ATLAS (Automatic Tool for Local Assembly Structures) - a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data

Richard Allen White III[1*#], Joseph Brown[1#], Sean Colby[1], Christopher C. Overall[2], Joon-Yong Lee[1], Jeremy Zucker[1], Kurt R. Glaesemann[3], Christer Jansson[4], Janet K. Jansson[1*]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

[2]Department of Neuroscience, University of Virginia, Charlottesville, VA, United States

[3]Infomation technology, High Performance Computing (HPC) and Cloud services, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

[4]Environmental and Molecular Sciences Laboratory (EMSL), Pacific Northwest National Laboratory, Richland, WA 99352, USA.

*To whom correspondence should be addressed

#These authors contributed equally to this work.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

Abstract

Summary: ATLAS (Automatic Tool for Local Assembly Structures) is a comprehensive multi-omics data analysis pipeline that is massively parallel and scalable. ATLAS contains a modular analysis pipeline for assembly, annotation, quantification and genome binning of metagenomics and metatranscriptomics data and a framework for reference metaproteomic database construction. ATLAS transforms raw sequence data into functional and taxonomic data at the microbial population level and provides genome-centric resolution through genome binning. ATLAS provides robust taxonomy based on majority voting of protein-coding open reading frames (ORFs) rolled-up at the contig level using modified lowest common ancestor (LCA) analysis. ATLAS is user-friendly, easy install through bioconda maintained as open-source on GitHub, and is implemented in Snakemake for modular customizable workflows.

Availability and implementation: ATLAS is written in python and distributed under a BSD license. ATLAS is compatible with python 3.5+ and anaconda 3+ versions. ATLAS functions on both MacOS and Linux. The source code of ATLAS is freely available at https://github.com/pnnl/atlas.

Contact: Richard Allen White III and Janet Jansson, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. Email: Richard.white@pnnl.gov or raw937@gmail.com, Janet.jansson@pnnl.gov

Keywords: Next-generation sequencing, Metagenomics, Metatranscriptomics, Genome binning, Lowest Common Ancestor (LCA) analysis, Snakemake, Modified Majority Voting-Method (MMVM)

1 Introduction

Whole community sequencing of DNA (i.e., metagenomics) and RNA (i.e., metatranscriptomics) has provided a wealth of information about microbial communities in a variety of habitats, including community compositions, predicted functions, and metabolic potential and activities (Jansson, 2011; Mason *et al.,* 2014; Prosser, 2015; Hultman *et al.,* 2015; Butterfield *et al*., 2016; White III *et al.,* 2016a). Recent improvements in metagenome assembly have enabled direct assembly of large and complex metagenomes (Howe *et al.*, 2014; Li *et al.,* 2015; White III *et al.,* 2016b). In addition, new algorithms have been developed and applied for binning genomes from metagenome data (Albertsen *et al.,* 2013; Imelfort *et al.,* 2014; Wu *et al.,* 2016). These approaches provide valuable insight into the function of microbial populations that are yet to be cultivated.

Current sequencing technologies can reach very high throughput >1 Terabytes (TB) of data in a single run (White III *et al.,* 2016a). With increasing sequencing throughput, a framework for rapid, modular, customizable workflows, and integrated data analysis is needed to obtain meaning from microbial community derived sequencing data. While some metagenomic data analysis pipelines and frameworks exist, such as IMG (Chen *et al.,* 2017), Parallel-META (Su *et al.,* 2014), MG-RAST (Meyer *et al.,* 2008), MetaAMOS (Treangen *et al.,* 2013), and MetaPathways2 (Konwar *et al.,* 2013), none include every key element required for metagenome and metatranscriptome analysis. These key elements include quality control of raw data, assembly, genomic binning, coverage estimation, functional annotation, taxonomic annotation using lowest common ancestor (LCA) and quantitative analysis of reads. Here we introduce ATLAS (Automatic Tool for Local Assembly Structures) as an

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

integrated and customizable pipeline for metagenome/metatranscriptome data quality control, assembly and annotation, metagenome binning, coverage estimation, and expression analysis.

2 DESCRIPTION OF TOOL

ATLAS has five analysis steps: (1) quality control, (2) assembly (3) annotation, (4) genome binning, and (5) taxonomic, functional and expression quantification analyses (Figure 1). Default input data are Illumina paired-end reads in FASTQ format; however, single-end Illumina, Ion Proton, and SOLiD reads in FASTQ format are also supported.

