Metabolic Network Analysis and Metatranscriptomics of a Cosmopolitan and Streamlined Freshwater Lineage

Joshua J. Hamilton1,\*, Sarahi L. Garcia2, Brittany S. Brown1, Jeffrey R. Dwulit-Smith1, Francisco Moya4, Ben O. Oyserman4, Sarah L.R. Stevens1, Stefan Bertilsson2, Katrina T. Forest1, Susannah G, Tringe3, Tanja Woyke3, and Katherine D. McMahon1,4,\*

1 Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA; 2 Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden; 3 United States Department of Energy Joint Genome Institute, Walnut Creek, CA, USA; 4 Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA

\* Correspondence: Joshua J. Hamilton, jjhamilton2@wisc.edu and Katherine D. McMahon, trina.mcmahon@wisc.edu

# Abstract

# Introduction

## acI

### What we know from tag data

* persistent and abundant through time and across lakes
* three clades and thirteen tribes

### What we know from SAG data

* small and streamlined genome
* conclusions on substrate utilization from SAG studies, incl. rhodopsin
* conclusions on substrate utilization from FISH studies
* no genomes from acI-C

## Reverse Ecology

* genome-gazing is a bottleneck, FISH studies are laborious - need an alternative
* reverse-ecology is an emerging paradigm to infer ecology from genomic information w/ a focus on systems-level signatures
* metabolism one aspect of ecology, long tradition of metabolic network analysis
* example of reverse ecology - seed compounds and environment

## Overview

* assembled largest catalog of acI genomes, including acI-C
* use existing reverse ecology to predict seed compounds for acI genomes
* high-throughput approach to leverage incomplete genome data
* confirm conserved gene functions and revealed clade-specific differences
* integrate with metatranscriptomic data
* many diverse transporters, all highly expressed. possible key to success of passive & planktonic lifestyle

# Materials and Methods

## Single-Cell Genome Generation, Selection, and Sequencing

Single-cell genomes were collected from the top of the water column (depth <1m) from each of two lakes, Mendota (Madison, WI, USA) and Damariscotta (Lincoln County, ME USA), in 2009. Samples were cryopreserved and sent to the Single Cell Genomics Center at the Bigelow Laboratory for Ocean Sciences for sorting, as previously described (Martinez-Garcia et al. 2012, Garcia et al. (2013)). Partial 16S rRNA genes amplified previously (Martinez-Garcia et al. 2012) were phylogenetically classified using a controlled nomenclature for freshwater bacteria (Newton et al. 2011) by insertion into references trees created in the ARB software package (W. Ludwig et al. 2004).

Actinobacterial SAGs used in this study were then sent to the JGI for sequencing and assembly, also as previously described (Ghylin et al. 2014). Briefly, shotgun libraries were constructed for each of the SAGs from re-amplified MDA products and sequenced on an Illumina HiSeq2000. All general aspects of and detailed protocols for library construction and sequencing can be found on the JGI website (http://www.jgi.doe.gov/).

For assembly, raw sequence data was first passed through a filtering program developed at JGI to eliminate known sequencing and library preparation artifacts. Assembly was then performed using Velvet [Zerbino and Birney (2008) and ALLPATHS-LG (Butler et al. 2008). Additional details of the assembly process have been previously described (Ghylin et al. 2014) and are available through the JGI Genome Portal (http://genome.jgi.doe.gov) Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi). Genome-specific information can be accessed in both databases by searching for the IMG Taxon OIDs given in Table 1.

## Metagenome Sampling, Sequencing, Assembly, and Binning

Sample collection, DNA sequencing, metabgenomic assembly, and genomic binning for the Trout Bog samples have been described previously (Bendall et al. 2016), and similar procedures were followed for Lake Mendota samples. A summary is provided here.

For Lake Mendota, depth-integrated water samples were collected from the top 12 meters at 96 time points during ice-free periods from 2008 to 2011. For Trout Bog, depth-integrated water samples were collected from the epilimnion (44 samples) and hypolimnion (45 samples) layers during ice-free periods from 2007 to 2009. All samples were filtered on 0.2 μm polyethersulfone filters (Supor, Pall Corp) prior to storage at -80°C, as described previously (Bendall et al. 2016). DNA was extracted from these filters using the FastDNA kit (MP Biomedicals) and sent to the JGI for sequencing, as described previously (Bendall et al. 2016).

