Cell Systems

Transparent Peer Review Record

A proteogenomics data-driven knowledge base of human cancer

Yuxing Liao, Sara R. Savage, Yongchao Dou, Zhiao Shi, Xinpei Yi, Wen Jiang, Jonathan T. Lei, Bing Zhang

Summary

Initial Submission: Received March 2, 2023

Scientific editor: Bernadett Gaal, DPhil

First round of review: Number of reviewers: 3
3 confidential, 0 signed
Revision invited April 11, 2023
Minor changes anticipated
Revision received May 11, 2023

Second round of review: Number of reviewers: 2
2 original, 0 new
2 confidential, 0 signed
Accepted: July 25, 2023

Cell Systems Transparent Peer Review Record

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Zhang,

I hope this email finds you well. The reviews are back on your manuscript and I've appended them below. You'll see that the reviewers find the manuscript compelling and their comments are intended to strengthen an already strong piece of work. We're happy to invite a revision.

If you have any questions or concerns about the revision, I'd be happy to talk about them, either over email or by phone. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Bernadett

Bernadett Gaal, DPhil
Editor-in-Chief, *Cell Systems*

Reviewers' comments:

Reviewer #1: This manuscript by Liao Y. presents LinkedOmicsKB, an online resource consisting of ~40,000 precomputed gene-, protein-, mutation-, and phenotype-centric web pages, based on the analysis results using the harmonized pan-cancer proteogenomic profiles from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Its utilities to explore these complex and interconnected resources were demonstrated by three case studies, to reveal novel insights for an understudied druggable protein, TP53 mutations, and cancer-related phenotypes. Overall, it is well-designed and user-friendly, which could be a valuable resource for the cancer community to discover clues for different questions. However, there are some improvements and clarifications needed and listed below.

Major issues:
1. Although the current version of LinkedOmicsKB includes over 1000 tumors across 10 cancer types, it would be useful to keep updated regularly as more proteogenomic studies are expected to get published in the future. Therefore, it is important to provide the updated versions of LinkedOmicsKB after the current initial version, and the update plan would be mentioned in the discussion section. Moreover, it is necessary to provide a document page on the website introducing the data sources and summary statistics for each version, including the current version.
2. Figure descriptions is lacking for many plots on the website. Although it is mentioned somewhat in the manuscript and bioinformatics veterans may understand from the legend panels, it may be inadequate and inconvenient for users who are not that familiar like clinicians. Thus, it is

recommended to provide the descriptions for most plots in the appropriate places in the web pages.
3. Regarding genome and transcriptome data pre-processing, are there standard pipelines utilized in CPTAC, as the one used for processing and quantifying the mass spectrometry data? Will the mutation calling results be comparable with the original publications?

Minor issues:
The current version of LinkedOmicsKB was analyzed on the level of cancer types. If multiple studies for a same cancer type was included in the future version, how to integrate the different studies of a cancer type?

Reviewer #2: In this manuscript, Liao et al, present a bioinformatics online platform for analyzing and visualizing pan-cancer proteogenomics data and precomputed analytic results. To evaluate its performance and clinical utility, the authors specifically focused on three case studies, i.e. using the platform to analyze (i) understudied druggable proteins, (ii) clinical phenotypes, and (iii) TP53 mutations. The bioinformatics platform is well-designed. Its interface is intuitive and user-friendly. The analytic and visualization modules are comprehensive and very useful. It is expected that the platform will highly benefit the cancer research field. Here, the reviewer only has a few minor comments.

1) Please indicate which test was used to calculate the p-values in Fig. 4i and 4j.
2) In Fig5b, in addition to the labels, please also highlight the data points in the scatter plot.
3) For all the scatter plots, please also add p-values and correlation coefficients and indicate them in the main text (at least for the main figures). Only showing the directions (negative or positive) in the main text is not enough.

Reviewer #3: The authors have created a web-based knowledge base called LinkedOmicsKB, which is built upon CPTAC's pan-cancer proteogenomics data. The primary objective of LinkedOmicsKB is to transform this data into meaningful and intelligible insights into cancer biology. Through three case studies, the authors demonstrate how LinkedOmicsKB can be used to develop novel hypotheses on genes, phosphorylation sites, somatic mutations, and cancer phenotypes.

The paper is well-written, and I enjoyed reading it. I only have one comment to make. The author stated that the data were recently harmonized and referred to Fig1a and Methods. However, I found that neither the figure nor the methods section provided a detailed description of how the harmonization was carried out.