The quality control module (Step 1) involves quality filtering of the metagenome/metatranscriptome sequence read data using the decontamination tool BBduk2 within the BBMap tool suite (https://sourceforge.net/projects/bbmap/). This approach uses k-mers to find and trim adapter sequences, performs quality based read trimming, and filters reads based on a minimum length threshold. The reads have the option of error correction based on both k-mer overlaps and read pair overlaps using Tadpole within BBMap. Decontamination can be performed across any reference read set and reads will be grouped into reference bins or non-hits using BBSplit. ATLAS provides references for common Illumina DNA spike-ins (i.e., bacteriophage phiX) and ribosomal RNA as default contaminant databases. Any additional contamination references in FASTA format are supported and be user supplied. Following decontamination, quality controlled read sets are used in read quantification of Step 5. A future version of ATLAS will include MerCat (i.e., “Mer-Cat enate”), a *de novo* assembly free direct read analysis module plug-in (Figure 1, White III *et al.*, 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

MerCat will provide alpha diversity and feature abundance calculations from quality controlled reads supplied by ATLAS using k-mer counting of any length k, specified by end user, without a reference sequence database dependency (i.e., database independent property analysis - DIPA) (White III *et al.*, 2017).

The assembly module uses quality controlled sequence reads for *de novo* assembly (Step 2). Pre-assembly sub-setting uses the quality controlled reads as input then uses a read coverage normalization step based on k-mer frequency. The data is then subset to a target coverage using BBNorm (in BBMap tool suite (https://sourceforge.net/projects/bbmap/). This subset of high-quality reads is then used as input to ATLAS default assemblers SPAdes (i.e., metagenomic mode) (Bankevich *et al.,* 2012) for datasets <100 GB and MEGAHIT (Li *et al.,* 2015) for larger more complex datasets (e.g., soil). Assembled contigs are assessed for total length and percent read coverage. The final contigs can optionally be trimmed prior to determining open reading frames. Assembly output defaults include quality controlled contigs >1 kbp in length, with read coverage estimations >2x per contig, and with at least 40% coverage of reads across the entire contig.

The annotation module (Step 3) performs functional and taxonomic annotation of quality control contigs. Quality controlled contigs are translated to protein coding open reading frames (ORFs) using Prodigal (Hyatt *et al.,* 2012) in metagenome mode and annotated using DIAMOND (Buchfink *et al.,* 2015) blastp for protein-protein searching. DIAMOND blastp high-scoring pairs are filtered to user specified bitscore and e-value cut-offs (defaults >200 and $<1x10^{7}$, respectively). Functional annotation utilizes non-redundant RefSeq (O'Leary *et al.,* 2016), EggNOG (Huerta-Cepas *et al.,* 2016), dbCAN for CAZy families (Yin *et al.,* 2012),

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

ENZYME for enzyme commission number (EC) (Bairoch, 2000), and COG (Tatusov *et al.,* 2003) databases. ATLAS obtains KEGG (Kanehisa and Goto, 2000) (i.e., KO number) annotations from EggNOG reference database. ATLAS provides pre-formatted databases via FTP with subcommands to simplify the process of database downloading, formatting, and version tracking. The database ontologies and hierarchies are included within the annotation references for downstream analysis. A DNA-DNA database search module using the Lambda search tool (Hauswedell *et al.,* 2014) will be added to a future version ATLAS (Figure 1). This DNA-DNA database search module will annotate ribosomal internal transcribed spacers (ITS), small subunit (SSU), and large subunit (LSU) genes using Unite ITS (Abarenkov *et al*., 2016) and the Silva (SSU/LSU) (Pruesse *et al.,* 2007) databases (Figure 1).

For taxonomic annotation, ATLAS uses RefSeq high-scoring pairs along with NCBI's taxonomy assignments reference tree via a modified majority voting-method (MMVM) that utilizes lowest common ancestor (LCA) (Hanson *et al.,* 2016), to determine the lowest common ancestor represented across all ORFs present within a single contig. Assembly and annotation outputs from ATLAS can be directly used to create databases for proteome searches or as inputs for quantitation analysis (step 5, below).

The binning module (Step 4) of ATLAS uses MaxBin2 (Wu *et al.,* 2016) to bin genomes from metagenomes. There are two binning parameters for MaxBin2 in ATLAS; (1) differential coverage estimation by user specified samples or (2) within a single sample without multi-sample differential coverage mapping. For quality control of bins, we recommend the CheckM package (Parks *et al.,* 2015). However, future versions of ATLAS will include a bin quality control and annotation integrated into our MMVM taxonomic assignment

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

package.

