Visualization of gene expression and expression as a phenotype with the XPO, XAO and DO using a combination of experimental data sources

Troy Pells², Malcolm E. Fisher¹, Erik Segerdell¹, Joshua D. Fortriede¹, Christina James-Zorn¹, Kevin A. Burns¹, Virgilio Ponferrada¹, Praneet Chaturvedi¹, Vaneet Lotay², Mardi Nenni¹, Stanley Chu², Ying Wang², DongZhuo Wang², Kamran Karimi², Peter D. Vize², and Aaron M. Zorn¹.

¹Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Abstract—Xenbase (www.xenbase.org) is a knowledge base for researchers and biomedical scientists that employ Xenopus (X. laevis and X. tropicalis) as a model organism in gaining a deeper understanding of developmental and disease processes. Through expert curation and automated data provisioning from various sources this MOD (model organism database) strives to integrate the Xenopus body of knowledge together with the visualization of biologically significant interactions. We present a vision for the usage of various ontologies that facilitate the visualization of gene expression from a combination of experimental data sources and the linking thereof back to human disease modeling.

In keeping with the philosophy to foster greater insight through the use of visualization techniques that bring together data from different sources and show trends in the large body of experimental and curated data in Xenbase. The approach has been taken to represent key developmental stages and selected embryonic/adult tissue anatomy terms with a modified heat map rendering. This gives us the combined results from the in-situ gene expression summary as well as normalized TPM read counts (Sessions et al.) and UMI counts from the single cell experiment data (Peshkin et al.). Through the acquisition of RNA-seq and ChIP-seq Xenopus data from GEO and subsequent processing through a bioinformatics pipeline to obtain average TPM read counts and differential expression readings. This enables the subsequent construction and visualization of EaP (Expression as a Phenotype) in conjunction with the XPO (Xenopus Phenotype Ontology) and DO (Disease Ontoloty). This data manipulation has practical application in making the information accessible from gene pages and linking back to the source (eg:article/image).

If you use Xenbase resources in your research please consider citing us, for example: Nucleic Acids Res. 2018 46(D1):D861-D868.

Keywords— xenopus phenotype ontology; xenopus anatomy ontology; model organism database; gene expression; genomics; human disease modeling; biocuration

I. INTRODUCTION

The curation interface in Xenbase is designed to facilitate the application of various ontologies for performing the annotations to various sources including images and articles. Through the common language and consistency achieved by using these ontologies the benefits of interoperability across organisms is accomplished. The XAO has made it possible to classify and compare developmental stages and tissues applicable to xenopus. In order to capture the observed phenotype and set the stage for interoperability and computational reasoning a deeper understanding is required to satisfy these goals. While the flexibility of post-composition of entity-quality (EQ) statements has merit. The wisdom of using pre-composition like the Mammalian Phenotype (MP) offered a more resilient approach. The advocates of pre-composition gave rise to the construction of the XPO (Xenopus Phenotype Ontology) as cloned from the MP and tailored to the nuances of Xenopus.

In the development of the XPO a mapping of EQ statements for Xenopus phenotype to the MP was used to identify common statements to base the first version of the XPO on. In keeping with the design patterns and interoperability defined for phenotype ontologies, by applying the structure of anatomy and physiology ontologies (Gkoutos et al. [1]). This structure would persist in the XPO as inherited from the MP and for new classes and relationships added there to. Cross species interoperability is accomplished through applying these rules for the structuring of the phenotype ontology such that the ontologies for anatomy, phenotype and disease being employed leverage the cross references to UBERON (Uber Anatomy Ontology) or GO (Gene Ontology), and PATO (phenotype and trait ontology). This in turn integrates well with those species for which these are used directly or are mapped to pre-composed phenotype ontologies.

