Keeping it light: (Re)analyzing community-wide datasets without major infrastructure

Harriet Alexander, Lisa K. Johnson, C. Titus Brown

May 26, 2018

1 Background

Advances in high-throughput, next-generation sequencing technologies have catapulted biology into a new computational era. In fields of biology that leverage sequencing data, the primary limiting step in the earlier stages of biological inquiry has increasingly shifted away from data generation to data analysis. Concomitant with the increasing emphasis on the computational processing of these data is the advancement of the computational tools available for such analyses: new computational approaches for the analysis of these data are constantly being created, tested, and proved worthy of use. Yet, outside of computational lab groups, the practice of reanalysis of previously generated data with new tools and approaches is not commonplace. Such reanalysis has great utility and will become more routine within the life sciences, yet reanalysis necessitates a new set of considerations for best practices and resource development.

Our interest in the issues surrounding reanalysis was spurred by a large-scale sequencing project: the Marine Microbial Transcriptome Sequencing Project (MMETSP), which generated 678 transcriptomes, spanning 396 different strains of eukaryotic microbial eukaryotes isolated from marine settings [1]. This dataset is an invaluable resource within the oceanographic community [2, 1], as it exponentially expands the accessible genetic information base of marine protistan life. Moreover, the MMETSP has created a uniquely useful test dataset for computational biologists. The MMETSP dataset spans a large evolutionary history of organisms, and all of the 678 transcriptomes were prepared and sequenced in a consistent way [2]. The sequencing project, which was completed in 2014, was originally assembled by the National Center for Genome Resources using a custom pipeline that employed the best available computational tools at the time [3, 4].

Since the original MMETSP analysis, new tools and techniques for the assembly of *de novo* transcriptomes from RNAseq data have been described and preexisting tools have been improved upon [5]. Moreover, new annotation tools and databases have become available. The transcriptome assembly project described in Johnson et al. [**?**] was designed to create a push-button assembly framework that not only enables the reanalysis of these data sets, but creates a framework to facilitate easy and rapid reanalyses in the future.

These secondary data products of sequencing, such as annotated assemblies, should be viewed as hypotheses generated from the underlying biology, rather than some immutable 'truth'. As such these secondary data products can continue to be improved as new tools are developed. For example, we note that MacManes (2018) described several limitations and challenges of current assembly technology and developed an improved Oyster River Protocol, which we could use to generate another, perhaps improved, MMETSP assembly.

Ultimately, such iterations on the original raw data have the potential to improve upon the secondary data products – the assembled transcriptomes and associated annotations that are relied upon by the broader community for biological inquiry.

Through this process, we developed several practices that we believe to be broadly applicable when reanalyzing data, especially when done by small research groups.

2 Main text

2.1 Storage of secondary data products

Funding agencies and academic journals now mandate the deposition of raw data into digital repositories (e.g. NCBI Sequence Read Archive (SRA) and Gene Expression Omnibus, European Nucleotide Archive). Thus, to date, the majority of the sequence data that has been generated and published are openly available online for reference and use in other studies. The sharing and availability of raw data from high-throughput sequencing studies has been largely managed through the development of archival services such as the SRA, which was established as part of the International Nucleotide Sequence Database Collaboration (INSDC)[6, 7]. The SRA currently contains more than 1.8e16 bases of information (~7e15 are open access)[1]. While a tremendous resource for biological inquiry, a major problem remains in that raw sequencing data is not the most directly useful form of sequencing data. Rather, biologists rely heavily upon the computationally generated

[1] As of 17 May 2018.

secondary products of sequencing reads (e.g. assembled transcriptomes or genomes, annotations, associated count-based data, etc.). There is a dearth of these secondary products in central, publicly accessible databases, such as the Transcriptome Shotgun Assembly (TSA) Sequence Database. In fact, a substantial proportion of these data products might be aptly categorized as 'dark data,' as they are largely undiscoverable and often archived independently in association with a publication or on private servers. Even more limiting, however, is that the guidelines for public databases such as the TSA specifically state that "Assemblies from sequences not directly sequenced by the submitter" should not be uploaded to the TSA, thereby excluding the potential for reassembled datasets to be made available and directly linked to preexisting BioProjects, BioSamples, TSAs, and SRA entries.[2]

From the perspective of our MMETSP reanalysis, we argue the community needs more than a place to put the primary and secondary data products associated with a single publication. Ideally, the results of each reanalysis would be deposited in a discoverable location, but would have a coherent archival procedure that is lab-independent, easily searchable, and "forward discoverable" (i.e. when a new version of a data product is released, old versions can point to the new version). Moreover, such an archival platform would ideally document the full provenance of the secondary data product. Movement towards this kind of data archival system are being made both with the development of alternative scientific data publication models (e.g. the Research Object[8]) as well as integration of metadata models (such as the Resource Description Framework) onto existing scientific databases like the European Bioinformatics Institute (EBI) [9], but policies surrounding secondary data products will need to change.

2.2 Directly linking secondary data products to provenance of work-flow

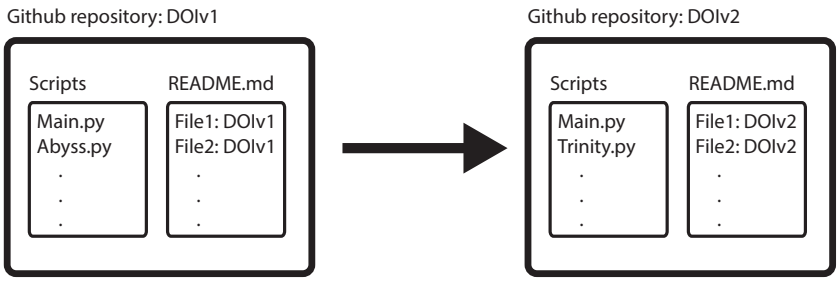

Figure 1: Flow of coupled script and data product evolution.

