Title

The DOE JGI Metagenome Workflow

Permalink

https://escholarship.org/uc/item/6gv1f5h8

Authors

Clum, Alicia

Huntemann, Marcel

Bushnell, Brian

et al.

Publication Date

2020

DOI

10.1101/2020.09.30.320929

Peer reviewed

The DOE JGI Metagenome Workflow

Alicia Clum,*[1] Marcel Huntemann,*[1] Brian Bushnell,[1] Brian Foster,[1] Bryce Foster,[1] Simon Roux,[1] Patrick P. Hajek,[1] Neha Varghese,[1] Supratim Mukherjee,[1] T.B.K. Reddy,[1] Chris Daum,[1] Yuko Yoshinaga,[1] Rekha Seshadri,[1] Nikos C Kyrpides,[1] Emiley A. Eloe-Fadrosh,[1] I-Min A. Chen,[1] Alex Copeland,[1] Natalia N. Ivanova[1]

Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA*
First Author and Second Author contributed equally to this work. Author order was determined randomly.

Compiled September 30, 2020
This is a draft manuscript, pre-submission

Address correspondence to Alicia Clum, aclum@lbl.gov.

ABSTRACT

The DOE JGI Metagenome Workflow performs metagenome data processing, including assembly, structural, functional, and taxonomic annotation, and binning of metagenomic datasets that are subsequently included into the Integrated Microbial Genomes and Microbiomes (IMG/M) comparative analysis system (I. Chen, K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J. White, R. Seshadri, et al, Nucleic Acids Rsearch, 2019) and provided for download via the Joint Genome Institute (JGI) Data Portal (https://genome.jgi.doe.gov/portal/). This workflow scales to run on thousands of metagenome samples per year, which can vary by the complexity of microbial communities and sequencing depth. Here we describe the different tools, databases, and parameters used at different steps of the workflow, to help with interpretation of metagenome data available in IMG and to enable researchers to apply this workflow to their own data. We use 20 publicly available sediment metagenomes to illustrate the computing requirements for the different steps and highlight the typical results of data processing. The workflow modules for read filtering and metagenome assembly are available as a Workflow Description Language (WDL) file (https://code.jgi.doe.gov/BFoster/jgi_meta_wdl.git). The workflow modules for annotation and binning are provided as a service to the user community at https://img.jgi.doe.gov/submit and require filling out the project and associated metadata descriptions in Genomes OnLine Database (GOLD) (S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, H. Katta, A. Mojica, I Chen, and N. Kyrpides, and T. Reddy, Nucleic Acids Research, 2018).

IMPORTANCE

The DOE JGI Metagenome Workflow is designed for processing metagenomic datasets starting from Illumina fastq files. It performs data pre-processing, error correction, assembly, structural and functional annotation, and binning. The results of processing are provided in several standard formats, such as fasta and gff and can be used for subsequent integration into the Integrated Microbial Genome (IMG) system where they can be compared to a comprehensive set of publicly available metagenomes. As of 7/30/2020 7,155 JGI metagenomes have been processed by the JGI Metagenome Workflow.

KEYWORDS: metagenomics, assembly, annotation, binning, SOP, IMG, JGI.

INTRODUCTION

Metagenomics, the study of the genetic content of natural microbial communities, provides a wealth of information about the structure and dynamics, perturbation, and resilience of ecosystems. Many tools are available for processing and analyzing metagenomic datasets including metaSPAdes(1) and MEGAHIT(2) for assembly, Prokka(3) and MG-RAST(4) for annotation and Kraken 2(5) for taxonomic identification, as well as integrated workflows such as SqueezeMeta(6) and MGnify(7). Here we present a metagenome workflow developed at the JGI which generates rich data in standard formats and has been optimized for downstream analyses ranging from assessment of functional and taxonomic composition of microbial communities to genome-resolved metagenomics and identification and characterization of novel taxa. This workflow is currently being used to analyze thousands of metagenomic datasets in a consistent and standardized manner.

RESULTS

The DOE JGI Metagenome Workflow aims to provide consistently processed metagenome data in standard formats suitable for a wide variety of analyses and interpretations across many studies and environmental samples. The workflow performs multiple quality checks and artifact removal, and provides a variety of summary statistics to assist users with the assessment of data quality and consistency. We illustrate the workflow using microbiomes from the Loxahatchee Nature Preserve in the Florida Everglades(8) as an example. In this follow-up study, sediment samples were collected and DNA was isolated by the students of Boca Raton Community High School, Boca Raton, from 4 different sites in the Loxahatchee Nature Preserve with 5 replicates at each site as previously described. DNA isolated from these samples was sequenced at the JGI using Illumina NovaSeq and standard library and sequencing protocols (Kapa HyperPrep library preparation kit, see Methods). Raw 2x150 reads were then processed by the DOE JGI Metagenome Workflow. The metadata for these samples can be found in Genomes OnLine Database (GOLD) (9) using GOLD study Gs0136122. Raw reads, as well as intermediate results and final assembly and annotation data can be found in the JGI Data Portal (https://genome.jgi.doe.gov) using JGI sequencing project identifiers linked to the GOLD study and Integrated Microbial Genome (IMG)(10) taxon identifiers provided in Table 1.

Read prefiltering and assembly results. The target amount of raw sequence data was 45 Gb per sample (300M reads). The number of high quality raw reads per sample after quality trimming, filtering, artifact and contamination removal is shown in Table 1. While the replicates from Loxahatchee West were sequenced somewhat more deeply than other samples, there is no significant difference in the amount of sequence generated for the other three sites. The prefiltering and assembly modules of the workflow automatically generate several conventional measures of assembly quality that are provided in the README files and can be accessed via the JGI Data Portal. A subset of these measures, which helps with assessing the consistency of the samples and identifying the outliers and artifacts, is shown in Table 1. Despite the fact that the samples from Loxahatchee North, South, and East received a very similar amount of sequence, as shown in Figure 1a, assembly statistics indicate that the replicates collected at the South site differ from the rest. Box-and-whisker plots for the L50 metric (Fig. 1b, the smallest length of contigs for which the sum of lengths makes up half of the dataset size) and percent of reads mapped to the assembly (Fig. 1c) demonstrate that assemblies of South site replicates are significantly more fragmented, as indicated by much lower L50, and have fewer reads mapped to them. This may be due to the fact that the sediment at the South site has a large amount of sand, which hindered isolation of sufficient quantities of high-quality DNA (Jonathan B. Benskin, personal communication) thereby resulting in a suboptimal library and poor assembly. Variation of library quality due to the quality and quantity of the source DNA may not be immediately obvious with a functional and/or taxonomic analysis of unassembled reads but is prominently brought to the researcher's attention by the DOE JGI Metagenome Workflow. It highlighted the differences between the South site and other sites due to the inconsistent performance of a sampling protocol, which may confound statistical analysis and obfuscate the true differences in functional and taxonomic composition.

