The JGI Pipeline for Annotation of Microbial Genomes and Metagenomes

Marcel Huntemann[1*], Konstantinos Mavrommatis[1], Natalia Ivanova[1], Natalia Mikhailova[1], Galina Ovchinnikova[1], Andrew Schaumberg[1], Jim Tripp[1], Krishna Palaniappan[2], Ernest Szeto[2], I-Min Chen[2], Nikos Kyrpides[1], Amrita Pati[1],

[1] LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA USA

[2] Lawrence Berkeley National Laboratory Computational Research, Berkeley, CA USA

**To whom correspondence should be addressed*: Email: mhuntemann@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

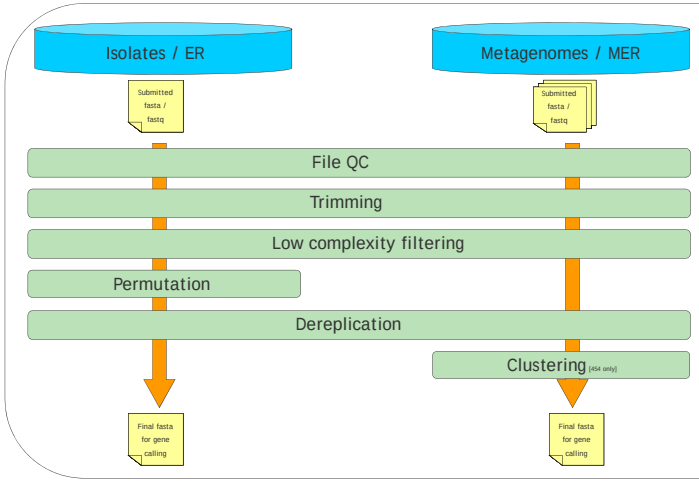
The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The JGI Pipeline for Annotation of Microbial Genomes and Metagenomes

Marcel Huntemann[1]*, Konstantinos Mavrommatis[1], Natalia Ivanova[1], Natalia Mikhailova[1], Galina Ovchinnikova[1], Andrew Schaumberg[1], Jim Tripp[1], Krishna Palaniappan[2], Ernest Szeto[2], I-Min Chen[2], Nikos Kyrpides[1], Amrita Pati[1]

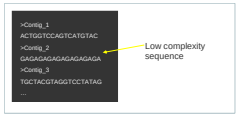
[1] DOE Joint Genome Institute, Walnut Creek, CA, USA
[2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

* Correspondence: Marcel Huntemann (mhuntemann@lbl.gov)

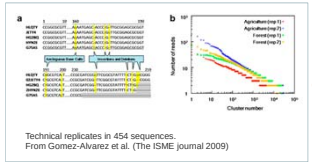


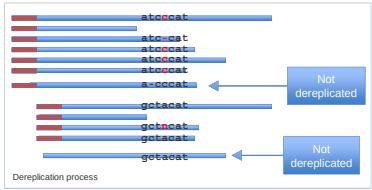
Dataset Pre-Processing

- **File QC**
  Validates correctness of input format and sequence alphabet, renames sequences (mapping file provided) and removes sequences that don't meet length thresholds.
- **Low complexity filtering**
  Removes sequences with too many low complexity parts.



- **Dereplication**
  Removes very similar sequences that have the same starting sequence.





- **Clustering**
  Keeps only the longest one of almost identical sequences (N matches any character and one gap allowed to account for polyN in 454 sequences).

- **Trimming**
  Trims off terminal Ns or if fastq file submitted uses appropriate method/cutoff depending on type (Sanger or Solexa).
- **Permutation**
  For finished genomes we try to predict the origin of replication and then permute the sequence so that it begins there.

