relevant interactions, but do not establish biological significance of any individual interaction outside those that are validated with the follow up experiments themselves.

2nd Revision - authors' response 27 September 2016

Response to Reviewer #1:

The authors have addressed most of my criticisms. The remaining issue is the reliability of the individual predictions, which is also pointed out by the other referee. To address this, I would suggest that the authors add statement to the discussion that more explicitly states that the validation experiments establish that the Y1H network is enriched in biologically relevant interactions, but do not establish biological significance of any individual interaction outside those that are validated with the follow up experiments themselves.

Thank you very much for the final comments. We respectfully argue that we have discussed at great length the topic raised once again, see the following in the Discussion:

Indeed, 30-60% of the PDIs tested in TF mutant animals resulted in changes in the expression of target genes. This result is similar to what has been previously reported for Y1H-derived interactions in Arabidopsis thaliana and Droshophila melanogaster (Brady et al, 2011; Hens et al, 2011). There are several reasons why we failed to detect a regulatory role for some PDIs detected by eY1H (Walhout, 2011). First, some PDIs identified in yeast may not occur in vivo. Second, some physical interactions, although occurring in vivo, may be neutral and have no regulatory consequence (MacNeil et al, 2015). Third, some regulatory interactions may have been missed as
they were tested in a single stage and condition, and because expression changes were measured in whole animals losing tissue resolution. Finally, some regulatory interactions may have been missed due to redundancy with other TFs.

We don’t feel that stating another point as indicated by the reviewer will do justice to the paper, we are up front and honest and open throughout.

3rd Editorial Decision 28 September 2016

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.
### A. Figures

1. **Data**

   - The data shown in figures should satisfy the following conditions:
     - the data were obtained and processed according to the field’s best practice and are presented to reflect the results of the experiments in a scientifically meaningful way.
     - figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
     - graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
     - if n < 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified.
     - Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the authorship guidelines on Data Presentation.

2. **Captions**

   - Each figure caption should contain the following information, for each panel where they are relevant:
     - a specification of the experimental system investigated (e.g. cell line, species name).
     - the assay(s) and method(s) used to carry out the reported observations and measurements.
     - an explicit mention of the biological and chemical entities that are being measured.
     - an explicit mention of the biological and chemical entities that are altered/changed/modified in a controlled manner.
     - the exact sample size (n) for each experimental group/condition, given as a number, not a range.
     - a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.)
     - a statement of how many times the experiment shown was independently replicated in the laboratory.
     - definitions of statistical methods and measures:
       - common tests, such as t-test (please specify whether paired or unpaired), single t-tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section.
       - are tests one-sided or two-sided?
       - are there adjustments for multiple comparisons?
       - exact statistical test results, e.g., P-value < 0.05 but P-value < 0.01?
       - definition of “center values” as median or average;
       - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

In the pink boxes below, provide the page number(s) of the manuscript draft or figure legend(s) where the information can be located. Every question should be answered. If the question is not relevant to your research, please write NA (non-applicable).