Annotation results. The DOE JGI Metagenome Workflow performs feature prediction (also known as structural annotation) on the assembled sequence and functional annotation of the predicted protein-coding genes (CDSs). Similar to the filtering and assembly modules, the annotation module generates summary statistics helpful for identification of artifacts and outlier samples. These statistics are provided in README files via the JGI Data Portal and can be also found in the IMG database on the Metagenome Details page of each dataset. A subset of the annotation measures for Loxahatchee samples is provided in Table 2. The results of functional annotation of CDSs appear to be highly consistent across the four sites, with 65.75+/-1.2% of all CDSs assigned to Clusters of Orthologous Genes (COGs)(11), 14.25+/-0.44% assigned to TIGRfams(12), 62.65+/-0.81% assigned to Pfams(13), and 40.2+/-1.85% assigned to KEGG Orthology (KO) Terms(14). However, the results of feature prediction summarized in Figure 2 paint a different picture. Again, the South site is different from the other three sites, having more predicted CDSs per Kb of assembled sequence (Fig. 2a), and a much higher number of predicted rRNAs per Mb of assembled sequence (Fig. 2b). Remarkably, there is no significant difference in tRNA counts (Fig. 2c). These observations are consistent with lower contiguity South site assemblies, as reflected in their lower L50 (Fig. 1b), which in turn results in fragmentation of longer protein coding genes, as well as long 16S/18S and 23S/28S rRNA genes. On the other hand, tRNAs, which are on average less than 100 nt long, are largely unaffected by the fragmentation of assembled sequences. Importantly, protein-coding genes, which span a large interval of sequence lengths, will be affected unevenly, with the copy number of longer proteins appearing to be higher in more fragmented assemblies, while shorter proteins will show no differences. These factors have to be taken into account when comparing functional composition of different samples and attempting to correlate it with various environmental factors. The feature prediction and functional annotation module of the DOE JGI Metagenome Workflow provides other indicators of the quality and consistency of metagenomic data: the counts of eukaryotic 18S and 28S rRNAs suggest the presence and abundance of eukaryotic genomes in the sample, which could derive from the eukaryotic members of the microbial community and/or host DNA in host-associated microbiomes. On the other hand, the relatively low percent of CDSs assigned to COGs and Pfams may indicate the presence of a large viral fraction in the community, since viral proteins are poorly represented in these protein and domain classification systems. All of these characteristics of the assembled metagenome need to be taken into account in comparative analyses, as they may affect the results of the taxonomic and functional annotation of the communities.

Binning results. The DOE JGI Metagenome Workflow includes automated binning of assembled sequences, as well as an initial characterization of bins in terms of completeness and contamination and quality. The bins are assigned to high-quality (HQ) and medium-quality (MQ) categories based on Minimum Information about a Metagenome-Assembled Genome (MIMAG) standards(15). Bins that do not meet the standards for HQ or MQ are discarded. For HQ and MQ bins additional data processing is performed: bins are assigned a predicted lineage based on the NCBI(16) and GTDB-tk(17) taxonomy. The results of genome binning for the Loxahatchee samples are summarized in Table 3. The vast majority of the bins generated for these datasets are MQ, and represent a minor portion of the total assembly typical of high-complexity metagenomes from soil and sediment samples. Binning results for each dataset can be accessed via the JGI Data Portal and in IMG, where a number of tools for search, analysis, and comparison of metagenome bins are available.

Runtimes. We illustrate the typical computational requirements of the DOE JGI Metagenome Workflow on 20 samples from Loxahatchee Nature Preserve in Table 4. Filtering used Intel Xeon Gold 6140 processors using 32 vCPU and 324GB of RAM. For error correction, assembly, and mapping a mix of configurations was used. Some datasets were run on Intel Xeon Platinum 8000 series processors with different amounts of memory depending on the stage (16 vCPU and 128 GB of RAM for error correction, 64 vCPU and 512 GB of RAM for assembly, 32 vCPU and 256 GB of RAM for mapping). For others Intel Xeon Gold 6140 processors were used with 72 vCPU, 1.5 TB of RAM, and 5 TB of local disk. Runtime Assembly in Table 4 represents CPU hours for filtering, error correction, assembly, and mapping. For annotation, assembled metagenomic sequences were split into 10 MB shards. The splitting is performed by a wrapper script for optimal utilization of the JGI compute infrastructure and is not required to run the workflow. These 10 MB shards were then processed in parallel with each shard running on its own 2.3 GHz Haswell processor node with 128 GB of RAM. Binning was run on 2.3 GHz Haswell processor nodes with 128 GB of RAM.

DISCUSSION

The DOE JGI Metagenome Workflow provides automatic assembly, annotation and binning of metagenome datasets. It is largely based on publicly available software and databases supplemented with custom scripts and wrappers to control the workflow and to enable seamless integration of the input and output of different programs. Filtering, read correction, assembly, and mapping use a median of 2,004 CPU hours for current metagenomes such as the Loxahatchee sediment metagenomes, and can be performed on standard high-performance computing nodes such as Intel Xeon Platinum 8000 series processor with 256 GB memory. On average, the annotation module of the workflow (feature prediction, functional annotation and product name assignment) can process 1 million bp in 9 CPU hours on a 2.3 GHz Haswell processor (Intel Xeon Processor E5-2698 v3) node with 128 GB DDR4 2133 MHz memory. On the same Haswell node the entire binning workflow, from initial bin prediction, scaffold level cleanup, bin-level phylogenetic prediction, and estimation of contamination and completion, can process 100,000 scaffolds in an average of 13 CPU hours. The workflow modules for read filtering and metagenome assembly are available as a WDL file (https://code.jgi.doe.gov/BFoster/jgi_meta_wdl.git). The annotation and binning modules of the workflow are publicly available via the IMG system's submission site (https://img.jgi.doe.gov/submit), which accepts assembled metagenome sequences in fasta format and requires submission of sample and project metadata as a condition of annotation and binning services. We plan to continue to improve the workflow by updating reference database versions, extending the existing software and adding new tools that allow the identification and characterization of more features in the metagenome datasets, as well as improving the performance by making changes geared towards exploiting the specific infrastructure the workflow is utilizing.