Shotgun libraries were constructed for each of the samples and sequenced on an Illumina GA IIx (four Trout Bog samples) or an Illumina HiSeq2000 (all other samples), following a 2x150 indexed run recipe as previously described (Bendall et al. 2016). All general aspects of and detailed protocols for library construction and sequencing can be found on the JGI website (http://www.jgi.doe.gov/). Metagenomic sequence reads are publicly available on the JGI Genome Portal (http://genome.jgi.doe.gov/) under Proposal ID 394.

Raw sequence data was passed through a filtering program developed at JGI to eliminate known sequencing and library preparation artifacts. Prior to assembly, reads were merged with FLASH (Magoc and Salzberg 2011), as previously described (Bendall et al. 2016). Merged reads were pooled by lake and layer into three co-assemblies using SOAPdenovo (Luo et al. 2012), and contigs from the resulting assemblies were assembled into a final assembly using Minimus (Sommer et al. 2007), as previously described (Bendall et al. 2016). Additional details of the assembly process and metagenomic sequence reads are available through the JGI Genome Portal (http://genome.jgi.doe.gov) under Proposal ID 394.

Genomes were binned from each metagenomic co-assembly using MetaBat (D. D. Kang et al. 2015), as described previously (Bendall et al. 2016). Briefly, contigs were classified into bins using tetranucleotide frequency and coverage patterns across the time-series and then manually curated, as previously described (Bendall et al. 2016). Genome sequences are available through IMG (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi) by searching for the IMG Taxon OIDs given in Table 1. Genomes were classified using taxonomic assignments from a set of 37 highly-conserved single-copy marker genes using Phylosift (Darling et al. 2014), as previously described (Bendall et al. 2016). acI genomes (class Actinobacteria) were further classified using a defined ontology for freshwater bacteria (Newton et al. 2011) as described below. Final bin size and number of contigs are reported in Table 1.

## Metatranscriptome Sampling and Sequencing

Four samples were collected from the top of the water column (depth <1m) from Lake Mendota (Madison, WI, USA) over a twenty-four hour period on August 20 and 21, 2015. For each sample, between 200 and 400 mL lake water was filtered onto a 0.2 μm polyethersulfone filter (Supor, Pall Corp) and stored at -80°C until extraction.

Prior to extraction, three samples were spiked with an internal standard to enable quantification of total transcript abundance, following an established protocol (Satinsky et al. 2014). Briefly, a 970-nucleotide-long mRNA standard was synthesized using a a T7 RNA polymerase and the Riboprobe In Vitro Transcription System (Promega, Madison, WI), according to the manufacturer’s protocol. A fixed quantity of each standard (1.172 x 10^10 copies) was added independently to each lysis tube immediately prior to the addition of the sample filter.

Samples were subject to TRIzol-based RNA extraction (Thermo Fisher Scientific, Waltham, MA) followed by on-column DNAse digestion and RNA purification using an RNeasy Mini Kit (Qiagen, Venlo, Netherlands). RNA was then sent to the University of Wisconsin-Madison Biotechnology Center (https://www.biotech.wisc.edu) for sequencing. There, samples were prepared for sequencing using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, CA), with the addition of a step for selective ribosomal RNA depletion using the Ribo-Zero rRNA Removal Kit (Bacteria) (Illumina). The resulting cDNA libraries were pooled in an equimolar ratio, and sequenced on an Illumina HiSeq2500.

Raw paired-end reads were then merged using FLASH (Magoc and Salzberg 2011) using default parameters. Finally, additional rRNA and ncRNA sequences were removed using SortMeRNA (Kopylova, Noe, and Touzet 2012) using default parameters. SortMeRNA was run using eight built-in databases for bacterial, archaeal, and eukaryotic small and large ribosomal subunits and ncRNAs, derived from the SILVA 119 (Quast et al. 2013) and RFAM (Nawrocki et al. 2015) databases.

Additional information, including all protocols and scripts for RNA analysis, can be found on Github (https://github.com/McMahonLab/OMD-TOILv2). Raw RNA sequences can be found on the National Center for Biotechnology Information (NCBI) website under BioProject PRJNA######.

## Genome Completeness and Phylogenetic Relationships

CheckM (Parks et al. 2015) was used to estimate genome completeness based on 204 single-copy marker genes conserved across the phylum Actinobacteria. Phylogenetic analysis of Actinobacterial SAGs and MAGs was performed using a concatenated alignment of single-copy marker genes obtained via Phylosift (Darling et al. 2014). Maximum likelihood trees were generated using RAxML (Stamatakis 2014) using the automatic protein model assignment option (PROTGAMMAAUTO) and 100 bootstraps.

