The Standard Operating Procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4)

Marcel Huntemann[1], Natalia N. Ivanova[1], Konstantinos Mavromatis[1], H. James Tripp[1], David Paez-Espino[1], Krishnaveni Palaniappan[2], Ernest Szeto[2], Manoj Pillay[2], I-Min A. Chen[2], Amrita Pati[1], Victor M. Markowitz[2], Nikos C. Kyrpides[1]

[1]Genome Biology Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA
[2]Biosciences Computing, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA

Abstract

The DOE-JGI Metagenome Annotation Pipeline (MAP v.4) performs structural and functional annotation for metagenomic sequences that are submitted to the Integrated Microbial Genomes with Microbiomes (IMG/M) system for comparative analysis. The pipeline runs on nucleotide sequences provided via the IMG submission site (http://img.jgi.doe.gov/submit). Users must first define their analysis projects in GOLD (https://gold.jgi-psf.org/) and then submit the associated sequence datasets consisting of scaffolds/contigs with optional coverage information and/or unassembled reads in fasta and fastq file formats. The MAP processing consists of feature prediction including identification of protein-coding genes, non-coding RNAs and regulatory RNAs, as well as CRISPR elements. Structural annotation is followed by functional annotation including assignment of protein product names and connection to various protein family databases.

Introduction

The DOE-JGI Metagenome Annotation Pipeline (MAP) supports the structural and functional annotation of metagenomic datasets submitted to the Integrated Microbial Genomes with Microbiomes (IMG/M) system [1]. The annotation includes the prediction of CRISPR elements, non-coding and protein-coding genes, and ends with the assignment of a product name and the prediction of functions for each gene. The annotated metagenomic datasets produced by MAP are integrated into IMG/M where they can be analyzed or revised in the context of a comprehensive set of publicly available genomes and metagenomes.

The DOE-JGI MAP requires a multi-FASTA file of assembled nucleotide sequences and/or a fastq file containing unassembled 454, Illumina or PacBio reads as input. To submit sequence datasets for annotation they need to be linked with an analysis project that previously has been specified in the Genomes OnLine Database [2] (GOLD: https://gold.jgi-psf.org/).

Annotation of metagenomic sequences in MAP is organized in three stages: sequence data pre-processing, gene prediction, functional annotation and phylogenetic lineage prediction for scaffolds/contigs. Identification of genes and repeats produces a GFF file without any functional information for the predicted genes. These protein coding genes are then assigned with a function followed by integration into IMG.

Implementation

The MAP stages and individual steps are further described below.

Sequence Data Preprocessing

In order to reduce noise and eliminate low-quality and low-complexity sequences, such as oligomers of sequencing primers and adaptors, a preprocessing step is implemented for all the metagenomic datasets (Figure 1(i)). First, ambiguous nucleotides in the sequence datasets are replaced by N's, while sequences having any characters other than {A,C,G,T,N} are removed from further processing. Additionally, all sequences are renamed in order to ensure that there are no duplicate sequence names and the names comply with the requirements of all the tools employed in subsequent stages. The pipeline creates a file that maps the old sequence names to the new ones.

For fastq files with unassembled reads a check for the matching lengths of each sequence and its quality values is applied. Sequences shorter than 150 bp are removed; unassembled 454 reads longer than 1000 bp are also removed.

Second, the sequences are trimmed in order to remove low-quality regions and trailing 'N's. In the case of unassembled reads, quality data from fastq files is used with Lucy [3] with a threshold of Q13 for Illumina reads and Q20 for 454 reads in order to identify and trim regions of low-quality at the ends of the reads. In the case of Illumina reads the longest contiguous sub-sequence that passes the Q13 threshold for all residues is retained. Unassembled 454 and Illumina reads containing more than five occurrences of 'N's are removed. Sequences shorter than 150 bp after trimming are also removed. The trimmed sequences then go through a low complexity filtering where sequences containing low complexity regions are identified and removed using the DUST [4] application.

For unassembled 454 reads, MAP performs a de-replication step to remove replicated sequences shown to be an artifact of the 454 pyrosequencing technique [5]. When two or more sequences are at least 95% identical, with their first 3 bps being identical as well, those two sequences are considered to be replicates and only the longer copy is retained. For unassembled Illumina reads, the same method is used to reduce the dataset size, except that the first 5 bps need to be identical for two sequences to be considered as replicates.

To further reduce the 454 dataset size we execute an additional clustering step. Since 454 data is known for erroneous calls whenever there are three or more nucleotides of the same type, unassembled 454 reads are clustered by first splitting each sequence in kmers, with each kmer reduced to a non-tandem sequence kmer, and then comparing if two sequences have identical kmers (DeClust, Mavrommatis K, unpublished).