MATERIALS AND METHODS

Data input. Standard metagenomes at JGI currently use 100 ng of genomic DNA, sheared to 300 bp using the Covaris LE220 and size selected with SPRI using TotalPure NGS beads (Omega Bio-tek). The fragments are treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-HyperPrep kit (KAPA biosystems) to create an unamplified Illumina library which is then sequenced 2x150 bp on the Illumina NovaSeq 6000 using S4 flowcells. The workflow can be used on paired-end Illumina datasets; kmer sizes for assembly should be adjusted if reads are shorter than 150 bp.

Sequence data preprocessing. Data is processed using Real-time Analysis (RTA) version 3.4.4 (https://support.illumina.com/downloads.html). BBDuk version 38.79 from the BBTools package (https://jgi.doe.gov/data-and-tools/bbtools/) is used to remove contamination, trim reads that contain adapter sequence and quality trim reads where quality drops to 0. Furthermore, it is used to remove reads that contain 4 or more 'N' bases, have an average quality score across the read of less than 3 or have a minimum length less than or equal to 51 bp or 33% of the full read length. Homopolymer streches of 5 Gs or more at the ends of reads are removed. Reads that can be mapped with BBMap from BBTools to masked human, cat, dog and mouse references at 93% identity are separated into a "chaff" file and not used in assembly. In an abundance of caution, reads aligned to common microbial contaminants described in the literature such as *Ralstonia pickettii* and *Acinetobacter calcoaceticus*(18, 19, 20, 21) are also separated into a "chaff" file. Masked references can be found at https://portal.nersc.gov/dna/microbial/assembly/bushnell/fusedERPBBmasked2.fa.gz. For convenience chaff files are provided on JGI's data portal.

Assembly. Filtered reads are error corrected using bbcms version 38.44 from BBTools with a minimum count of 2 and a high count fraction of 0.6. Bbcms uses a count-min sketch to store kmer counts, making it a scaleable solution for error correction of metagenomic datasets. For computational efficiency, interleaved fastq files are split into two separate files. These split error corrected files are assembled with metaSPAdes version 3.13.0 using the "metagenome" flag, running the assembly module only (i. e. without error correction) with kmer sizes 33,55,77,99,127. Contigs that are smaller than 200 bp are discarded. Filtered reads are mapped back to contigs larger than 200 bp using BBMap 38.44 with "interleaved" as true, "ambiguous" as random, and "covstats" option specifying a contig coverage file for subsequent analysis of abundance of various populations and genes. The coverage file contains information on average fold coverage, length, GC content, percent of bases covered, number of reads by strand, read GC, median fold and standard deviation of coverage.

Feature prediction. The assembled contigs are passed on to the annotation module of the workflow, which first predicts non-coding RNA genes (tRNAs, rRNAs and other RNAs), followed by the identification of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and protein coding genes (CDSs) as shown in Fig. 3a. Prediction of tRNAs is performed using tRNAscan-SE 2.0.6(22) in "bacterial" and "archaeal" search modes. This allows the workflow to select the best annotation mode and ensure higher annotation accuracy for metagenomic contigs of different taxonomic origin, since many archaeal tRNAs cannot be predicted in "bacterial" or "general" modes. For each contig the number of tRNAs with known isotype returned by each mode is compared. The results from the mode with the higher number of tRNAs with known isotype get reported and if both modes have returned the same number, the results from the "bacterial" mode are included in the final annotation. Ribosomal RNA genes (5S, 16S, 23S) as well as other non-coding RNA genes (ncRNAs) including tmRNA, antisense RNAs, etc. and RNA regulatory features, such as various binding sites and motifs ("misc_bind", "misc_feature", "regulatory") are identified by comparing the contigs via cmsearch from the INFERNAL 1.1.3 package(23) against the Rfam 13.0 database(24) using the trusted cutoffs parameter (–cut_tc). If any reported hits are overlapping even by 1 bp and they belong to the same Rfam class, the lower scoring of the two is discarded. CRISPR elements are identified using a version of CRT-CLI 1.2 modified in-house as described previously(25). For the search parameter the minimum and maximum repeat lengths are set to 20 and 50 bp, respectively, whereas the minimum and maximum spacer length is set to 20 and 60 bp, respectively. The search window size is set to 7 bp and an element needs to have at least three repeats to get reported. Protein-coding genes are predicted via a combination of Prodigal 2.6.3(26) and GeneMarkS-2 1.07(27). Prodigal is executed in "meta" mode and with the '-m' argument so that genes won't be built across runs of Ns. GeneMark is run with '–Meta mgm_11.mod' and '–incomplete_at_gaps 30'. CDS shorter than 75 bp (25 amino acids) are discarded. The last step of the feature prediction combines the results from all tools and attempts to resolve overlaps between features of different types. Two features are considered to overlap if they share more than 10 bp or more than 90 bp in the case of two CDSs. The regulatory RNA features (misc_bind, misc_feature and regulatory) are allowed to overlap with any other feature type. In case of an overlap between other types of features the lower-ranked feature gets removed. The feature ranking order is rRNA > tRNA > ncRNA, tmRNA > CRISPR > GeneMarkS-2 > Prodigal. Before deleting a CDS that overlaps with another feature over its 5' end, first an attempt is made to find an alternative start site for the protein-coding gene that removes the overlap. Functional annotations of RNA features are based on their descriptions provided by the tool or database used to predict them: tRNA isotype (amino acid and codon) as well as potential pseudogene annotation is provided by tRNAscan-SE, while product names for rRNAs, ncRNAs and regulatory RNA features are derived from the corresponding Rfam models. Functional annotation and product name assignment for protein sequences of the non-overlapping CDSs is performed by the functional annotation module.

