

Reprocessing the Microbial Genomic Goldmine: Winner of the ICG13 Prize

Scott Edmunds 

Published December 13, 2018

Citation

Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. <https://doi.org/10.59350/22rps-x6446>

Keywords

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

Copyright

Copyright © Scott Edmunds 2018. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*[Out today microbial genomics goldmine](#) is the winner of our ICG13 Prize, presenting work that can aid in revealing new biologically relevant findings and missed genes from previously generated transcriptome assemblies. Teaching old data new tricks, and maximising every last nugget of information from previously funded research. Here we present some insight into why the reviewers and judges felt this work was so novel and these efforts to reprocess the microbial genomics goldmine has such promise in the field of reproducibility and data reuse.

*

Analyses of genomic data often miss a large amount of information that is present in the data due to the presence of so-called genomics "dark data". This is information that is actually contained within the data, but due to limitations in the analysis methods and variation in the analyses tools used, it is usually missed. Titus Brown and his lab have now created an automated pipeline to assemble and annotate previously-analyzed raw data to dig out this hidden information. This study mines a huge marine microbial dataset from the [Microbial Transcriptome Sequencing Project \(MMETSP\)](#), demonstrating that re-analysing old data with new tools can yield new results.

Previous [work on the MMETSP](#) sequenced 678 transcriptomes and assembled genes that spanned 396 different strains of marine eukaryotes. This dataset has been an invaluable resource within the oceanographic community, exponentially expanding the accessible genetic information base of marine protistan life. In the 5 years since the original analysis was completed, tools, techniques and databases have been improved on. While analysis of this historical data could potentially be carried out again to produce new and more accurate findings, re-analysis of previously generated data with new tools is not commonplace, and it is unclear what the best practice would be. Running analyses again produces different results, and the effects of using different pipelines are poorly understood, making it difficult to determine the usefulness of the new results relative to the previous findings.

The authors of this study tackled this challenge in a systematic manner. They created an automated pipeline to assemble and annotate the original raw data from the MMETSP data. The resulting new transcriptome assemblies were then automatically evaluated in the pipeline and compared against previously-generated assemblies from the original assembly pipeline developed by the National Center for Genome Research. As there is no one-size-fits-all protocol for transcriptome assembly, and as software tools are constantly improving, this pipeline enabled improvements to be tested and quantified. The new assemblies generated containing the majority of the previous data as well as new content. On average, 7.8% of the annotated sequence in the new assemblies had novel gene names not found in the historical assemblies, demonstrating that new findings can be gleaned from old data. Asked why having the most accurate and up-to-date assembly is important, author Lisa Johnson stated: "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."

Microbial goldmine

Lisa Johnson presenting the work at ICG13 in Shenzhen.

While raw sequencing data is commonly shared and well cared for by government funded public archives, the resulting assembled genomes, annotations and results generally are not. For the microbial genomics work carried out in this article, the processed data and results of each re-analysis would ideally be deposited in a discoverable location, with versions and automated redirection to newer versions. The authors have attempted to do just that, with the resulting outputs archived [in the public Zenodo repository](#) hosted by CERN, as well as snapshots from the study archived [in our GigaDB database](#). If we are to take full advantage of public data, this work demonstrates that researchers need to make these products "forward discoverable", automatically notifying users when a dataset is updated or changed. For researchers low on resources, the benefit would make it possible to improve downstream work without significant additional funding, experimentation or sequencing.

This microbial genomics work was selected by our international panel of judges as the winner of our [second GigaScience prize](#), and the first author Lisa Johnson came out to the International Conference on Genomics in Shenzhen to present the work. In a follow up posting we'll present a Q&A with Lisa going into more detail on this work, but we also have [accompanying commentary](#) by the authors expanding on the important lessons for reproducibility. The [video of her talk](#) is also available to view below, and the slides are also [available to view](#).

**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](https://doi.org/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](https://doi.org/10.1093/gigascience/giy159)

The post [Reprocessing the Microbial Genomic Goldmine: Winner of the ICG13 Prize](#) appeared first on [GigaBlog](#).