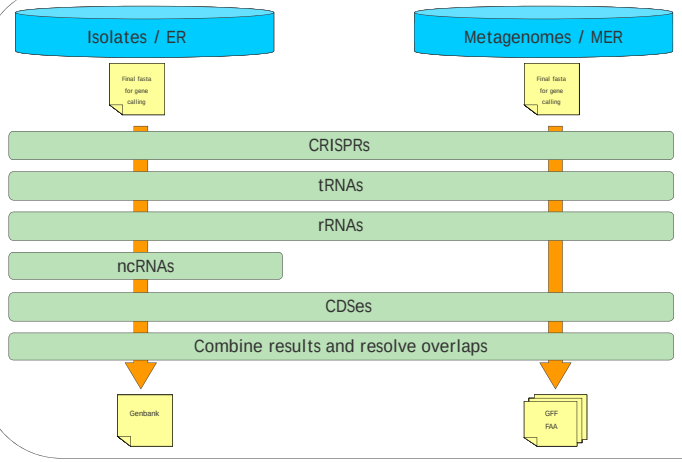
Computational Setup

Structural and functional annotation are embarrassingly parallel steps. JGI-PAMM v1.0, the JGI Pipeline for Annotation of Microbial genomes and Metagenomes, is set up on computational systems at NERSC (www.nersc.gov) and makes efficient use of the provided HPC infrastructure.

The current pipeline has been deployed in an integrated manner since the beginning of 2012 and has processed over **19.5 Billion genes** till date in an fully automated fashion.

The functional annotation of metagenomes is implemented within the Hadoop framework. It uses an on-the-fly approach on the Genepool Phase 2 system and runs out of GPFS instead of HDFS. All compute processes are implemented in the map phase to account for the dynamic set of TaskTracker nodes.

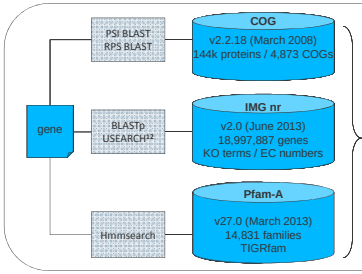
The pipeline is available to users submitting their datasets via IMG's submission system (http://img.jgi.doe.gov/submit).

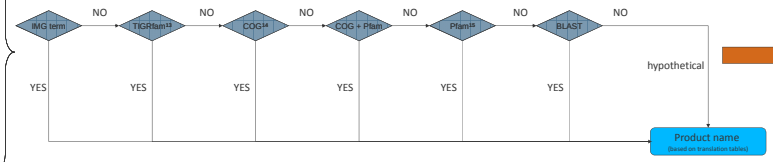


Feature Prediction

- **CRISPR prediction**
  Uses a modified **CRT CLI**[1] and **PileRCR**[2] to check for CRISPR elements.
- **tRNA prediction**
  Uses **tRNAscan**[3] and additionally **Blast**[4] on the first and last 150 bp of each sequence to account for partial tRNAs.
- **rRNA prediction**
  Uses **Hmmer**[5] and an in-house curated set of rRNA models.
- **ncRNA prediction**
  Uses **Blast** to determine which sequence regions have potential hits to **Rfam**[6] models, extracts those regions and runs **cmsearch**[7] on them. [genomes side only]
- **CDS prediction**
  Uses **Prodigal**[8] for genomes and a combination of **Prodigal**, **GeneMark**[9], **Metagene**[10] and **FragGeneScan**[11] for metagenomes.
- **Overlap resolution**
  Combines prediction results and resolves overlaps based on a curated set of rules for each feature type combination.

Functional Annotation





References

[1] Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
[2] Edgar, R.C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007).
[3] Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
[4] Altschul, S.F. et al. Basic local alignment search tool. *J Mol Biol* **215** (3): 403–410.
[5] Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 2009 Oct;23(1):205-11.
[6] Gardner, P.P. et al. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**, D136-140 (2009).
[7] Nawrocki, E.P. et al. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335-1337 (2009).
[8] Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
[9] Lukashin, A.V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-15 (1998).
[10] Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**, 5623-30 (2006).
[11] Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**, e191 (2010).
[12] Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2009).
[13] Haft, D.H., Selengut, J.D. & White, O. The TIGRFAMs database of protein families. *Nucl. Acids Res.* **31**, 371-373 (2003).
[14] Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
[15] Finn, R.D. et al. The Pfam protein families database. *Nucl. Acids Res.* **36**, D281-288 (2008).

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.