Functional annotation. Functional annotation for metagenomes consists of associating protein-coding genes with KO terms, Enzyme Commission (EC) numbers, COG assignments, SMART domains, SUPERFAMILY assignments, CATH-FunFam annotations, Pfam and TIGRFAM annotations as shown in Fig. 3b. Genes are associated with KO terms and EC numbers based on results of sequence similarity search of metagenome proteins against a reference database of isolate proteomes using lastal 1066 from the LAST package(28) with default parameters. The reference database of isolate proteomes (IMG-NR) is composed of all non-redundant protein sequences encoded by public, high quality genomes in the current version of the IMG database. For each metagenome protein the top five LAST hits are considered. At least two of the top five hits need to have a KO assignment and all hits that have a KO assignment need to list the same combination of KO terms. If both conditions are met, the same combination of KO terms is assigned to the query gene if the alignment length for any of the hits with KO assignment covers at least 70% of the shorter one of query and subject. Proteins are associated with COGs by comparing protein sequences to the COG Hidden Markov Models (HMMs) created from the updated 2014 models using HMMER 3.1b2(29), and a thread-optimized version of hmmsearch(30), with a per-domain e-value cutoff (–domE) of 0.01. Since an alignment of a protein to the model may be fragmented, i. e. there may be multiple aligned segments of the two, these are concatenated and their cumulative alignment length calculated. If the cumulative alignment length is less than 70% of the shorter of the two (the protein or the model), such a hit is discarded. In addition, if a protein has hits to different COG models and their alignments overlap significantly (by more than 10% of the length of the shorter model), the hit to the model with the lower full sequence bitscore is discarded; for significantly overlapping hits with the same bit score, the hit with the higher e-value gets removed. The same thread-optimized version of hmmsearch, as well as parameters, filtering and overlap resolution rules are used to assign protein sequences to the 01_06_2016 version of the SMART database(31), the 1.75 version of the SUPERFAMILY database(32) and the frozen set of the 4.2.0 version of the CATH-FunFam database(33). Proteins are associated with Pfam-A by comparing protein sequences to version 30 of the Pfam database using thread-optimized version of hmmsearch from HMMER 3.1b2. Model-specific trusted cut-offs are used with (–cut_tc option in hmmsearch) and for overlapping hits that belong to the same Pfam clan the lower scoring one is removed. Proteins are associated with TIGRFAMs using version 15.0 of the TIGRFAM database and hmmsearch with a per-domain e-value cutoff (–domE) of 0.01. All hits that don't cover at least 70% of the shorter of the protein or model get discarded. Furthermore if two hits overlap for more than 10% of the length of the shorter model, the hit to the lower scoring model (by bit score) is discarded. Protein product names are assigned based on the name of their associated protein families in the order of priority KO term > TIGRFAM > COG > Pfam. If multiple TIGRFAMs with different isology types are associated with a protein, only one TIGRFAM is assigned in the order equivalog > hypoth_equivalog > paralog > exception > equivalog_domain > hypoth_equivalog_domain > paralog_domai n> subfamily > superfamily > subfamily_domain > domain > signature > repeat. Proteins without any of the above assignments are annotated as "hypothetical protein". Proteins associated with multiple protein families of the same type (KO term, TIGRFAM, COG or Pfam) are annotated with a product name consisting of concatenation of individual protein family names joined with "/". Multiple repetitions of the same protein family are collapsed to a single instance. The contig coverage information is used to calculate so-called "estimated gene copies", whereby the number of genes in a certain group, such as a COG or Pfam protein family, is multiplied by the average coverage of the contigs from which these genes were predicted. This step is important for accurate estimation of abundance of protein families and takes into account different abundance of populations found in the assembled metagenome sequences.

Taxonomic annotation. For taxonomic annotation of metagenomes the best LAST (28) hits of CDSs computed as described above for KO term assignment are used. The taxonomy of the best hit is assigned to each metagenome protein. The taxonomy of metagenome contigs ("scaffold lineage") is predicted based on the majority rule, whereby the lineage at the lowest taxonomic rank to which at least 50% CDSs encoded by the metagenomic contig have hits is assigned. Similar to protein family annotations, contig coverage information is used to estimate the abundance of various lineages in the community by multiplying contig counts by their average coverage.

Binning. The assembled contigs and coverage file generated per metagenome is used as input to the MetaBAT v2.12.1(34) program to generate genome bins based on the consistency of coverage and tetranucleotide frequency. The genome bins then undergo contamination removal, wherein the per scaffold phylum information generated by the annotation module ("scaffold lineage") is used to remove scaffolds per bin that are not assigned to the predominant phylum. The post- processed bins are fed to the CheckM v1.0.12(35) program to determine genome completion and contamination estimates. These estimates along with the per scaffold rRNA and tRNA information generated by the annotation module, is used to assign HQ or MQ value to each bin, per MIMAG standards. The HQ and MQ bins are then subject to phylogenetic lineage determination by two methods. First, an internal IMG program computes the phylogenetic lineage per genome bin using the per scaffold lineage generated by the annotation module. Next, the GTDB-tk v0.2.2 program computes per bin lineage by placing them into domain-specific, concatenated protein reference trees. The high and medium quality bins, along with the corresponding data processing metadata, are loaded into IMG for user access and download.

Pre-formatted tables. To assist with preparing publications 9 tables are generated. Information on what is contained in each table is described in Table 5.

Availability of data and materials. The metadata for these samples can be found in GOLD (https://gold.jgi.doe.gov/) using GOLD study Gs0136122. Raw reads, as well as intermediate results and final assembly and annotation data can be found in the JGI Data Portal (https://genome.jgi.doe.gov) by following links from the GOLD study or by using IMG taxon identifiers provided in Table 1. A WDL for filtering and genome assembly (v1.0) is available at https://code.jgi.doe.gov/BFoster/jgi_meta_wdl.git. IMG for annotation (v5.0.19) and binning (v1.0) is available at https://img.jgi.doe.gov/.

ACKNOWLEDGMENTS

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank the Advanced International Certificate of Education (AICE) Biology class at Florida's Boca Raton Community High School and Mr. Jonathan Benskin for partnering with the JGI on the Everglades metagenome studies.