## Metabolic Network Reconstruction and Reverse Ecology

### Genome Annotation and Model Processing (including Genome Merging)

Genome annotations and metabolic network reconstructions were performed using KBase (http://kbase.us/). Unannotated contigs for each genome were pushed to KBase and annotated using the “Annotate Microbial Contigs” method using default options, which uses components of the RAST toolkit (Brettin et al. 2015, Overbeek et al. (2014)) for genome annotation. Genome-scale metabolic network reconstructions were performed using the “Build Metabolic Model” app using default parameters, which relies on the Model SEED framework (Henry et al. 2010) to build a draft metabolic model without gapfilling. As acI genomes are known to be small with numerous auxotrophies (Garcia et al. 2015), we opted to forgo the gapfilling step and instead leverage multiple genomes to fill gaps, as described below.

Metabolic models were then downloaded via KBase, pruned, and converted to metabolic network graphs. In particular, biomass, exchange, transport, spontaneous, and DNA/RNA biosynthesis reactions were removed from the model, and reactions were mass- and charge-balanced. Next, currency metabolites (compounds used to carry electrons and functional groups) were removed following an established procedure (Ma and Zeng 2003). Finally, the network model was converted to a metabolic network graph, in which nodes denote compounds and edges denote reactions. A directed edge from compound *a* to compound *b* means that *a* is a substrate in a reaction which produces *b*. The metabolic network graph representation of glycolysis is shown in Supplemental Figure S2.

Finally, all genome-level metabolic network graphs for a single acI clade were combined to generate a composite clade-level metabolic network graph for that clade. Beginning with two genomes, nodes and edges unique to the second genome are identified and appended to the network graph for the first genome, giving a composite metabolic network graph. The process is repeated for each genome belonging to the clade, until all of the network graphs have been incorporated into the composite.

### Calculation and Evaluation of Seed Compounds

Seed compounds for each composite clade-level metabolic network graph were calculated using established methods (Borenstein et al. 2008). Briefly, the metabolic network is decomposed into its strongly connected components (SCCs), sets of nodes where each node in the set is reachable from every other node. Seed compounds can then be found by identifying source components (components with no incoming edges) on the condensation of the original graph: each source component represents a collection of seed compounds. The process is illustrated for an artificial network in Supplemental Figure S2. Finally, all predicted seed compounds were manually evaluated to identify those which may be biologically meaningful. Examples are given in the Results section.

### Re-annotation of Peptidases and Glycoside Hydrolases

Many seed compounds were associated with reactions catalyzed by peptidases or glycoside hydrolases, and genes associated with these reactions were re-annotated. Peptidase sequences were annotated using the MEROPS batch BLAST interface using default parameters (Rawlings, Barrett, and Finn 2015). Glycoside hydrolases were first annotated using dbCAN (Yin et al. 2012) to assign these genes to glycoside hydrolase families, as defined in the Carbohydrate-Active enZYmes Database CAZY (Lombard et al. 2014). Hidden Markov Models for these sub-families were then downloaded from dbCAN, and HMMER3 (S. R. Eddy 2011) was used to assign these genes to individual sub-families using default parameters. Re-annotated peptidases and glycoside hydrolases are given in Supplementary Tables S2 and S3, respectively.

## Integrating Reverse Ecology with Metatranscriptomics

### Protein Clustering, Metatranscriptomic Mapping, and Clade-Level Gene Expression

OrthoMCL (L. Li, Stoeckert, and Roos 2003) was used to identify clusters of orthologous genes (COGs) in the set of acI genomes. OrthoMCL was run using default options. Then, metatranscriptomic reads were mapped to a single fasta file containing all acI genomes using BBMap (https://sourceforge.net/projects/bbmap/) with the ambig=all and minid=0.85 options. With this command, reads are allowed to map equally well to multiple sites. An 85% percent identity cutoff was chosen as it minimizes the percent of reads which map to multiple sites.

Next, a custom implementation of HTSeq-Count (Anders, Pyl, and Huber 2014) was used to count the total number of reads which map to each (genome, gene) pairing. Any read which mapped to multiple sites within the collection of acI genomes was discarded. Using the COGs identified by OrthoMCL, the total number of reads which map to each (clade, COG) pairing was then counted. This gives a measure of gene expression for the clade-level composite genome.