²Department of Biological Science, University of Calgary, Calgary, Alberta, Canada

II. DATA PROVISIONING

Prior to the development of a phenotype curation interface in Xenbase. A Phenote version customized for the use of the XAO provided the curators with a means to explore phenotype annotations as applied to Xenopus related articles and images. This early work provided a wealth of insight into, and working with, pre/post composed statements. As an alternative to manual curation, especially for EaP (Expression as a Phenotype) the GEO (Gene Expression Omnibus) was identified as a data source for Xenopus RNA-seq and ChIP-seq raw data. A GEO metadata interface was developed to enable the curators to work with the GSE/GSM metadata to review the experiment and control samples together with the capturing of their background, manipulation and assay information. Once again utilizing ontologies including BFO, XAO and a subset of the PubChem small molecules identified in the Xenopus small molecules ontology (XSMO) as applicable to the experiments being curated. Through the NCBI eUtilitize the MINiML xml for the GEO metadata was sourced for Xenopus related experiments (approx. 220). The SRA run tables provided the link between the GSE, GSM and the raw data files. The bioinformatics pipeline for processing the RNA-seq or ChIPseq data was constructed in order to map to the current xenopus genome and to generate TPM read counts and DE (Differential Expression) reads based on the GEO metadata that had been reviewed and annotated by the curators.

The GEO loader is a java application developed to consume the GEO metadata as well as provide the staging of the TPM read counts and DE data. The population of a Star schema table structure in Xenbase is slated for the storage of this data in a Fact and Dimension table structure for subsequent consumption by the gene expression visualization that is being contemplated. The EaP statements derived from the DE reads would be stored in tables linked to the GSE and in turn the article and relevant image for the experiment.

III. GENE EXPRESSION VISUALIZATION

The Insitu gene expression data that exists in Xenbase provided the basis upon which to develop a modified heat map to identify developmental stage and tissue expression of a gene on a gene page. The TPM read counts (Session et al. [2]) supported the identification of gene expression in select developmental stages and adult tissue stages. For the embryonic tissue stages (Briggs et al. [3]) single cell experiments provided UMI read counts suitable for identifying gene expression for these tissues. By bringing together the gene expression data and combining them in a heat map visualization, meaningful identification of expression of gene's at various stages and tissues can be accomplished.

IV. CONCLUSION

Bridging the pre/post composition phenotype statements represents one of the issues that remains to be fleshed out. The application of a graph path to map one to the other and determine the accuracy, remains to be explored. The handling of copyright images (embargo) related to phenotype statements and the timing of the curation effort is another process flow topic that is under reviewed.

As the development of the XPO progresses and additional disease (DO) relationships are constructed, the benefits derived from the combined availability of expression as a phenotype, anatomical phenotype and gene ontology phenotype to contribute towards human disease modeling is anticipated. Additional data integration with Monarch via their dipper pipeline is also contemplated.

ACKNOWLEDGMENT

Major funding for Xenbase is provided by the National Institute of Child Health and Human Development, grant P41 HD064556

REFERENCES

- [1] Gkoutos et al. "The anatomy of phenotype ontologies: principles, properties and applications" Brief Bioinform. 2017 Apr 6. doi: 10.1093/bib/bbx035.https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbx035/3108819
- [2] Session et al. "Genome evolution in the allotetraploid frog Xenopus laevis." Nature 2016 Oct 20;538(7625):336-343. doi: 10.1038/nature19840 https://www.ncbi.nlm.nih.gov/pubmed/27762356
- [3] Briggs, Peshkin et al. "The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution." Science 2018 Jun 1;360(6392). pii: eaar5780. doi: 10.1126/science.aar5780. https://www.ncbi.nlm.nih.gov/pubmed/29700227
- [4] Karimi et al. "Xenbase: a genomic, epigenomic and transcriptomic model organism database", Nucleic Acids Res. 2018 46(D1):D861-D868. https://www.ncbi.nlm.nih.gov/pubmed/29059324
- [5] James-Zorn et al. "Navigating Xenbase: An Integrated Xenopus Genomics and Gene Expression Database", Eukaryotic Genomic Databases: Methods and Protocols, 2018, Volume 1757, Chapter 10, pp. 251-305. https://www.ncbi.nlm.nih.gov/pubmed/29761462