[2] @CTB do we have a citation/link for this policy?

In the absence of a community database specifically for the type of secondary product that was produced in this analysis, we opted to upload the assemblies, annotations, and counts to Zenodo (https://zenodo.org), a scientific data repository founded by CERN, which provided DOIs for the assemblies. We then created a GitHub repository containing the scripts used to generate the assemblies, and added the output data DOIs (Figure 1). The entirety of the repository was then archived with Zenodo, which generated a single DOI for the project.

As such, the scripts used in the generation of transcriptomes are directly linked through a unique DOI to the data products that are listed in the directory. Since the scripts are easily accessible, they can be tweaked to reanalyze the primary sequence data using different parameters or tools, and the new pipeline and output files can be archived again with Zenodo using the same approach as above. Moreover, the Zenodo archival system will then automatically indicate the presence of other versions of a given repository such that a user might be sure to use the newest version of an assembly. In future, such an approach might be further complemented by the integration of a JSON Linked Data file detailing the metadata for the assembly product, such as the pipeline used and previous versions of the assemblies.

3 Conclusion

The Github-Zenodo framework presented here represents a relatively low cost, undemanding way for small research groups (i.e. a graduate student,) to perform large-scale reanalysis projects in an efficient and publicly accessible way. The direct linking of protocols and metadata to output data products is paramount in the data heavy future of scientific advancement.

We also identified several lingering issues surrounding large scale reanalysis.

Actual computation on these large datasets is a non-trivial issue, as it requires access to facilities with sufficiently large, high-memory machines. Amazon Web Service instances and other "cloud" platforms, including XSEDE, provide flexible computating options, and are broadly accessible. Cloud-based systems, however, tend to be more expensive per computation hour than local resources. High Performance Computing (HPC) resources at local institutions represent another potential site of compute ability. However, HPCs can be temperamental and potentially balk at larger, more node-consuming procedures; moreover, bioinformatics tools may poorly optimized for HPCs: Trinity, used in our pipeline, creates many small files for each run, and this repeatedly caused disk slowdowns on our HPC. The reanalysis by Johnson et al. [**?**] attempted to use both but ultimately found that the HPC provided the most consistent scalable automation for running hundreds of jobs in a cost efficient manner. However, more generally, we see no global solution

for identifying and optimizing the global scientific cyberinfrastructure requirements for projects which require significant scaling; such considerations must be made on a project-by-project basis given the resources available to each lab.

Beyond the optimization of computational resources, we feel that there is a significant opportunity for scientific advancement with high-throughput sequencing projects in making data products 'forward discoverable', because this makes it possible to improve downstream work without significant upstream investment. In an ideal future, a researcher might be automatically notified when a dataset that she is actively working on is updated or changes. This presents many social and technical challenges that will need to be solved if we are to take full advantage of public data sets.

References

[1] Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, Bell CJ, Bharti A, Dyhrman ST, Guida SM, Heidelberg KB, Kaye JZ, Metzner J, Smith SR, Worden AZ: **Probing the evolution, ecology and physiology of marine protists using transcriptomics**. *Nature Reviews Microbiology* 2016, **15**:6–20, [http://www.nature.com/doifinder/10.1038/nrmicro.2016.160].

[2] Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, Nadathur G, Nagai S, Ngam PB, Palenik B, Pawlowski J, Petroni G, Piganeau G, Posewitz MC, Rengefors K, Romano G, Rumpho ME, Rynearson T, Schilling KB, Schroeder DC, Simpson AGB, Slamovits CH, Smith DR, Smith GJ, Smith SR, Sosik HM, Stief P, Theriot E, Twary SN, Umale PE, Vaulot D, Wawrik B, Wheeler GL, Wilson WH, Xu Y, Zingone A, Worden AZ: **The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing**. *PLoS Biology* 2014, **12**(6):e1001889, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4068987{&}tool=pmcentrez{&}rendertype=abstract].

[3] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: A parallel assembler for short read sequence data**. *Genome Research* 2009, **19**(6):1117–1123, [http://www.ncbi.nlm.nih.gov/pubmed/19251739http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2694472http://genome.cshlp.org/cgi/doi/10.1101/gr.089532.108].

[4] Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome research* 1999, **9**(9):868–77, [http://www.ncbi.nlm.nih.gov/pubmed/10508846http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC310812].

[5] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nature biotechnology* 2011,

29(7):644–52, [http://dx.doi.org/10.1038/nbt.1883].

[6] Kodama Y, Shumway M, Leinonen R: **The sequence read archive: Explosive growth of sequencing data**. *Nucleic Acids Research* 2012, **40**(D1):D54–6, [http://www.ncbi.nlm.nih.gov/pubmed/22009675http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3245110].

[7] Shumway M, Cochrane G, Sugawara H: **Archiving next generation sequencing data**. *Nucleic Acids Research* 2009, **38**(SUPPL.1):D870–1, [http://www.ncbi.nlm.nih.gov/pubmed/19965774http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2808927].

[8] Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C: **Why linked data is not enough for scientists**. *Future Generation Computer Systems* 2013, **29**(2):599–611.

[9] Callahan A, Cruz-Toledo J, Ansell P, Dumontier M: **Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data**. In *The Semantic Web: Semantics and Big Data*, Springer, Berlin, Heidelberg 2013:200–212, [http://link.springer.com/10.1007/978-3-642-38288-8{_}14].