For each clade, read counts were then normalized by sequencing depth and ORF length and expressed on a reads per kilobase million (RPKM) basis (Mortazavi et al. 2008), while accounting for different ORF lengths within a COG. RPKM counts were averaged across the four metatranscriptomes to give the average expression across a 24-hour period. Finally, within each clade, the percentile rank expression for each COG was calculated.

### Re-annotation of Transporter Genes

Many highly-expressed genes were annotated as transport proteins, and these proteins were re-annotated to assign function to COGs identified by OrthoMCL. Protein sequences were BLASTed against the TCDB reference database (Saier et al. 2014) using blastp and given the annotation of the best-BLAST hit. If a COG contained genes with multiple annotations, the majority annotation was selected for that COG. Examples are given in the Results section.

## Availability of Data and Materials

All genomic and metatranscriptomic sequences are available through IMG and NCBI, respectively. Reverse ecology calculations were performed using the Python package reverseEcology, written expressly for this purpose and available on the Python Package Index. A reproducible version of this manuscript is available at https://github.com/joshamilton/reverseEcologyMS.

# Results

## Phylogenetics (Figure 1)

* assemble Actino genomes from our extensive FW genome collection and other sources
* identify acI as a monophyletic group w/in phylum Actinobacteria, acI-C as a group within acI - Figure 1 shows just these genomes

## Completeness Estimates (Figure 2)

* reverse ecology tested on complete genomes, SAGs and MAGs are incomplete
* construct composite genomes at tribe- and clade-level
* tribe-level remain incomplete, choose clade-level for Analysis

## Protein Clustering and Metatranscriptomics

* total number of protein clusters
* core and accessory genome for lineage based on presence/absence w/in clades
* read recruiting - % of MG/MT reads which map to acI reference genome collections (still need to map the MGs)
* what does this say about…
* are acI active or dormant?
* quality of our reference genomes wrt conditions that day?

## Workflow (Figure 3)

* developed a computational pipeline for reverse ecology analysis on incomplete genomes
* annotate using KBase, no gapfilling (explain justification)
* size of models, fraction of total genes, comparison to other organisms
* metabolic models converted to network graphs and merged
* check size of SCC and reduce to largest component (compare to other studies re: size, % of nodes)
* seed compound calculations - # of seed sets, % which contain one compound, max size

## Evaluation of Potential Seed Compounds (Figure 4a)

* anticipate “noise” in results and evaluate individual compounds
* screen seed compounds to identify a subset for further investigation
* examples of reasons to reject
  + network pruning (carbamoyl phosphate)
  + missing annotations (fatty acids)
* examples of reasons to retain
  + homoserine auxotrophy
  + peptide degradation
* manually curate selected compounds
* false negative - genes missing from the model - lysine auxotrophy
* false negative - alternative pathways - threonine auxotrophy
* true result - gene missing from model and genome - homoserine auxotrophy

## Re-annotation of Peptidases and Glycoside Hydrolases (Figure 4b)

* peptidases
* families and sub-families
* consistency with MAR-FISH and genomic studies
* similarities / differences btw clades - presence/absence and expression
* glycoside hydrolases
* families and sub-families
* consistency with MAR-FISH and genomic studies
* similarities / differences btw clades - presence/absence and expression

## Transporters

* summary of transporters
* diverse array of N-rich compounds
* transport of simple and complex sugars
* nucleotides and vitamins
* other transporters not shown in figure
* actinorhodopsin also highly expressed

## Actinorhodopsin (possible Figure 5 or new analysis)

* among most highly-expressed genes
* coupled with high expression of sugar transporters confirms acI are photoheterotrophs
* retinal biosynthesis pathway expressed (and functional in acI-A and acI-B)
* other reasons for such high expression (MMBR review, need to investigate in our genomes)
* promoter strength?
* co-localized w/ other highly-expressed genes?

# Discussion

## (Dis)Agreement w/ previous findings

* diverse array of N-rich compounds & transport of sugars consistent w/ literature
* cyanophycinase?

## Ecology of acI

* ecological niche
* scavenger of cell lysate, exudates/partially degraded plant materials
* ecological successful b/c ready to eat whatever comes its way
* actR hypothesis: establishes a protein gradient to drive ATP transport
* discuss Sarahi’s vitamin auxotrophy idea?

## Potential and Limitations of Reverse Ecology

* potential applications
* limitations
* only as good as database, known functions of interest are missing
* lacks pathway information, enabling alternative metabolic routes - may miss true auxotrophies

# Acknowledgements

# Conflict of Interest

The authors declare no conflict of interest.

