PomBase SAB Meeting

10 DEC 2024 / 4:00PM BST / ZOOM

Attendees

Present:
*Li-Lin Du (*National Institute of Biological Sciences, Beijing, China), *Kathleen Gould* (Vanderbilt University, Nashville, TN, USA), *Sophie Martin,* (University of Lausanne, Lausanne, Switzerland), *Alison Pidoux* (Allshire Lab, Centre for Cell Biology, University of Edinburgh, Edinburgh, UK) *Sigurd Braun* (Institute of Genetics, Justus-Liebig-University Giessen, Giessen, Germany). Jurg Bahler (PomBase, UCL), Juan Mata (PomBase, Cambridge), Valerie Wood (PomBase, Cambridge), Pascal Carme (PomBase, UCL), Kim Rutherford (PomBase, Cambridge))

Apologies: n/a

Zoom https://cam-ac-uk.zoom.us/j/84154500685?pwd=rUySw6RZuVt7BV88N5KafJY2T0zLT4.1

Resources

PomBase 2024 update for SAB

Feedback

Areas for improvement

Li-Lin suggestions:

1. Top of the page focus

Both AlphaFold view and Genome browser view are very useful to me, and I often need to access both of them during one visit to PomBase. However, toggling between them can take a bit of time. With the addition of PATHWAY VIEWS, it will possibly be more of a problem. I wonder whether it is desirable and feasible to load all these views when opening the page, rather than let the user toggling between the views.

Ideas from the discussion with the SAB:

The slower loading of the widgets display on Li-Lin's browser is probably due to geographical constraints that are out of PomBase's control.

The idea of having an option for users to choose the default widget to display on their browser has been raised. This solution would be possible, but would need to implement some type of options/preferences page.

Li-Lin suggested having the genome browser widget always displayed, and the other widgets collapsible. We are wondering wether this solution would be preferred by the majority of the community, and consider asking about it in the next user survey.

2. Protein localization

- protein localization" is a key functional aspect of a gene. Currently, this information is listed under "GO cellular component", where in silica prediction and experimental evidence are mixed together. There might be ways to improve this.

  As demonstrated during the discussion, PomBase provides a filter allowing to display based on the type of evidence supporting the annotation. During our post-meeting discussion, we have decided to make this filter clearer for all GO annotations on PomBase by:

  - *Changing the label of the [Throughput] filter to [Evidence type]*
  - *Changing the evidence groups filters to:*

    High throughput experiment = (HDA/HMP)

    Low throughput experiment = (EXP/IMP/IGI/IDA)

    All experimental evidence = (HDA/HMP) + (EXP/IMP/IGI/IDA)

    Inferred = (IC/TAS/NAS/ISO/IBA/ISS)

  Adding a filter for EXP data pombase/website/issues/2297 DONE

  The idea of having a separate type of localisation annotations, different from GO CC, to capture protein localisations that are not related to their activity has been mentioned. As we want to avoid having separate annotations describing the same type of informations, we have decided to start using the "located_in" GO relation https://wiki.geneontology.org/Located_in for CC annotations not related to an activity, instead of the "is_active_in" relation.

  Recording locations where gene product is inactive /pombase/website/issues/2298

- "SPD / RIKEN - Orfeome Localization Data", which is listed as "Strain and reagents", but is actually a localization database.

  ACTION ITEM FIXED https://github.com/pombase/website/issues/2295

- RIKEN web pages may disappear one day, as has happened to other useful resources (such as the Broad Institute web pages). Maybe it is worth considering hosting the dataset at PomBase.

  To mitigate against a disaster, we made a copy of the RIKEN dataset

- Another large-scale protein localization dataset is the following one, which may have not been linked at PomBase. Again, if possible, hosting the dataset at PomBase will be useful in case these web pages disappear one day.
  https://www2.nict.go.jp/bio/seibutsu/DATA/GFP-lib-New/indexGFP.html

  For the Hayashi data the authors would need to provide mapping between their image files to PomBase systematic identifiers for us to be able to provide gene-specific links to the entries in this dataset. Kim has made a local copy of the images and tables in the dataset in case the website disappears in the future.

  ACTION ITEM FIX PROPOSED https://github.com/pombase/website/issues/2295

Comments from Alison's team:

- One (cytology-focused) person suggested: One thing that would be useful for me is to collate various microscopy images of the same protein, i.e., to have a microscopy page with any published live-cell, IF, Super-Res, etc. imaging.

Possible Solutions:

External links and image hosting

https://www.ebi.ac.uk/bioimage-archive/ might be a good home

We could then hopefully use some Web app to host relevant images on gene pages

[Note that inferences and HTP are removed if they are known to be incorrect, However, GO is supposed to represent only the localization of activity, not transit locations (for example as cargo through the ER), although we have not followed this mandate fully, and still represent experimental locations that are no location of activity. We need to obtain clarification on this.

