GigaBlog

Meet the GigaScience ICG Prize Winner, Pt. 2: Lisa Johnson Q&A

Scott Edmunds

Published December 14, 2018

Citation
Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. https://doi.org/10.59350/6jgh5-bbh50

Keywords
Biology, Genomics, ICG, ICG Prize, ICG13
Feature Image

Lisa Johnson ICG prize winnerYesterday we published the winning paper of the second *GigaScience* prize, with additional detail and coverage in GigaBlog describing why we and the judging panel found it so novel. This was an impressively case study in reproducibility, reassembling & reannotating around 700 microbial eukaryotic transcriptomes to demonstrate this approach can aid in revealing new biologically relevant findings and missed genes from old data. Here in a follow-up blog is one of our author Q&A's with first author Lisa Johnson. Lisa presented her work at our prize track session at ICG13 (the 13th International Conference on Genomics) in Shenzhen, and we have the video and slides from her talk available to view online. Lisa is in the Molecular, Cellular, and Integrative Physiology lab of our Editorial Board Member C. Titus Brown at UC Davis.

With additional detail in an accompanying commentary on what the broadly applicable take home messages for re-analyzing data are, in this Q&A we ask Lisa for some more insight into why they carried out the work, what the technical and cultural challenges were, and what the ideal research infrastructure would be for this type of work.

While raw sequencing data is well catered for why do you think processed data such as assemblies annotations and results haven't been supported and shared in the same manner?

There are public repositories such as NCBI-TSA and NCBI-Genome for depositing data products such as assemblies and annotations generated by assemblers and annotation pipelines. But, if you aren't the owner of the original raw data, you can't submit assemblies to these "official" repositories.

The original MMETSP transcriptomes assembled by the National Center for Genome Research (NCGR) that we re-assembled are hosted by a privately-hosted website, iMicrobe and are available on an ftp site. These addresses and files can change without a change log. (And they did throughout the three-year process of this paper.)

Re-assemblies are improvements to the original assemblies. The processes, i.e. software and pipeline management for assemblies and annotations are improving. We found that reference assemblies can change and potentially provide additional useful information with newer versions and different assembly tools.

Why do people keep creating new sequencing data rather than going back to re-use previously produced data?

ICG prize winner

Lisa presenting this work at ICG13.

The collection of RNA sequences expressed in a sample of cells is dynamic – unlike DNA sequences. New RNAseq data generated under experimental conditions to help answer specific biological questions allow us to see the overall physiological status of cells. But, a high quality reference is needed to characterize RNAseq expression data. For those of us whose study organisms that do not have a well-characterized reference genome like the mouse or human,

we have to generate *de novo* assemblies and annotations to generate our own reference transcriptomes or genomes. Having the best possible quality reference is necessary to be able to accurately characterize new RNAseq data. This is especially true if significant investments will be made downstream based on differential expression results, as is sometimes the case with biomarker and drug discovery in the agriculture, food and pharmaceutical fields.

You provide some suggestions in your commentary to overcome the technical challenges of data re-use, but what do you think the cultural challenges are?

I see funding, computing resources and training as the main challenges for re-assembling old data to generate better references.

Domain-specific research funding seems to be primarily focused on novelty: generating new data to answer new questions in the field. If algorithms or software tools are developed in the pursuit of these questions, great, but ongoing work to continue software development or reference improvements are rarely supported, I believe. This work was thankfully supported by the Moore Foundation's Data-Driven-Discovery initiative Investigator award to my advisor, C. Titus Brown. This Moore Foundation initiative had the foresight to support scientific discovery specifically in the area of data science towards advancing solutions that can be used by others.

Right now, *de novo* assemblers and annotation pipelines are computationally expensive, time-consuming and require training to use. Running them again might seem like insanity. The pipeline I ran to re-assemble the 678 MMETSP samples took ~8,000 wall-clock hours (678 assemblies * avg 12 hours each) to run. The main de novo transcriptome assembler, Trinity requires at least 1 GB of RAM per 1 million reads. Depending on the organism, more RAM than this is usually better. I had to run our whole MMETSP pipeline twice and a collection of samples a few times more to update the Trinity version and re-assemble some of the problem samples. I ran the majority of samples on the Michigan State University's iCER high performance computer (HPC), for legacy reasons, and experienced some technical challenges associated with running hundreds of de novo transcriptome assemblies in parallel with the Trinity assembler. Some of the assemblies were also run on NSF-XSEDE's Jetstream cloud platform. NSF-XSEDE has been a great avenue for providing free computing resources for US researchers, and I highly recommend the Bridges HPC resource available through XSEDE.

We run training workshops at UC Davis and meet people with data from many interesting organisms and questions. We can help with generating assemblies and annotations. But for many groups this is the only reference that they will generate as it can be technically challenging and time-consuming to even get to that point. Time is particularly valuable when pursuing answers to the biological questions, funded by domain-specific research grants and often doesn't allow for re-assemblies and evaluations.

You used Zenodo to host the resulting outputs and data. What were the advantages of that platform and what would the ideal repository or hosting body be for this?

The main advantage of zenodo is that it offers persistent storage, funded by CERN and will likely not go away in the next several decades. A DOI is generates for the repository, which is

versionable, and interfaces with GitHub releases. We have direct control of the displayed text and uploaded file content without having to go through a party to make changes. These changes are automatically logged.

The main downfall of zenodo is the 50GB repository limit. While this is an improvement from their previous 2 GB/file limit, for me the ideal platform is one with no file size or repository limits so that raw data could be included. (I realize this is expensive to maintain.) Another ideal feature is automated upload/download, as zenodo has available through their rest api.

You demonstrated this approach works with MMETSP data but where should it go next?

Automated and programmable pipelines that are extensible for an arbitrary number of samples and supports software updates can be applied to any data set. We've had some success in our lab automating pipelines with snakemake, which is where I would like to see the MMETSP pipeline go next.

ICG prize winner

Finally it was great having you visit us in Shenzhen to present at the ICG Prize track (pictured above). How did it feel being announced as the 2018 ICG prize winner?

Thank you! I was surprised. All the prize track winners who presented at ICG were very impressive and interesting. I really appreciated the opportunity to attend ICG and meet the other speakers, attendees and *GigaScience* staff.

**Further Reading
**Johnson, LK et al. (2018): Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*. doi: 10.1093/gigascience/giy158

Alexander, H et al. (2018): Keeping it light: (Re)analyzing community-wide datasets without major infrastructure. *GigaScience*. doi: 10.1093/gigascience/giy159

The post Meet the GigaScience ICG Prize Winner, Pt. 2: Lisa Johnson Q&A appeared first on GigaBlog.