# References

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. “HTSeq A Python framework to work with high-throughput sequencing data.” *Bioinformatics* 31 (2): 166–69. [doi:10.1101/002824](http://doi.org/10.1101/002824).

Bendall, Matthew L, Sarah LR Stevens, Leong-Keat Chan, Stephanie Malfatti, Patrick Schwientek, Julien Tremblay, Wendy Schackwitz, et al. 2016. “Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations.” *The ISME Journal* 10 (7). Nature Publishing Group: 1589–1601. [doi:10.1038/ismej.2015.241](http://doi.org/10.1038/ismej.2015.241).

Borenstein, Elhanan, Martin Kupiec, Marcus W Feldman, and Eytan Ruppin. 2008. “Large-scale reconstruction and phylogenetic analysis of metabolic environments.” *Proceedings of the National Academy of Sciences* 105 (38): 14482–87. [doi:10.1073/pnas.0806162105](http://doi.org/10.1073/pnas.0806162105).

Brettin, Thomas, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Gary J Olsen, Robert Olson, et al. 2015. “RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes.” *Scientific Reports* 5: 8365. [doi:10.1038/srep08365](http://doi.org/10.1038/srep08365).

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. 2008. “ALLPATHS: De novo assembly of whole-genome shotgun microreads.” *Genome Research* 18 (5): 810–20. [doi:10.1101/gr.7337908](http://doi.org/10.1101/gr.7337908).

Darling, Aaron E, Guillaume Jospin, Eric Lowe, Frederick A Matsen, Holly M Bik, and Jonathan A Eisen. 2014. “PhyloSift: phylogenetic analysis of genomes and metagenomes.” *PeerJ* 2 (January): e243. [doi:10.7717/peerj.243](http://doi.org/10.7717/peerj.243).

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195. [doi:10.1371/journal.pcbi.1002195](http://doi.org/10.1371/journal.pcbi.1002195).

Garcia, Sarahi L, Moritz Buck, Katherine D McMahon, Hans-Peter Grossart, Alexander Eiler, and Falk Warnecke. 2015. “Auxotrophy and intra-population complementary in the ‘interactome’ of a cultivated freshwater model community.” *Molecular Ecology* 24 (17): 4449–59. [doi:10.1111/mec.13319](http://doi.org/10.1111/mec.13319).

Garcia, Sarahi L, Katherine D McMahon, Manuel Martinez-Garcia, Abhishek Srivastava, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, and Falk Warnecke. 2013. “Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton.” *The ISME Journal* 7 (1). Nature Publishing Group: 137–47. [doi:10.1038/ismej.2012.86](http://doi.org/10.1038/ismej.2012.86).

Ghylin, Trevor W, Sarahi L Garcia, Francisco Moya, Ben O Oyserman, Patrick Schwientek, Katrina T Forest, James Mutschler, et al. 2014. “Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage.” *The ISME Journal* 8 (12). Nature Publishing Group: 2503–16. [doi:10.1038/ismej.2014.135](http://doi.org/10.1038/ismej.2014.135).

Henry, Christopher S, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. 2010. “High-throughput generation, optimization and analysis of genome-scale metabolic models.” *Nature Biotechnology* 28 (9). Nature Publishing Group: 977–82. [doi:10.1038/nbt.1672](http://doi.org/10.1038/nbt.1672).

Kang, Dongwan D, Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.” *PeerJ* 3: e1165. [doi:10.7717/peerj.1165](http://doi.org/10.7717/peerj.1165).

Kopylova, Evguenia, Laurent Noe, and Helene Touzet. 2012. “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.” *Bioinformatics* 28 (24): 3211–17. [doi:10.1093/bioinformatics/bts611](http://doi.org/10.1093/bioinformatics/bts611).

Li, Li, Christian J Stoeckert, and David S Roos. 2003. “OrthoMCL: identification of ortholog groups for eukaryotic genomes.” *Genome Research* 13 (9): 2178–89. [doi:10.1101/gr.1224503](http://doi.org/10.1101/gr.1224503).

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M Coutinho, and Bernard Henrissat. 2014. “The carbohydrate-active enzymes database (CAZy) in 2013.” *Nucleic Acids Research* 42 (D1): 490–95. [doi:10.1093/nar/gkt1178](http://doi.org/10.1093/nar/gkt1178).