REFERENCES

1. **Nurk S, Meleshko D, Korobeynikov A, Pevzner PA**. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27 (5):824–834. doi:10.1101/gr.213959.116.
2. **Li D, Liu CM, Luo R, Sadakane K, Lam TW**. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31 (10):1674–1676. doi:10.1093/bioinformatics/btv033.
3. **Seemann T**. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069. doi: 10.1093/bioinformatics/btu153.
4. **Keegan KP, Glass EM, Meyer F**. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Microb Environ Genom (MEG) Methods Mol Biol p 207–233. doi: 10.1007/978-1-4939-3369-3_13.
5. **Wood DE, Lu J, Langmead B**. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. doi:10.1101/762302.
6. **Tamames J, Puente-Sánchez F**. 2019. SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. Front Microbiol 9. doi:10.3389/fmicb.2018.03349.
7. **Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, et al**. 2019. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res doi:10.1093/nar/gkz1035.
8. **Abraham BS, Caglayan D, Carrillo NV, Chapman MC, Hagan CT, Hansen ST, Jeanty RO, Klimczak AA, Klingler MJ, Kutcher TP, et al**. 2020. Shotgun metagenomic analysis of microbial communities from the Loxahatchee nature preserve in the Florida Everglades. Environ Microbiome 15 (1). doi:10.1186/s40793-019-0352-4.
9. **Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A, Chen IMA, Kyrpides NC, Reddy T**. 2018. Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res 47 (D1). doi:10.1093/nar/gky977.
10. **Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, et al**. 2018. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47 (D1). doi:10.1093/nar/gky901.
11. **Galperin MY, Makarova KS, Wolf YI, Koonin EV**. Jan 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Oxford University Press https://www.ncbi.nlm.nih.gov/pubmed/25428365.
12. **Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E**. 2012. TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res 41 (D1). doi:10.1093/nar/gks1234.
13. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al**. 2015. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44 (D1). doi:10.1093/nar/gkv1344.
14. **Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K**. Jan 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Oxford University Press https://www.ncbi.nlm.nih.gov/pubmed/27899662.
15. **Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al**. Aug 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Publishing Group https://www.nature.com/articles/nbt.3893.

16. **Federhen S**. 2011. The NCBI Taxonomy database. Nucleic Acids Res 40 (D1). doi:10.1093/nar/gkr1178.
17. **Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH**. Nov 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press https://academic.oup.com/bioinformatics/article/36/6/1925/5626182.
18. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW, et al**. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12 (1). doi:10.1186/s12915-014-0087-z.
19. **Tanner MA, Goebel BM, Dojka MA, Pace NR**. 1998. Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. Appl Environ Microbiol 64 (8):3110–3113. doi:10.1128/aem.64.8.3110-3113.1998.
20. **Onstott TC, Moser DP, Pfiffner SM, Fredrickson JK, Brockman FJ, Phelps TJ, White DC, Peacock A, Balkwill D, Hoover R, et al**. 2003. Indigenous and contaminant microbes in ultradeep mines. Environ Microbiol 5 (11):1168–1191. doi:10.1046/j.1462-2920.2003.00512.x.
21. **Kulakov LA, Mcalister MB, Ogden KL, Larkin MJ, Ohanlon JF**. 2002. Analysis of Bacteria Contaminating Ultrapure Water in Industrial Systems. Appl Environ Microbiol 68 (4):1548–1555. doi: 10.1128/aem.68.4.1548-1555.2002.
22. **Chan PP, Lowe TM**. 2019. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. U.S. National Library of Medicine https://www.ncbi.nlm.nih.gov/pubmed/31020551.
23. **Nawrocki EP, Eddy SR**. Nov 2013. Infernal 1.1: 100-fold faster RNA homology searches. Oxford University Press https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810854/.
24. **Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI**. 2017. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46 (D1). doi:10.1093/nar/gkx1038.
25. **Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P**. Jun 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BioMed Central https://www.ncbi.nlm.nih.gov/pubmed/17577412.
26. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ**. Mar 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BioMed Central https://www.ncbi.nlm.nih.gov/pubmed/20211023.
27. **Lomsadze A, Gemayel K, Tang S, Borodovsky M**. Jul 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Cold Spring Harbor Laboratory Press https://www.ncbi.nlm.nih.gov/pubmed/29773659/.
28. **Kielbasa SM, Wan R, Sato K, Horton P, Frith MC**. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res 21 (3):487–493. doi:10.1101/gr.113985.110.
29. **Mistry J, Finn RD, Eddy SR, Bateman A, Punta M**. Jul 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Oxford University Press https://www.ncbi.nlm.nih.gov/pubmed/23598997.
30. **Arndt W**. 2016. Modifying HMMER3 to Run Efficiently on the Cori Supercomputer Using OpenMP Tasking. 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) Pages: 239-246. http://www.hicomb.org/papers/HICOMB2018-04.pdf.
31. **Letunic I, Bork P**. 2017. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46 (D1). doi:10.1093/nar/gkx922.
32. **Gough J, Karplus K, Hughey R, Chothia C**. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313 (4):903–919. doi:10.1006/jmbi.2001.5080.
33. **Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A, et al**. 2018. CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res 47 (D1). doi:10.1093/nar/gky1097.
34. **Kang DD, Froula J, Egan R, Wang Z**. Aug 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ Inc. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4556158/.
35. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW**. Jul 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Cold Spring Harbor Laboratory Press https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4484387/.

TABLE 1 Sequencing and assembly statistics for 20 samples (4 sites, 5 replicates each) from the Loxahatchee Nature Preserve.