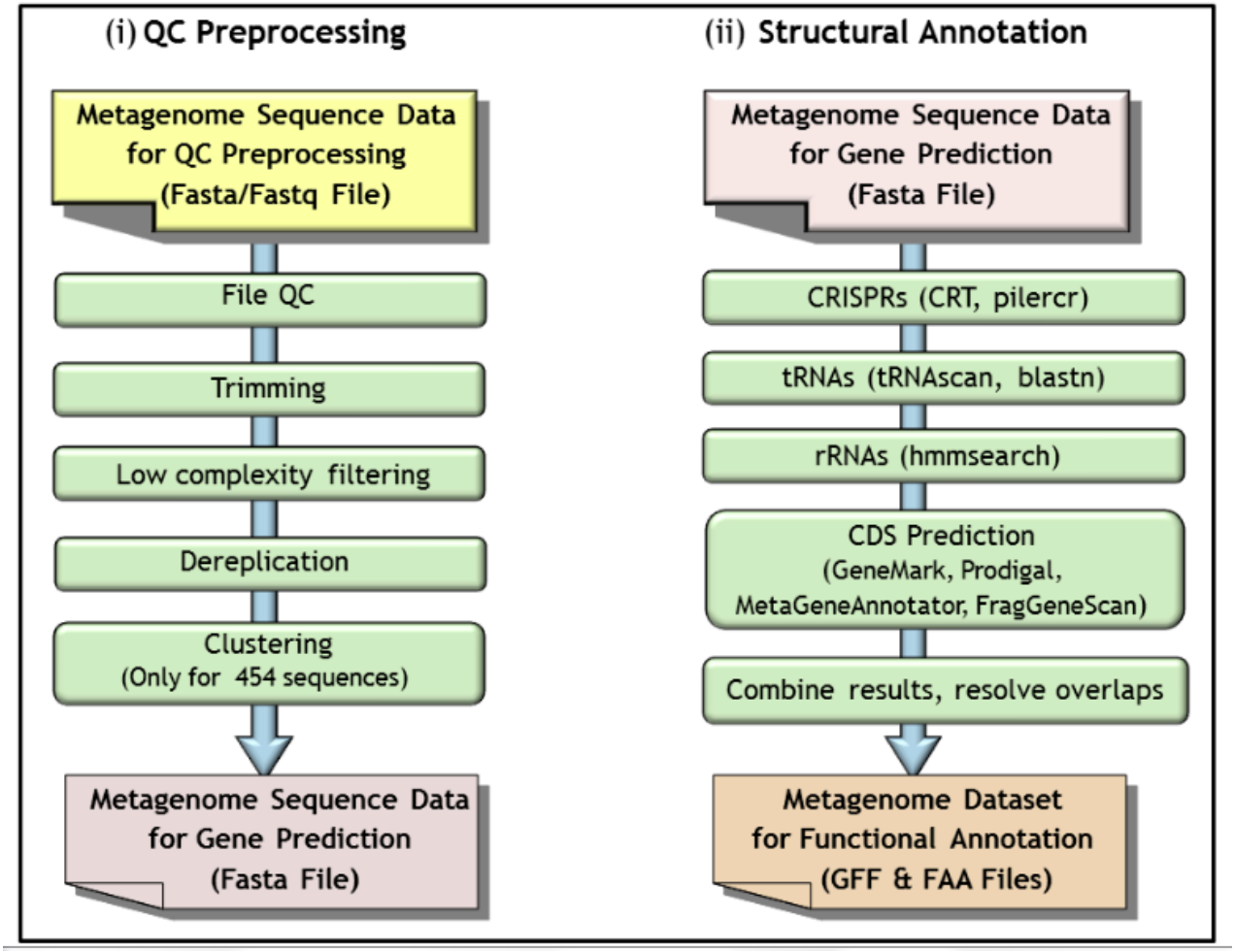

Figure 1. Metagenome sequence data preprocessing and structural annotation steps.

Feature Prediction

Scaffolds that have stretches of 50 Ns or more are separated into contigs in order to facilitate gene prediction. Scaffolding information is retained and contigs are assembled back into scaffolds after structural annotation. The first step in feature prediction is the identification of CRISPRs and non-coding RNA genes (tRNA, rRNA and other RNA genes), followed by prediction of protein coding genes, as shown in Figure 1(ii).