3. The DNA binding sites for transcription factors reported in the following preprint is a useful resource, and worth being considered to be added in PomBase in someway.https://www.biorxiv.org/content/10.1101/2024.08.20.607873v2

Action: We will prioritize this dataset

4. Gene structure annotation changes can have a big impact on the researchers working on the affected genes.

- Currently, this information is under the sections “Warnings” and "Gene structure history". I wonder whether this information can be more prominently displayed, for example, in the topmost “Summary” section.
- In addition, I wonder whether it is possible to add a description of what type of changes had occurred, for example, “Start codon revision” or “The currently annotated start codon corresponds to the second methionine codon in the previous annotation”.

We have new history section but we have not announced yet (waiting for finalization).

We have made these links moore prominent, th new history pahges are available from the genome status menu

https://www.pombase.org/status/gene-coordinate-changes-protein-coding

https://www.pombase.org/status/gene-coordinate-changes-RNA

We still need to announce this.

Alison Pidoux suggestions
OverviewVery positive. Essentially everyone loves PomBase. Superior to other genome databases which with respondents familiar. Breadth and depth of annotation valued. I have always found PomBase team to be very responsive to requests for information and tools.

5. Fungal Pathogens Having more connections to Fungal Pathogens could widen usefulness of PomBase.
At present it’s useful in one direction:
‘Working on a fungal pathogen (*Cryptococcus neoformans*), we usually come in to look at 'the pombe homologue' and find out a lot about it, quickly and efficiently.’
But would be beneficial to have information in the other direction:
Would be nice to have links to C*andida* and C*ryptococcus neoformans* homologues (other fungal pathogens too?), rather than just other models and human. ie would be great to have a 'health link' that included better links to fungal pathogens as well as human diseases!

Ideas that were raised in the discussion with the SAB: Providing a list of fungal-specific proteins accessible through an easy to find link. We could also host a fungal drug targets dataset.

We also discussed the possibility of linking to orthologs from pathogenic fungi from gene pages. Discussing this idea post-meeting discussion, we decided to look at the World Health Organization priority fungal pathogens (+ input from the community) and to link from gene pages to UniProt orthologues predicted by PANTHER.

Find out which WHO priority fungal pathogens are in PANTHER and request any exclusions

Use PANTHER to provide ortholog data store in Chado

Link to UniPort entries in ortholog page section

Make UniProt IDs searchable in the PomBase simple search and the gene mapper

Kim also suggested adding the fungal pathogens species for which we get a mapping to the orthologs identifier mapper.

Potential fungal pathogen ortholog links /website/issues/2305

6. Swamping of phenotype and interaction annotation

This generated the most frequent feedback on '*what aspects of PomBase annoy you most?'* Strong opinion that some kind of ranking in phenotype data is needed. For instance, the phenomics data from the Bahler lab (Rodriguez-Lopez et al, 2023) is an excellent resource, but the more subtle effects seen (e.g. 10% reduction in growth) crowd out the more severe effects for initial searches of phenotype on PomBase, often giving many hundreds of hits for e.g. resistance to chemical X, with no sense of the strength of phenotype/impact.

Discussing this issue post-SAB, we decided to display the Phenotype Scores (already stored in our annotation dataset and displayed on the article page) from the Rodriguez-Lopez et al. dataset in the Advanced search results (list of genes) when searching specifically for the phenotypes covered in the article.

Ditto for the physical interaction data—some of the studies included have an extremely long list of supposedly interacting proteins. Because of this, we generally don't pay much attention to physical interactions. We are aware that there is a high/low throughput filter for genetic/physical interactions.

Vw: This has bothered me- I only noticed today that you can use the filter "cell population viability" to filter all of the sensitivity and resistance terms away (which seems slightly odd but the viability terms only capture whether cells are inviable or viable or if increased cell death is measured, rather than slow growth and sensitivity/resistance don't distinguish this)

7. Genome Sequence Update

Several people responded in the affirmative: Yes, need T2T sequence. Proper sequence across the centromeres, mating type. Maybe include several strains.

Comment from our (AP's) bioinformatician Pin Tong: "T2T will be very useful but new data will need to be generated. Strain selection important. As pombe have mating type, it's hard to just have one version of genome. May need to make multiple version - the solution is to make a graph genome. So far graph genome is not user-friendly for biologists without bioinformatic training."

'But if they can prove a chain file for liftover from other Array data's genome version will be very useful.' Like UCSC, https://genome-asia.ucsc.edu/cgi-bin/hgLiftOver

From the discussion with the SAB members, we took the decision to to postpone the updating of the genome sequence once PomBase become part of the Alliance of Genome Resources.

8. Suggestions for other tools / useful features

a) A great addition would be to add AI powered chat-boxes where users can simply write their queries and get responses. This way, the learning curve for how to use various tools on the website etc would be mostly gone.