Ludwig, Wolfgang, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier, Yadhukumar, Arno Buchner, et al. 2004. “ARB: a software environment for sequence data.” *Nucleic Acids Research* 32 (4): 1363–71. [doi:10.1093/nar/gkh293](http://doi.org/10.1093/nar/gkh293).

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.” *GigaScience* 1: 18. [doi:10.1186/2047-217X-1-18](http://doi.org/10.1186/2047-217X-1-18).

Ma, Hongwu, and An-Ping Zeng. 2003. “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.” *Bioinformatics* 19 (2): 270–77. [doi:10.1093/bioinformatics/19.2.270](http://doi.org/10.1093/bioinformatics/19.2.270).

Magoc, Tanja, and Steven L Salzberg. 2011. “FLASH: fast length adjustment of short reads to improve genome assemblies.” *Bioinformatics* 27 (21): 2957–63. [doi:10.1093/bioinformatics/btr507](http://doi.org/10.1093/bioinformatics/btr507).

Martinez-Garcia, Manuel, Brandon K Swan, Nicole J Poulton, Monica Lluesma Gomez, Dashiell Masland, Michael E Sieracki, and Ramunas Stepanauskas. 2012. “High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton.” *The ISME Journal* 6 (1). Nature Publishing Group: 113–23. [doi:10.1038/ismej.2011.84](http://doi.org/10.1038/ismej.2011.84).

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. [doi:10.1038/nmeth.1226](http://doi.org/10.1038/nmeth.1226).

Nawrocki, Eric P., Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, et al. 2015. “Rfam 12.0: Updates to the RNA families database.” *Nucleic Acids Research* 43 (D1): D130–37. [doi:10.1093/nar/gku1063](http://doi.org/10.1093/nar/gku1063).

Newton, Ryan J, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. 2011. “A guide to the natural history of freshwater lake bacteria.” *Microbiology and Molecular Biology Reviews* 75 (1): 14–49. [doi:10.1128/MMBR.00028-10](http://doi.org/10.1128/MMBR.00028-10).

Overbeek, Ross A, Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terrence Disz, Robert A Edwards, et al. 2014. “The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).” *Nucleic Acids Research* 42 (D1): 206–14. [doi:10.1093/nar/gkt1226](http://doi.org/10.1093/nar/gkt1226).

Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Phil Hugenholtz, and Gene W Tyson. 2015. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.” *Genome Research* 25 (7): 1043–55. [doi:10.1101/gr.186072.114.Freely](http://doi.org/10.1101/gr.186072.114.Freely).

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.” *Nucleic Acids Research* 41 (D1): D590–96. [doi:10.1093/nar/gks1219](http://doi.org/10.1093/nar/gks1219).

Rawlings, Neil D, Alan J Barrett, and Robert Finn. 2015. “Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors.” *Nucleic Acids Research* 44 (D1): D343–50. [doi:10.1093/nar/gkv1118](http://doi.org/10.1093/nar/gkv1118).

Saier, Milton H, Vamsee S Reddy, Dorjee G Tamang, and Åke Västermark. 2014. “The transporter classification database.” *Nucleic Acids Research* 42 (D1): D251–58. [doi:10.1093/nar/gkt1097](http://doi.org/10.1093/nar/gkt1097).

Satinsky, Brandon M, Brian L Zielinski, Mary Doherty, Christa B Smith, Shalabh Sharma, John H Paul, Byron C Crump, et al. 2014. “The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010.” *Microbiome* 2 (1): 17. [doi:10.1186/2049-2618-2-17](http://doi.org/10.1186/2049-2618-2-17).

Sommer, Daniel D, Arthur L Delcher, Steven L Salzberg, and Mihai Pop. 2007. “Minimus: a fast, lightweight genome assembler.” *BMC Bioinformatics* 8: 64. [doi:10.1186/1471-2105-8-64](http://doi.org/10.1186/1471-2105-8-64).

Stamatakis, Alexandros. 2014. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.” *Bioinformatics* 30 (9): 1312–13. [doi:10.1093/bioinformatics/btu033](http://doi.org/10.1093/bioinformatics/btu033).

Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “DbCAN: A web resource for automated carbohydrate-active enzyme annotation.” *Nucleic Acids Research* 40 (W1): W445–51. [doi:10.1093/nar/gks479](http://doi.org/10.1093/nar/gks479).

Zerbino, Daniel R, and Ewan Birney. 2008. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.” *Genome Research* 18 (5): 821–29. [doi:10.1101/gr.074492.107](http://doi.org/10.1101/gr.074492.107).