### B. Statistics and general methods

<table>
<thead>
<tr>
<th>Question</th>
<th>Page/NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?</td>
<td>NA</td>
</tr>
<tr>
<td>1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used</td>
<td>NA</td>
</tr>
<tr>
<td>2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?</td>
<td>NA</td>
</tr>
<tr>
<td>3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.</td>
<td>NA</td>
</tr>
<tr>
<td>For animal studies, include a statement about randomization even if no randomization was used.</td>
<td>NA</td>
</tr>
<tr>
<td>4. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.</td>
<td>NA</td>
</tr>
<tr>
<td>5. For animal studies, include a statement about blinding even if no blinding was done</td>
<td>NA</td>
</tr>
<tr>
<td>6. For every figure, are statistical tests justified as appropriate?</td>
<td>Yes, standard statistical tests were used in each figure</td>
</tr>
<tr>
<td>7. Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.</td>
<td>Yes, most data analysis was done using non-parametric tests or randomisations. Student's t-test was used on log-transformed expression data.</td>
</tr>
<tr>
<td>8. Is there an estimate of variation within each group of data?</td>
<td>NA, when parametric tests were used</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1.</td>
<td>To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Abhubedi (see link list at top right), IDTechmedia (see link list at top right).</td>
</tr>
<tr>
<td>2.</td>
<td>Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.</td>
</tr>
<tr>
<td>3.</td>
<td>Dual use research of concern</td>
</tr>
<tr>
<td>4.</td>
<td>Data Accessibility</td>
</tr>
<tr>
<td>5.</td>
<td>Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.</td>
</tr>
<tr>
<td>6.</td>
<td>For experiments involving live vertebrates, include a statement of compliance with ethical regulators and identify the committee(s) approving the experiments.</td>
</tr>
<tr>
<td>7.</td>
<td>We recommend consulting the ARRIVE guidelines (see link list at top right), J. Exp. Biol. 200, 1 (2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under Reporting Guidelines. See also NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.</td>
</tr>
<tr>
<td>8.</td>
<td>Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.</td>
</tr>
<tr>
<td>9.</td>
<td>To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Abhubedi (see link list at top right), IDTechmedia (see link list at top right).</td>
</tr>
<tr>
<td>10.</td>
<td>We recommend consulting the ARRIVE guidelines (see link list at top right), J. Exp. Biol. 200, 1 (2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under Reporting Guidelines. See also NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.</td>
</tr>
<tr>
<td>11.</td>
<td>Identify the committee(s) approving the study protocol.</td>
</tr>
<tr>
<td>12.</td>
<td>Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WHA Declaration of Helsinki and the Department of Health and Human Services Belmore Report.</td>
</tr>
<tr>
<td>13.</td>
<td>For publication of patient photos, include a statement confirming that consent to publish was obtained.</td>
</tr>
<tr>
<td>14.</td>
<td>Report any restrictions on the availability (and/or on the use) of human data or samples.</td>
</tr>
<tr>
<td>15.</td>
<td>Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.</td>
</tr>
<tr>
<td>16.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>17.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
<tr>
<td>18.</td>
<td>Provide accession codes for deposited data. See author guidelines, under Data Deposition.</td>
</tr>
<tr>
<td>20.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>21.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>22.</td>
<td>Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access controlled repositories such as dbGaP (see link list at top right) or EGA (see link list at top right).</td>
</tr>
<tr>
<td>23.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>24.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
<tr>
<td>25.</td>
<td>We recommend consulting the ARRIVE guidelines (see link list at top right), J. Exp. Biol. 200, 1 (2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under Reporting Guidelines. See also NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.</td>
</tr>
<tr>
<td>26.</td>
<td>Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WHA Declaration of Helsinki and the Department of Health and Human Services Belmore Report.</td>
</tr>
<tr>
<td>27.</td>
<td>For publication of patient photos, include a statement confirming that consent to publish was obtained.</td>
</tr>
<tr>
<td>28.</td>
<td>Report any restrictions on the availability (and/or on the use) of human data or samples.</td>
</tr>
<tr>
<td>29.</td>
<td>Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.</td>
</tr>
<tr>
<td>30.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>31.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
<tr>
<td>32.</td>
<td>Provide accession codes for deposited data. See author guidelines, under Data Deposition.</td>
</tr>
<tr>
<td>33.</td>
<td>Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Cryoelectronographic data for small molecules d. Functional genomic data e. Proteomic and molecular interactions</td>
</tr>
<tr>
<td>34.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>35.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>36.</td>
<td>Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access controlled repositories such as dbGaP (see link list at top right) or EGA (see link list at top right).</td>
</tr>
<tr>
<td>37.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>38.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
<tr>
<td>39.</td>
<td>We recommend consulting the ARRIVE guidelines (see link list at top right), J. Exp. Biol. 200, 1 (2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under Reporting Guidelines. See also NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.</td>
</tr>
<tr>
<td>40.</td>
<td>Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WHA Declaration of Helsinki and the Department of Health and Human Services Belmore Report.</td>
</tr>
<tr>
<td>41.</td>
<td>For publication of patient photos, include a statement confirming that consent to publish was obtained.</td>
</tr>
<tr>
<td>42.</td>
<td>Report any restrictions on the availability (and/or on the use) of human data or samples.</td>
</tr>
<tr>
<td>43.</td>
<td>Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.</td>
</tr>
<tr>
<td>44.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>45.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
<tr>
<td>46.</td>
<td>Provide accession codes for deposited data. See author guidelines, under Data Deposition.</td>
</tr>
<tr>
<td>47.</td>
<td>Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Cryoelectronographic data for small molecules d. Functional genomic data e. Proteomic and molecular interactions</td>
</tr>
<tr>
<td>48.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>49.</td>
<td>Deposition is strongly recommended for any datasets that are central and integral to the study, please consider the journal's data policy. If it is not structured to public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under Expanded View) or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).</td>
</tr>
<tr>
<td>50.</td>
<td>Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access controlled repositories such as dbGaP (see link list at top right) or EGA (see link list at top right).</td>
</tr>
<tr>
<td>51.</td>
<td>For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under Reporting Guidelines. Please confirm you have submitted this list.</td>
</tr>
<tr>
<td>52.</td>
<td>For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under Reporting Guidelines. Please confirm you have followed these guidelines.</td>
</tr>
</tbody>
</table>

* For all hyperlinks, please see the table at the top right of the document.