Identification of CRISPR elements is performed using the CRT [6] and PILER-CR v1.06 [7] tools. The PILER-CR settings include maximum spacer length of 100bp and the CRISPR element is required to have at least 5 repeats, which need to have at least 90% identity with each other and 75% identity with the consensus sequence. The MAP pipeline also runs a modified version of the CRT-CLI 1.2 version. The modified CRT has the capability to read multi-FASTA

files, detect truncated repeats at the ends of the contigs/scaffolds as well as the anchor repeat in the trail end and deal with spacer artifacts and repeats that contain Ns. This version also executes checks for repeat and spacer length ratios, while the length and similarity checks are performed as part of **"**all vs**.** all**"** spacer and repeat comparisons. Furthermore, the progression step of the sliding search window is reduced to 1, while threshold values and search ranges, which are strictly defined in the original software, can be changed from default values on the command line together with the new options and arguments. In the CRT version implemented in the MAP pipeline, the default values for the minimum and maximum repeat lengths are set to 20 and 50 bp, respectively, while the minimum and maximum spacer lengths are set to 20 and 60 bp, respectively. The ratio of the spacer lengths to the repeat lengths are required to be between 0.6 and 2.5. The default search window is 7 bp long and an element needs to have at least 3 repeats that have a minimum of 70% identity. The predictions from PILER-CR and CRT are concatenated and when overlapping, the CRT predictions are retained.

Protein-coding genes and non-coding RNA genes are identified using a combination of Hidden Markov Models and BLAST-based approaches. The first category of non-coding RNAs, tRNAs, are predicted using tRNAscan SE-1.3.1 [8] which requires the domain of the organism (*Bacteria*, *Archaea*, *Eukaryota*) as a parameter. A metagenome is a potential mixture of the three domains of life, so the program is run for each domain, that is three times, with the best scoring predictions selected. In order to identify fragmented tRNAs at the end of contigs/scaffolds (which tRNAscan fails to predict) sequences are compared to a BLAST database of nt sequences of tRNAs predicted in the isolate genomes. For sequences longer than 300bps, only the first 150bps and the last 150 bps are matched. Only hits with high sequence similarity (at least 85% identity and a minimum alignment length of 40) are kept.

Ribosomal RNA genes (5S, 16S, 23S) are predicted using hmmsearch tool from the package HMMER 3.0 [9]. Since the domain is a parameter that is required by hmmsearch, the pipeline runs it again three times against in-house curated models, derived from full length genes within IMG, while keeping the best scoring models.

The identification of protein-coding genes is performed using a consensus of four different *ab initio* gene prediction tools: prokaryotic GeneMark.hmm (v. 2.8) [10], MetaGeneAnnotator (v. Aug 2008) [11], Prodigal (v. 2.6.2) [12] and FragGeneScan (v. 1.16) [13]. The predictions from all the tools are combined and protein-coding genes with translations shorter than 32 amino acids are deleted. A majority rule-based decision schema is then followed in order to select gene calls. When there is a tie between two or more different gene models, selection is based on the preference order of gene callers determined by benchmarking of the individual gene finders on simulated metagenomic datasets (GeneMark > Prodigal > MetaGeneAnnotator > FragGeneScan). Overlaps between predicted features of different type (e. g. ncRNAs and protein-coding genes) get resolved based on an in-house curated set of rules.

Every annotated gene is assigned a locus tag of the form PREFIX_#####, where the prefix is the identifier of the GOLD Analysis Project associated with the metagenome dataset. The first "#"indicates the sequence type: 1=assembled, 2=unassembled 454 sequence, 3=unassembled Illumina sequence, 4=unassembled PacBio sequence. It is followed by one or more digits indicating the sequence number within the dataset and the number of the gene on this particular

sequence (which gets incremented by one for each following gene). Thereby each locus tag provides a unique identifier for every gene within a sequencing project.

The output of this stage consists of two files: a fasta formatted file containing all CDS protein sequences and a GFF formatted file placing predicted features on the metagenome sequences.

After the pre-processing and the structural annotation completed successfully, basic statics, e.g. number of sequences, sequence lengths distribution and number of genes predicted by each tool, can be viewed on the details page of every submission.

Functional Annotation

Functional annotation for metagenomes consists of associating protein-coding genes with COGs, Pfams, KO terms, EC numbers and phylogenetic lineage for scaffolds/contigs.

Genes are associated with COGs by comparing protein sequences to COG PSSMs from the CDD database [14], using RPS-BLAST 2.2.23 with an e-value of 10e-2 (0.1). If the overlap between two COG predictions is greater than half of the length of the shorter model, the hit having the largest bit score, lowest e-value, longer alignment length or higher percent identity, is retained.

Genes are associated with Pfam-A by comparing protein sequences to the Pfam database [15] using HMMER 3.0. Model specific trusted cut-offs are used with hmmsearch (--cut_tc), with output filtering following the same rules as mentioned above for COG assignments.