Authors' response to the reviewers' first round comments

Attached.

Cell Systems

Transparent Peer Review Record

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Zhang,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Bernadett

Bernadett Gaal, DPhil
Editor-in-Chief, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total!

Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Manuscript Text:

- We don't allow "priority claims" (e.g. new, novel, etc.). For a discussion of why, read: http://crosstalk.cell.com/blog/getting-priorities-right-with-novelty-claims, http://crosstalk.cell.com/blog/novel-insights-into-priority-claims.
- Please check that you use the word "significantly" in the statistical sense only.

STAR Methods: Note that Cell Press has recently changed the way it approaches "availability" statements for the sake of ease and clarity. Please revise the first section of your STAR Methods as follows, noting that the particular examples used might not pertain to your study. Please consult the STAR Methods guidelines for additional information.

RESOURCE AVAILABILITY

Lead Contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jane Doe (janedoe@qwerty.com).

Materials Availability: This study did not generate new materials. *-OR-* Plasmids generated in this study have been deposited at [Addgene, name and catalog number]. *-OR-* etc.

Data and Code Availability:

- **Source data statement** (described below)
- **Code statement** (described below)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Data and Code Availability statements **have three parts and each part must be present. Each part should be listed as a bullet point, as indicated above.**

Instructions for section 1: Data. The statements below may be used in any number or combination, but at least one must be present. They can be edited to suit your circumstance. Please ensure that all datatypes reported in your paper are represented in section 1. For more information, please consult this list of standardized datatypes and repositories recommended by Cell Press.

• [Standardized datatype] data have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

• [De-identified human/patient standardized datatype] data have been deposited at [datatype-specific repository]. They are publicly available as of the date of publication until [date or delete "until"]. Accession numbers are listed in the key resources table.

• [De-identified human/patient standardized datatype] data have been deposited at [datatype-specific repository], and accession numbers are listed in the key resources table. They are available upon request until [date or delete "until"] if access is granted. To request access, contact [insert name of governing body and instructions for requesting access]. [Insert the following when applicable] In addition, [summary statistics describing these data/processed datasets derived from these data] have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. These accession numbers are also listed in the key resources table.

• Raw [standardized datatype] data derived from human samples have been deposited at [datatype-specific repository], and accession numbers are listed in the key resources table. Local law prohibits depositing raw [standardized datatype] datasets derived from human samples outside of the country of origin. Prior to publication, the authors officially requested that the raw [adjective] datasets reported in this paper be made publicly accessible. To request access, contact [insert name of governing body and instructions for requesting access]. [Insert the following when applicable] In addition, [summary statistics describing these data/processed datasets derived from these data] have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. These accession numbers are also listed in the key resources table.

• This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

Instructions for section 2: Code. The statements below may be used in any number or combination, but at least one must be present. They can be edited to suit your circumstance. ***If you are using GitHub, please follow the instructions here to archive a "version of record" of your GitHub repo at Zenodo, then report the resulting DOI. Additionally, please note that the Cell Systems strongly recommends that you also include an explicit reference to any scripts you may have used throughout your analysis or to generate your figures within section 2.***

• All original code has been deposited at [repository] and is publicly available as of the date of publication. DOIs are listed in the key resources table.

• All original code is available in this paper's supplemental information.

• This paper does not report original code.

Instructions for section 3. Section 3 consists of the following statement: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

In addition,

STAR Methods follows a standardized structure. Please reorganize your experimental procedures to include these specific headings in the following order: LEAD CONTACT AND MATERIALS AVAILABILITY (including the three statements detailed above); EXPERIMENTAL MODEL AND SUBJECT DETAILS (when appropriate); METHOD DETAILS (required); QUANTIFICATION AND STATISTICAL ANALYSIS (when appropriate); ADDITIONAL RESOURCES (when appropriate). We're happy to be flexible about how each section is organized and encourage useful subheadings, but the required sections need to be there, with their headings. They should also be in the order listed. Please see the STAR Methods guide for more information or contact me for help.

Thank you!

Cell Systems

Transparent Peer Review Record

Reviewer comments:

Reviewer #1: The authors have made a diligent effort to address all the comments and suggestions made in the prior review. LinkedOmicsKB presents a timely and valuable resource to the cancer community. It is appropriate for publication in Cell systems.

Reviewer #3: The authors have addressed my concern.