The last module (Step 5) quantifies the coverage of the assembly by mapping reads using annotations from metagenomes and metatranscriptomes. Functional and taxonomic count data is obtained by mapping quality controlled reads to assembled contig annotations using BBMap, then parsed using featureCounts of the Subread package (Liao *et al*., 2014) to user specifications. This provides the final tabular output of functional annotations, expressed functions (if RNA-Seq is available), taxonomy, and taxonomy based functional annotations based on user specifications.

ATLAS is written in Python 3.5, implemented using Snakemake (Köster and Rahmann, 2012) workflow infrastructure for flexible scalability, trivial parallelism of workflow steps, and extensive data provenance for reproducibility. ATLAS is easily installed using bioconda (https://bioconda.github.io/): conda install --channel bioconda atlas. The source code of ATLAS is freely available at https://github.com/pnnl/atlas.

3 SUMMARY

ATLAS packages, databases, and workflows are easy to use, simple to install, modular, and user customizable. ATLAS provides a robust bioinformatics framework for metagenomic and metatranscriptomic data, where raw FASTQ files are fully processed into annotated tabular files for downstream analysis and visualization. ATLAS fills a major computational and analysis gap, namely the integration of quality control, assembly, annotation, binning and expression analysis, and provides a framework for integrated 'omics analysis.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

Acknowledgments

We thank Nathan Johnson for his assistance in preparing an excellent figure.

Funding

This research was provided by the PNNL Laboratory-Directed Research and Development (LDRD) Program at PNNL through the Microbiomes in Transition (MinT) Initiative and the Initiative integrated Plant-Atmosphere-Soil System (iPASS) Initiative. PNNL is a national laboratory operated by Battelle for the Department of Energy (DOE) under contract DE-AC06-76RL01830. A portion of the research was conducted using PNNL Institutional Computing (PIC) at PNNL and at EMSL, a national scientific user facility sponsored by the DOE Office of Biological and Environmental Research and located at PNNL.

References

Abarenkov,K. et al. (2016) Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden). *MycoKeys*, **16**, 1-15.

Albertsen,M. *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.,* **6**, 533-538.

Bankevich, A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.,* **19**, 455-477.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304-305.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.,* **12**, 59-60.

Butterfield,C.N. *et al.* (2016) Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ*, **4**, e2687.

Chen,I.A. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res*., **45**, D507-D516.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27-30.

Hanson,N.W. *et al.* (2016) LCA*: an entropy-based measure for taxonomic assignment within assembled metagenomes. *Bioinformatics,* **32**, 3535-3542.

Howe,A.C. *et al.* (2014) Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci*., **111**, 4904-4909.

Hauswedell,H. *et al.* (2014) Lambda: the local aligner for massive biological data. *Bioinformatics,* **30**, 49-355.

Huerta-Cepas,J. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*., **44**, D286-D293.

Hyatt,D. *et al.* (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics,* **28**, 2223-2230.

Imelfort,M. *et al.* (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ,* **2**, e603.

Jansson,J.K. (2011) Towards “Tera-Terra”: terabase sequencing of terrestrial metagenomes. *ASM Microbe,* **6**, 309-315.

Konwar,K.M. *et al.* (2013) MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics,* **14**, 202.

Köster,J. and Rahmann, S. (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520-2522.

Li,D. *et al.* (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics*, **31**, 1674-1676.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923-930.

Mason,O.U. *et al.* (2012) Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISMEJ*. **6**, 1715-1727.

Meyer,F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

taxonomic expansion, and functional annotation. *Nucleic Acids Res*., **44**, D733-D745.

Parks,D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.,* **25**, 1043-1055.

Prosser,J.I. (2015) Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology. *Nat. Rev. Microbiol*., **13**, 439-446.

Pruesse,E. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.,* **35**, 7188-7196.

Su,X. *et al.* (2014) Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. *PLoS One,* **9**, e89323

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* **4**, 41.

Treangen,T.J. *et al.* (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.,* **14**, R2.

White,R.A. III, *et al.* (2016a) Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. *Msystems*, **1**, e00045-16.