Genes are associated with KO terms [16] and EC numbers based on USEARCH results [17] comparing metagenome proteins against an isolate genome reference database with maxhits of 50 and an e-value of 0.1. An isolate genome reference database is assembled using all non-redundant protein sequences from public, high quality genomes in IMG. The top 5 hits to genes in the KO index are used, with an assignment made only if there is at least 30% identity and at least than 70% of the KO gene sequence is covered by the alignment. EC number assignments are derived from KO assignments using KEGG KO to EC mapping. One top USEARCH hit per gene is also retained for the Phylogenetic Distribution tool in IMG and assignment of phylogenetic lineage to scaffolds and contigs. The latter is assigned as the last common ancestor of USEARCH hits of the genes on the scaffold/contig provided that at least 30% of the genes have USEARCH hits.

Optional scaffold/contig coverage information, if provided by the user at the time of the submission, is used to calculate “estimated gene copies”, whereby the number of genes is multiplied by the average coverage of the contigs, on which these genes were predicted. This feature is important for accurate estimation of abundance of protein families, such as COGs, Pfams or KO terms, and phylogenetic lineages found in the metagenomes with assembled scaffolds/contigs, which collapse many unassembled reads into a single sequence. In the absence of coverage information most abundant and well-assembled phylogenetic lineages and protein families may appear underrepresented in the abundance analyses.

Product Name Assignment

Protein product names are assigned based on the name of their associated protein families, as follows:

1. If the gene has a COG assigned, the gene has at least 20% identity to the COG PSSM, and the alignment length is at least 70% of the COG consensus length, then the COG name is assigned as product name. If the COG name is “uncharacterized conserved protein” or contains “predicted”, the name has the format “COG name - COG ID”. If either the percent identity or alignment length condition is not satisfied, a check whether the COG and any Pfams assigned to the gene are found in a COG-Pfam correspondence table. This table has been generated by mapping COGs onto Pfams through the genes to which both are assigned. If at least one of the gene’s Pfams matches the respective COG in COG-Pfam correspondence table, the COG name is assigned as product name, even though the percent identity and alignment length for COG hit does not satisfy the above criteria.
2. For genes that were not associated with a product name using COG, product names are assigned based on the name of their associated Pfam, if a gene has at least one Pfam assigned to it.

Functional Annotation Sources

- COG 2014, November 2014
- KEGG Release 71.0, July 2014
- PFAM 27.0, March 2013
- IMG NR, March 2015

Summary

The MAP pipeline provides rapid automatic annotation of metagenome datasets. It is largely based on publicly available software supplemented with custom scripts for data handling and seamless integration of the input and output of different programs. The functional annotation is implemented within the Hadoop framework (https://hadoop.apache.org/). Consistency and reproducibility of the results produced by MAP depend on the databases and software used in the pipeline. New, updated versions of databases like Rfam, Pfam, and KEGG allow the prediction of more genes and more precise annotations. The pipeline is publicly available to the genomics community. We will continue to improve the MAP pipeline by extending the existing software and adding new tools that allow the identification and characterization of more features in the metagenome datasets. Additionally, we will be working on porting compute-intensive steps, e.g. the functional annotation, to the supercomputers located at the National Energy Research Scientific Computer Center (NERSC), such as Edison and Cori.

Acknowledgements

This work is funded by Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy (Contract No. DE-AC02-05CH11231).

References

1. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC. IMG/M 4 version of the integrated metagenome comparative analysis system.Nucleic Acids Res. 2015; 42(1):D568-73.
2. Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, et al. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2015;43:D1099-D1106.
3. Chou HH, Holmes MH. DNA sequence quality trimming and vector removal. Bioinformatics. 2001;17:1093-1104.
4. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. J Comput Biol. 2006;5:1028-40.
5. Gomez-Alvarez V, Teal TK, Schmidt TM. Systematic artifacts in metagenomes from complex microbial communities. ISME J. 2009;11:1314-7.
6. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.
7. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18.
8. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955-64.
9. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195
10. Lukashin A, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26:1107-15.
11. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. DNA Res. 2008;15:387-96.
12. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
13. Rho M, Tang H, Ye Y. FragGeneScan: Predicting Genes in Short and Error-prone Reads. Nucleic Acids Res. 2010;38:e191.
14. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, et al. CDD: a conserved domain database for inter-active domain family analysis. Nucleic Acids Res. 2007;35:D237-40.
15. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam Protein Families Database. Nucleic Acids Res. 2012;40:D290-301.

16. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42: D199–205.

17. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460-61.