Sample Name	IMG Taxon ID	Latitude and Longitude	Filtered Reads (M)	Contigs (M)	Contigs (Mb)	Contig L50	Reads Mapped to Assembly (%)
Lox_West_1	3300038551	26.469/-80.443	432.41	6.37	4281.10	783	61.64
Lox_West_2	3300038408	26.469/-80.443	335.90	5.01	3329.57	763	58.92
Lox_West_3	3300038552	26.469/-80.443	478.21	7.22	4968.38	814	65.04
Lox_West_4	3300038469	26.469/-80.443	447.07	6.49	4393.53	792	62.92
Lox_West_5	3300038470	26.469/-80.443	347.74	4.89	3172.10	734	53.95
Lox_North_1	3300038409	26.677/-80.375	265.39	3.12	2017.05	736	52.06
Lox_North_2	3300038421	26.677/-80.375	294.36	3.60	2255.03	697	52.26
Lox_North_3	3300038558	26.677/-80.375	355.61	4.86	2909.37	646	44.28
Lox_North_4	3300038550	26.677/-80.375	296.91	3.86	2361.02	666	43.15
Lox_North_5	3300038422	26.677/-80.375	240.01	3.14	1896.85	654	41.56
Lox_South_1	3300038401	26.358/-80.298	241.50	2.87	1328.12	445	23.17
Lox_South_2	3300038549	26.358/-80.298	335.62	4.83	2379.73	481	31.57
Lox_South_3	3300038402	26.358/-80.298	240.39	2.93	1406.77	469	25.33
Lox_South_4	3300038403	26.358/-80.298	244.71	3.00	1514.26	496	27.91
Lox_South_5	3300038663	26.358/-80.298	253.01	3.31	1771.86	538	33.78
Lox_East_1	3300038454	26.502/-80.223	299.62	3.99	2746.17	819	54.72
Lox_East_2	3300038455	26.502/-80.223	322.84	4.18	2834.88	795	52.17
Lox_East_3	3300038431	26.502/-80.223	292.44	3.65	2385.22	740	46.35
Lox_East_4	3300038410	26.502/-80.223	247.69	3.49	2320.95	761	52.70
Lox_East_5	3300038468	26.502/-80.223	266.29	3.75	2317.21	670	46.14

TABLE 2 Annotation statistics for 20 samples (4 sites, 5 replicates each) from the Loxahatchee Nature Preserve.

Sample Name	IMG Taxon ID	Contigs (Mb)	CRISPR	Predicted							CDSs assigned to (% total)			
				CDSs (M)	16S rRNA	18S rRNA	23S rRNA	28S rRNA	5S rRNA	tRNAs	COGs	TIGRfam	Pfam	KEGG
Lox_West_1	3300038551	2859.7	391	4.413	943	2	1559	8	384	18675	67	15	63	39
Lox_West_2	3300038408	2204.4	250	3.412	742	8	1209	14	377	19124	65	14	63	39
Lox_West_3	3300038552	3396.0	458	5.245	1084	8	1735	12	560	29892	64	14	62	38
Lox_West_4	3300038469	2949.2	420	4.534	957	5	1612	9	529	27020	65	14	62	39
Lox_West_5	3300038470	2061.2	242	3.218	722	6	1292	11	384	18675	66	15	63	40
Lox_North_1	3300038409	1293.3	339	1.994	574	15	973	22	289	13655	65	14	62	39
Lox_North_2	3300038421	1408.6	372	2.189	644	16	1029	20	292	14843	65	14	62	39
Lox_North_3	3300038558	1761.8	255	2.818	877	11	1432	9	382	19094	65	14	62	41
Lox_North_4	3300038550	1460.0	171	2.333	736	9	1209	13	345	16512	65	14	62	40
Lox_North_5	3300038422	1161.4	145	1.860	589	9	978	12	268	12534	66	14	62	40
Lox_South_1	3300038401	571.3	58	1.011	454	21	863	33	150	4992	67	14	63	44
Lox_South_2	3300038549	1139.7	137	1.977	622	15	1187	27	237	10120	67	14	63	42
Lox_South_3	3300038402	653.7	83	1.159	421	18	854	33	140	5752	68	14	64	44
Lox_South_4	3300038403	750.8	105	1.286	465	14	895	20	174	6767	67	14	63	43
Lox_South_5	3300038663	950.1	87	1.589	493	5	911	7	190	8662	68	14	64	42
Lox_East_1	3300038454	1852.5	219	2.803	691	11	1041	15	334	16789	65	15	63	39
Lox_East_2	3300038455	1891.4	259	2.879	678	10	1158	20	322	17682	65	15	63	39
Lox_East_3	3300038431	1551.8	156	2.396	615	8	1020	12	249	13642	66	15	64	40
Lox_East_4	3300038410	1529.8	208	2.359	557	8	942	12	246	14059	65	14	62	39
Lox_East_5	3300038468	1431.5	196	2.232	581	13	966	18	271	12773	64	14	61	38

TABLE 3 Binning statistics for 20 samples (4 sites, 5 replicates each) from the Loxahatchee Nature Preserve.

Sample Name	IMG Taxon ID	High Quality Bins			Medium Quality Bins		
		Number	Size (Mb)	Contigs	Number	Size (Mb)	Contigs
Lox_West_1	3300038551	0	0	0	9	18.97	3041
Lox_West_2	3300038408	0	0	0	12	24.90	3854
Lox_West_3	3300038552	0	0	0	11	27.55	3251
Lox_West_4	3300038469	0	0	0	10	19.56	2542
Lox_West_5	3300038470	0	0	0	6	16.68	2542
Lox_North_1	3300038409	0	0	0	4	12.25	2100
Lox_North_2	3300038421	0	0	0	4	15.51	2241
Lox_North_3	3300038558	1	1.25	35	12	22.80	3749
Lox_North_4	3300038550	1	1.29	46	6	7.24	1180
Lox_North_5	3300038422	1	1.26	39	6	10.36	1751
Lox_South_1	3300038401	0	0	0	1	3.14	498
Lox_South_2	3300038549	1	7.34	152	3	4.06	711
Lox_South_3	3300038402	0	0	0	0	0	0
Lox_South_4	3300038403	0	0	0	1	0.83	103
Lox_South_5	3300038663	0	0	0	2	3.50	528
Lox_East_1	3300038454	2	4.16	365	6	18.80	2485
Lox_East_2	3300038455	0	0	0	4	8.41	1150
Lox_East_3	3300038431	0	0	0	7	16.21	2177
Lox_East_4	3300038410	0	0	0	8	22.64	3269
Lox_East_5	3300038468	0	0	0	10	21.20	2753

TABLE 4 CPU hours for different modules in the JGI Metagenome Workflow on 20 samples from Loxahatchee Nature Preserve.