Vw: This will likely happen when we become part of the Genome Alliance

b) CRISPR4P (or something similar as a tool) integrated in PomBase when we look for a region.

Vw: In progress, we need direct links to Manu's share your cloning tool for CRISPR & primer design

c) For the search to find things on the website as well as in the categories genes, terms, publications. eg Centromere search term should direct you to centromere map /info.

Vw This has been on the to-do list for a while, reprioritise?
https://github.com/pombase/website/issues/929
But searching for centromeres will also require pages for non gene features
While demonstrating some of the features available on PomBase, the SAB suggested to do outreach about hidden features like the Documentation search, use of different filters on gene pages, and the query builder.

The SAB suggested to develop a tool to automatically build interaction networks (like STRING). As mentioned during the discussion, we probably won't do during this grant because we are waiting for a general tool developed by other resource.

Suggestion to especially flag genes with different phenotypes between S. pombe and S. japonicus. Post-meeting discussion: This might make a nice student project to consider once phenotype annotations in japonicus becomes more comprehensive.

9. New types of data

Single-cell RNA-Seq.

10. Curation

Comments generally along the lines of anticipating that the task will be performed by AI in near future.

Vw: Yes and no. LLMs are only good at highly represented knowledge, and if they don't know, they guess. They perform poorly on long-tail data in a small number of publications which is precisely what we are trying to capture. We are working with members of the AI community and will hopefully be able to import genes, alleles and standardized curation for some datatypes into Canto for review as part of our next grant. The review process will be critical because we aim for >99% accuracy and AI is achieving only 60-80% even for gene name recognition.*
** Physical interactions, localizations, modifications and catalytic activities are likely to work well because they have formal semantics. GO processes and phenotypes, not so well because they are described using highly variable language, and often embedded in figures. We also need the annotations connected to the correct entities (genes, alleles). Getting all of this correct will be a challenge. Our data will likely be used as gold standard training data for GO curation.*

Funding justification:

- How can we better demonstrate the "counterfactual" to funders—i.e., what would happen if PomBase ceased to exist?

Use Google Analytics to approximate the number of minutes PomBase is used by years (give metrics of the time people spend on PomBase). In our post-meeting discussion, Kim explained that this wouldn't be accurate as Google Analytics can't accurately measure the time a user spends on PomBase pages, their estimated average time spent is probably incorrect.

Circulate a user survey with directed questions: How much time do you spend on PomBase? How does PomBase save you time? What would it be like if PomBase didn't exist?

Involve members of the community in some sort of simulation: How long does it take them to find certain information with PomBase vs. How long without using PomBase?

- How can we capture evidence of PomBase's value, especially since knowledge

bases don't receive the same citation metrics as repositories? (For example, how do we convey the qualitative information in the attached letter from Paul Nurse for our last grant application.)

Ideas from the SAB discussion to get community support:

Have a reminder on the front page of how to cite PomBase with a link to the "How to cite us?" page, as well as examples of the cases in which authors should cite PomBase. Having a message/banner on the front page asking the users to remember citing and mentioning PomBase in Acknowledgments. Could add a countdown to the end of PomBase funding in the last year of the current grant. Could also be added to gene pages as a lot of users don't go through the front page.

During our post-meeting discussion, we discussed having such a banner displayed at rotating intervals on the gene pages, to remind users on when to cite PomBase and link to "How to cite us" page. Make the point that if they don't have a specific point on which they can cite PomBase, mentions of PomBase in Acknowledgements also count for funders metrics to demonstrate the value of Pombase.

During the discussion, several members of the SAB talked about various kinds of use cases. I.e "hidden data": data from high-throughput papers (like phenotypes) that can be hard to find by doing a literature search, data that wouldn't be found by literature search because they are not mentioned in the abstract, data that users wouldn't access easily from papers that are not Open Access. They expressed how PomBase is helpful in making this information findable through curation.

We had a discussion about how to collect testimonials and user stories. The SAB will continue to think about these.

Funding PomBase through user subscriptions has been mentioned as a last resort measure, as having users pay a subscription might not even be sustainable.

Questions not covered

Community curation:

- How can we improve community curation uptake (currently at 55% of all publications)?

Alignment with human genetics:

- How can we better connect *S. pombe* knowledge to human genetics and health?

Notes

Post-meeting discussion:

- Display rationale for each link to an orthology resource (in mouse-over)
- Also check when DIOPT was last updated
- DIOPT reference dated to September 2021: last update ? How about the predictors used by DIOPT DIOPT Version 9 - 2023
- Having a table of orthologs for model organisms linking to MODs, using PANTHER to map orthologs (this will only work for a subset of species).

Minutes approved by:

Jurg Bahler

Kathy Gould

Sophie Martin

Li-Lin Du