Re: CELL-SYSTEMS-D-23-00102 "A proteogenomics data-driven knowledge base of human cancer"

REVISIONS IN RESPONSE TO REVIEWERS' COMMENTS

We thank the reviewers for the insightful comments and constructive suggestions. We have considered all comments and suggestions and revised the manuscript accordingly. For your convenience, we highlighted the changes in blue in the revised manuscript. Please see below for a point-by-point response to each of the points made by the reviewers.

Reviewer #1:
This manuscript by Liao Y. presents LinkedOmicsKB, an online resource consisting of ~40,000 precomputed gene-, protein-, mutation-, and phenotype-centric web pages, based on the analysis results using the harmonized pan-cancer proteogenomic profiles from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Its utilities to explore these complex and interconnected resources were demonstrated by three case studies, to reveal novel insights for an understudied druggable protein, TP53 mutations, and cancer-related phenotypes. Overall, it is well-designed and user-friendly, which could be a valuable resource for the cancer community to discover clues for different questions. However, there are some improvements and clarifications needed and listed below.

Major issues:

1. Although the current version of LinkedOmicsKB includes over 1000 tumors across 10 cancer types, it would be useful to keep updated regularly as more proteogenomic studies are expected to get published in the future. Therefore, it is important to provide the updated versions of LinkedOmicsKB after the current initial version, and the update plan would be mentioned in the discussion section. Moreover, it is necessary to provide a document page on the website introducing the data sources and summary statistics for each version, including the current version.

Response: A high level description of the data source can be found on the front page. We added a download page (https://kb.linkedomics.org/download) to make all data used in this version available for download. As suggested, we also added a statistics page (https://kb.linkedomics.org/statistics) to provide summary statistics of the current version. It is our plan to continue updating LinkedOmicsKB as new data becomes available. We described our update plan in the Discussion section in the revised manuscript:

Revision: "*To ensure that LinkedOmicsKB remains up-to-date, we will follow a systematic plan for updating the portal with new data as it becomes available. This plan involves identifying relevant sources of data, assessing their quality, processing the data using our standardized pipelines, and integrating it into the existing database. We will prioritize maintaining the comprehensiveness, quality, and consistency of the portal throughout this process. Regular*

updates will be scheduled and communicated to users, providing relevant metadata, annotations, and summary statistics to aid in the interpretation of the new data."

2. Figure descriptions is lacking for many plots on the website. Although it is mentioned somewhat in the manuscript and bioinformatics veterans may understand from the legend panels, it may be inadequate and inconvenient for users who are not that familiar like clinicians. Thus, it is recommended to provide the descriptions for most plots in the appropriate places in the web pages.

Response: We added a manual page (https://kb.linkedomics.org/manual) that can be accessed from the front page of the website. This provides detailed information on how to navigate the gene-, protein-, mutation-, and phenotype-centric web pages and how to use and interpret the figures and tables. We have also added necessary description and information of the statistical test in the pop-up plot window.

3. Regarding genome and transcriptome data pre-processing, are there standard pipelines utilized in CPTAC, as the one used for processing and quantifying the mass spectrometry data? Will the mutation calling results be comparable with the original publications?

Response: All genomic, transcriptomic, and proteomic data used in LinkedOmicsKB were processed using standardized data processing pipelines. Standardized data processing and data harmonization is a major effort of the CPTAC pan-cancer working group, and it is described in detail in a companion paper that has been accepted in principle at Cancer Cell and cited in our manuscript. We have improved our Methods section (see below) to help readers understand the process without having to go back to the companion paper.

The overall consistency between these data and the originally published data is very high. However, there may be minor differences in mutation calling as well as other omics results between this version and the original publications. The differences are mostly due to pipeline improvements and standardization.

Revision: "*CPTAC pan-cancer proteogenomics data processed using standardized data processing pipelines by the CPTAC pan-cancer working group and harmonized by the BCM harmonization pipeline were used in this study. Detailed information on data processing and harmonization is described in a companion paper (Li et al., accepted, Cancer Cell). Briefly, the data included only cases and samples used in the flagship manuscripts[4–10,13,16,32]. For gene harmonization between omics, all data were processed using a common reference genome annotation, GENCODE V34 basic (CHR)[33]. A single primary isoform was selected for each gene. For coding genes, MANE Select and SwissProt were used to prioritize isoforms. If a gene did not have a single MANE Select and/or SwissProt isoform, then the isoform was prioritized using the longest protein sequence followed by the longest transcript. Additionally, remaining Swiss-Prot proteins and MANE Plus Clinical isoforms were retained as secondary isoforms for web portal display. For data harmonization across cohorts, standardized pipelines were used to process each omics type. For RNA and proteomics, expression levels were normalized to a*

common value. Below, we provide a brief overview of the data processing methods used for each omics data type, and further details are available in the companion paper (Li et al, accepted, Cancer Cell)."