Sample Name	IMG Taxon ID	Assembly	Feature Prediction	Functional Annotation	Binning
Lox_West_1	3300038551	3576.16	12423.68	8980.48	264.9
Lox_West_2	3300038408	2751.16	12572.8	6836.48	110.3
Lox_West_3	3300038552	4155.04	13522.56	10065.92	367.6
Lox_West_4	3300038469	3699.6	12163.84	9695.36	225.9
Lox_West_5	3300038470	2713.03	8332.16	7274.88	90.0
Lox_North_1	3300038409	1801.75	5659.52	3489.28	23.9
Lox_North_2	3300038421	2064.19	6092.85	3990.40	23.5
Lox_North_3	3300038558	2455.81	7430.4	6223.36	14.9
Lox_North_4	3300038550	1944.75	6147.2	4270.08	11.0
Lox_North_5	3300038422	1692.39	5338.8	3429.76	9.3
Lox_South_1	3300038401	1540.82	62.72	29.30	2.1
Lox_South_2	3300038549	1534.45	88.55	62.23	7.1
Lox_South_3	3300038402	1556.06	78.19	33.38	1.9
Lox_South_4	3300038403	1621.84	61.65	36.12	5.7
Lox_South_5	3300038663	1771.97	72.76	53.28	7.4
Lox_East_1	3300038454	2086.37	114.67	99.84	59.3
Lox_East_2	3300038455	2298.94	117.02	89.79	62.5
Lox_East_3	3300038431	2153.02	102.98	100.34	31.7
Lox_East_4	3300038410	1877.78	99.47	84.15	35.4
Lox_East_5	3300038468	1795.02	101.5	66.69	25.3

TABLE 5 Preformatted tables

Table number	Table information
1	Study information
2	Sample information
3	Library information
4	Sequence process
5	Assembly statistics
6	Annotation parameters
7	Functional diversity
8	Metagenome properties
9	Taxonomic composition

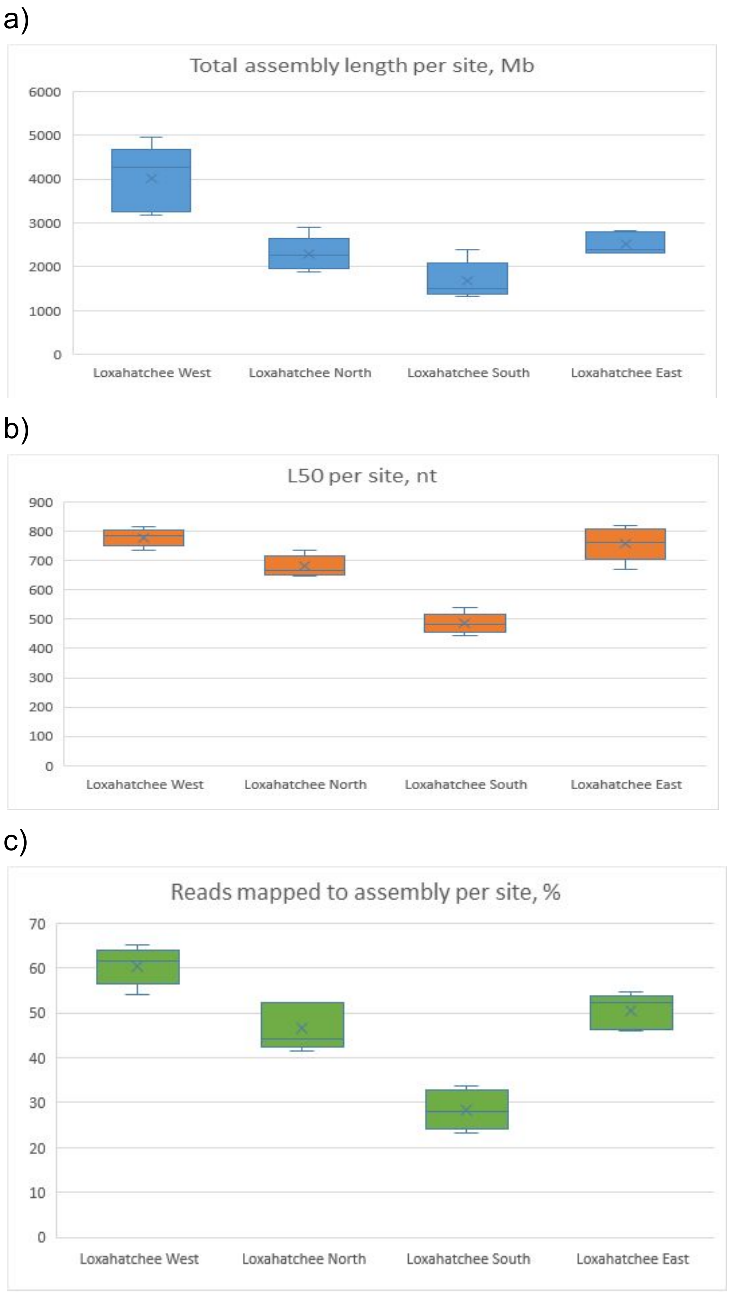

Figure 1. Box-and-whiskers plots of sequencing and assembly statistics for 4 sites in the Loxahatchee Nature Preserve. a) Total assembly length per site, Mb. b) L50 (the smallest length of contigs whose sum of lengths makes up half of the dataset size) per site, nt. c) Reads mapped to the assembly as percent of total number of reads generated for sample, per site, %.