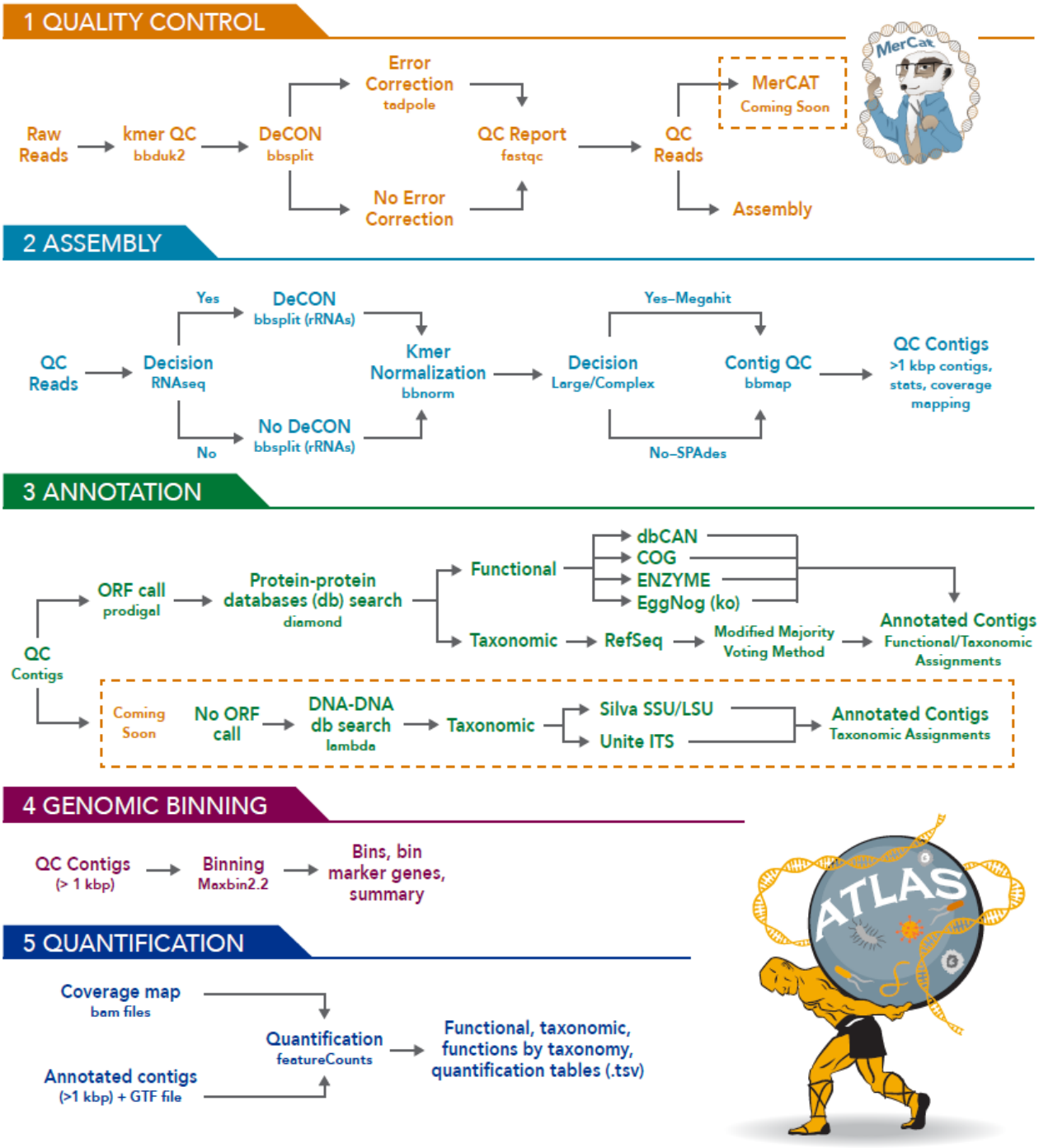
White,R.A. III, *et al.* (2016b) The past, present and future of microbiome analyses. *Nature Protocols*, 11, 2049-2053.

White,R.A. III, *et al*. (2017) MerCat: a versatile k-mer counter and diversity estimator for database-independent property analysis obtained from metagenomic and/or metatranscriptomic sequencing data. *PeerJ Preprints*, **5**, e2825v1

Wu,Y.W. *et al*. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.

Yin,Y. *et al*. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. **40**, 445-451.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017



244 **Figure 1: ATLAS workflow.** MerCat and Lambda based (DNA-DNA database) search
245 modules will be added to a future version of ATLAS.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2843v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017