Minor issues: The current version of LinkedOmicsKB was analyzed on the level of cancer types. If multiple studies for a same cancer type was included in the future version, how to integrate the different studies of a cancer type?

Response: This is an important direction for future development, and we added our plan in the Discussion section.

Revision: "*Additionally, we also plan to make the analysis pipeline and web portal customizable in the future, so that the framework can be applied to other cohort-based proteogenomic studies both within and outside the cancer community. For example, a breast cancer portal can be developed to integrate proteogenomics data generated from multiple independent breast cancer studies.*"

Reviewer #2:

In this manuscript, Liao et al, present a bioinformatics online platform for analyzing and visualizing pan-cancer proteogenomics data and precomputed analytic results. To evaluate its performance and clinical utility, the authors specifically focused on three case studies, i.e. using the platform to analyze (i) understudied druggable proteins, (ii) clinical phenotypes, and (iii) TP53 mutations. The bioinformatics platform is well-designed. Its interface is intuitive and user-friendly. The analytic and visualization modules are comprehensive and very useful. It is expected that the platform will highly benefit the cancer research field. Here, the reviewer only has a few minor comments.

1) Please indicate which test was used to calculate the p-values in Fig. 4i and 4j.

Response: P-values are from the logrank test. Legends of Fig. 4 are updated to add more details.

2) In Fig5b, in addition to the labels, please also highlight the data points in the scatter plot.

Response: Fig. 5b has been updated to highlight the labeled data points with color.

3) For all the scatter plots, please also add p-values and correlation coefficients and indicate them in the main text (at least for the main figures). Only showing the directions (negative or positive) in the main text is not enough.

Response: The p-values and correlation coefficients are now included in all scatter plots in the manuscript and on the website.

Reviewer #3:
The authors have created a web-based knowledge base called LinkedOmicsKB, which is built upon CPTAC's pan-cancer proteogenomics data. The primary objective of LinkedOmicsKB is to transform this data into meaningful and intelligible insights into cancer biology. Through three case studies, the authors demonstrate how LinkedOmicsKB can be used to develop novel hypotheses on genes, phosphorylation sites, somatic mutations, and cancer phenotypes. The paper is well-written, and I enjoyed reading it. I only have one comment to make.

The author stated that the data were recently harmonized and referred to Fig1a and Methods. However, I found that neither the figure nor the methods section provided a detailed description of how the harmonization was carried out.

Response: We thank the reviewer for the positive comments. All genomic, transcriptomic, and proteomic data used in LinkedOmicsKB were processed using standardized data processing pipelines. Standardized data processing and data harmonization is a major effort of the CPTAC pan-cancer working group, and it is described in detail in a companion paper that has been accepted in principle at Cancer Cell and cited in our manuscript. We have improved our Methods section to help readers understand the process without having to go back to the companion paper.

Revision: "*CPTAC pan-cancer proteogenomics data processed using standardized data processing pipelines by the CPTAC pan-cancer working group and harmonized by the BCM harmonization pipeline were used in this study. Detailed information on data processing and harmonization is described in a companion paper (Li et al., accepted, Cancer Cell). Briefly, the data included only cases and samples used in the flagship manuscripts[4–10,13,16,32]. For gene harmonization between omics, all data were processed using a common reference genome annotation, GENCODE V34 basic (CHR)[33]. A single primary isoform was selected for each gene. For coding genes, MANE Select and SwissProt were used to prioritize isoforms. If a gene did not have a single MANE Select and/or SwissProt isoform, then the isoform was prioritized using the longest protein sequence followed by the longest transcript. Additionally, remaining Swiss-Prot proteins and MANE Plus Clinical isoforms were retained as secondary isoforms for web portal display. For data harmonization across cohorts, standardized pipelines were used to process each omics type. For RNA and proteomics, expression levels were normalized to a common value. Below, we provide a brief overview of the data processing methods used for each omics data type, and further details are available in the companion paper (Li et al, accepted, Cancer Cell).*"