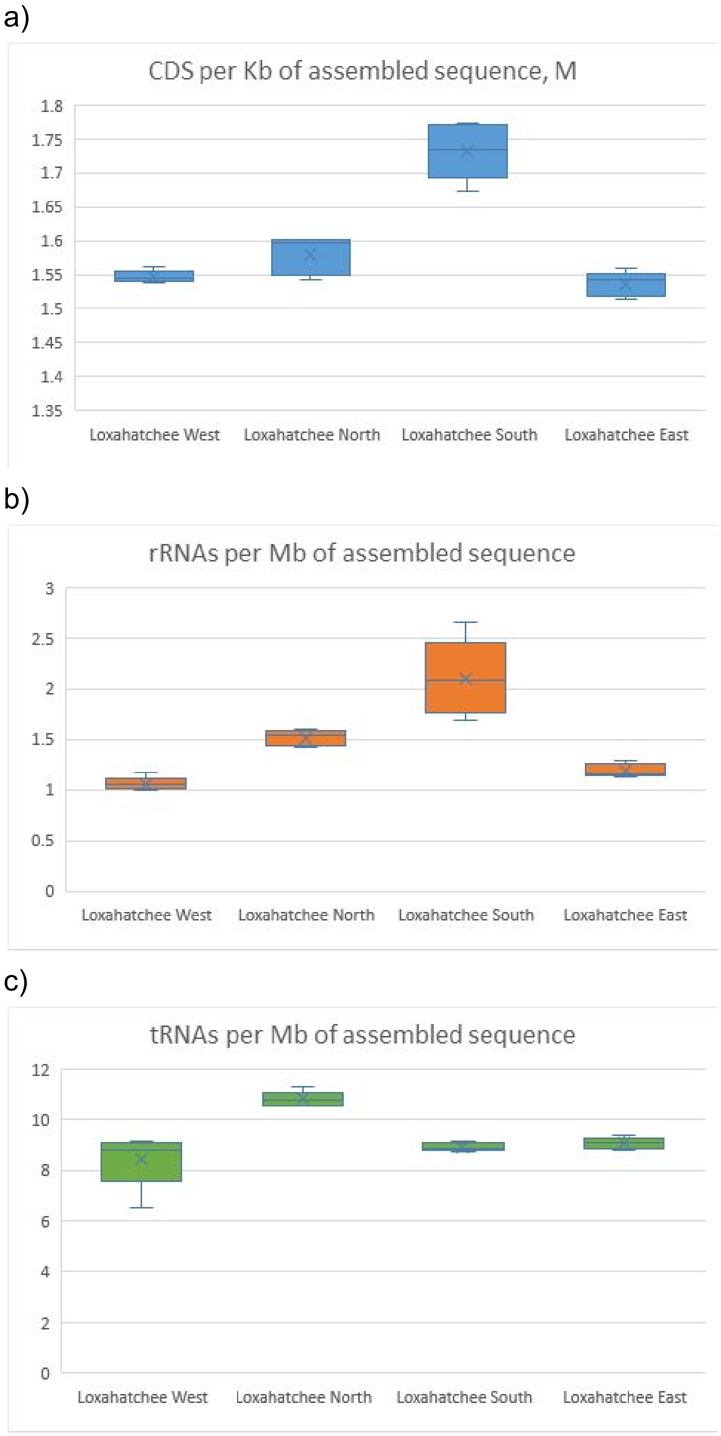

Figure 2. Box-and-whisker plots summarizing the results of structural annotation for 20 samples (4 sites, 5 replicates each) from the Loxahatchee Nature Preserve. a) Number of predicted CDSs per Kb of assembled sequence, millions. b) Number of predicted rRNA genes per Mb of assembled sequence. c) Number of predicted tRNA genes per Mb of assembled sequence.

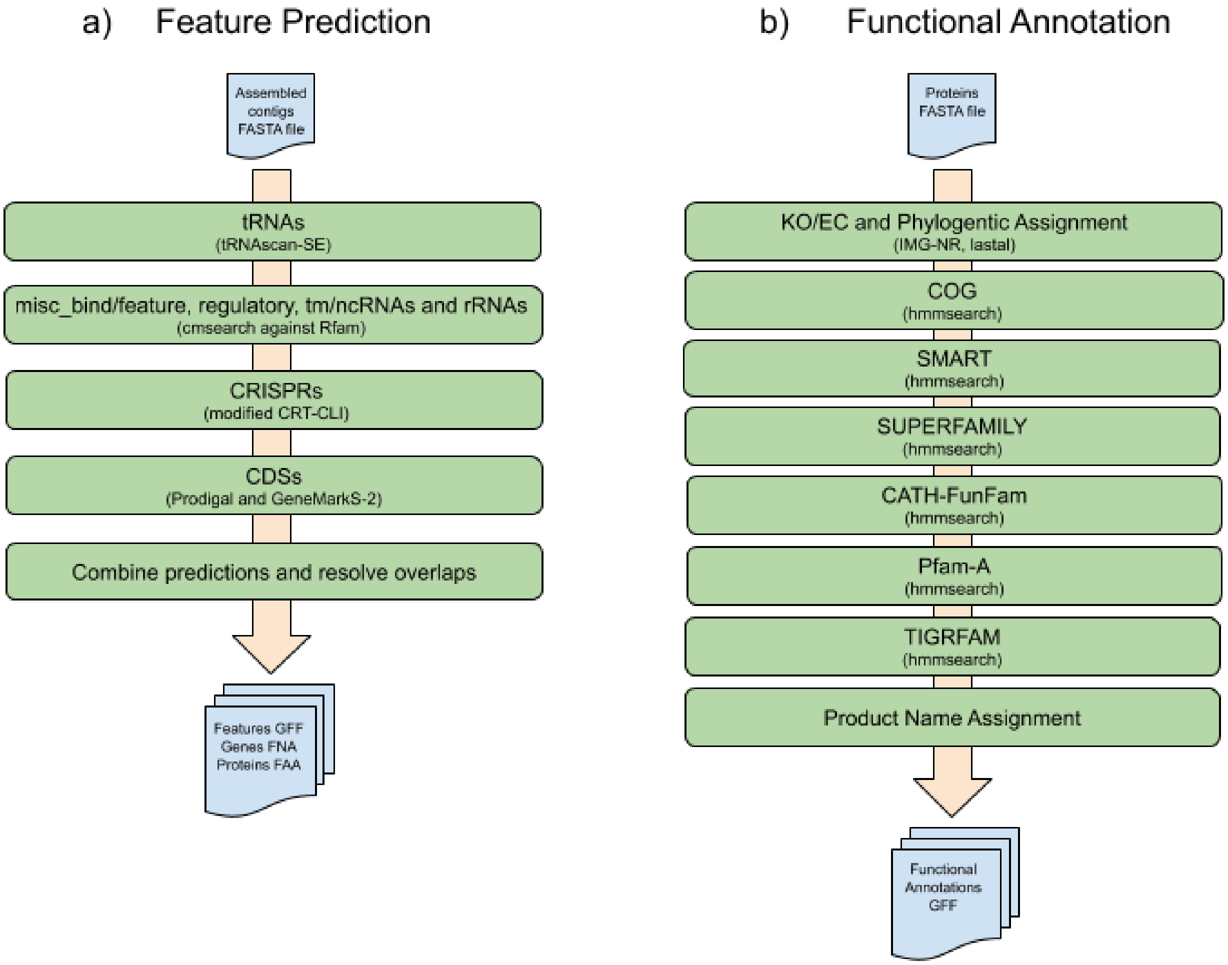

Figure 3. Workflow of a) Feature Prediction and b) Functional Annotation