DOE Joint Genome Institute FY15 Q3 Progress Report

Q3: Report on new computational approaches for the annotation of genomic data.

Background

Recent technological advancements have enabled the large-scale sampling of genomes from uncultured microbial taxa, through the high-throughput sequencing of single amplified genomes (SAGs; Rinke et al., 2013; Swan et al., 2013) and assembly and binning of genomes from metagenomes (GMGs; Cuvelier et al., 2010; Sharon and Banfield, 2013). The importance of these products in assessing community structure and function has been established beyond doubt (Kalisky and Quake, 2011). Multiple Displacement Amplification (MDA) and sequencing of single cells has been immensely successful in capturing rare and novel phyla, generating valuable references for phylogenetic anchoring. However, efforts to conduct MDA and sequencing in a high-throughput manner have been heavily impaired by contamination from DNA introduced by the environmental sample, as well as introduced during the MDA or sequencing process (Woyke et al., 2011;Engel et al., 2014; Field et al., 2014). Similarly, metagenome binning and assembly often carries various errors and artifacts depending on the methods used (Nielsen et al., 2014). Even cultured isolate genomes have been shown to lack immunity to contamination with other species (Parks et al., 2014; Mukherjee et al., 2015). As sequencing of these genome product types rapidly increases, contaminant sequences are finding their way into public databases as reference sequences. It is therefore extremely important to define standardized and automated protocols for quality control and decontamination, which would go a long way towards establishing quality standards for all microbial genome product types.

Progress

The DOE JGI developed ProDeGe (Tennessen et al., 2015) the first fully automated computational protocol for decontamination of genomes. ProDeGe uses a combination of homology-based and sequence composition-based approaches to separate contaminant sequences from the target genome draft (Figure 1). It has been pre-calibrated to discard at least 84%of the contaminant sequence, which results in retention of a median 84% of the target sequence. The standalone software is freely available at http://prodege.jgi-psf.org//downloads/src and can be run on any system that has Perl, R, Prodigal and NCBI Blast installed. A graphical viewer allowing further exploration of data sets and exporting of contigs accompanies the web application for ProDeGe at http://prodege.jgi-psf.org, which is open to the wider scientific community as a decontamination service (Figure 2).

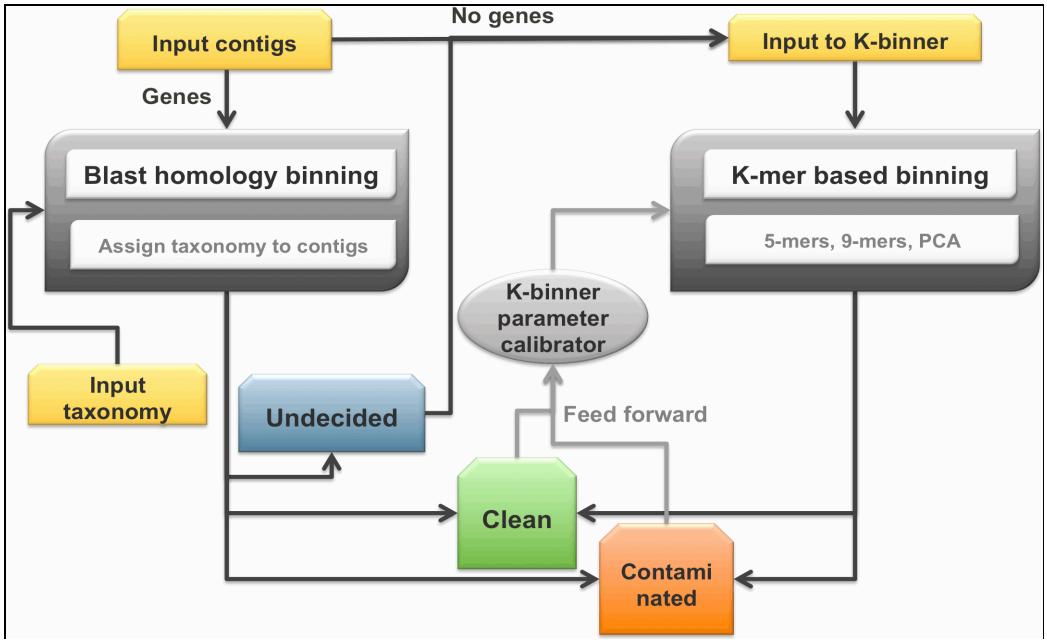

Figure 1: Schematic overview of the ProDeGe engine

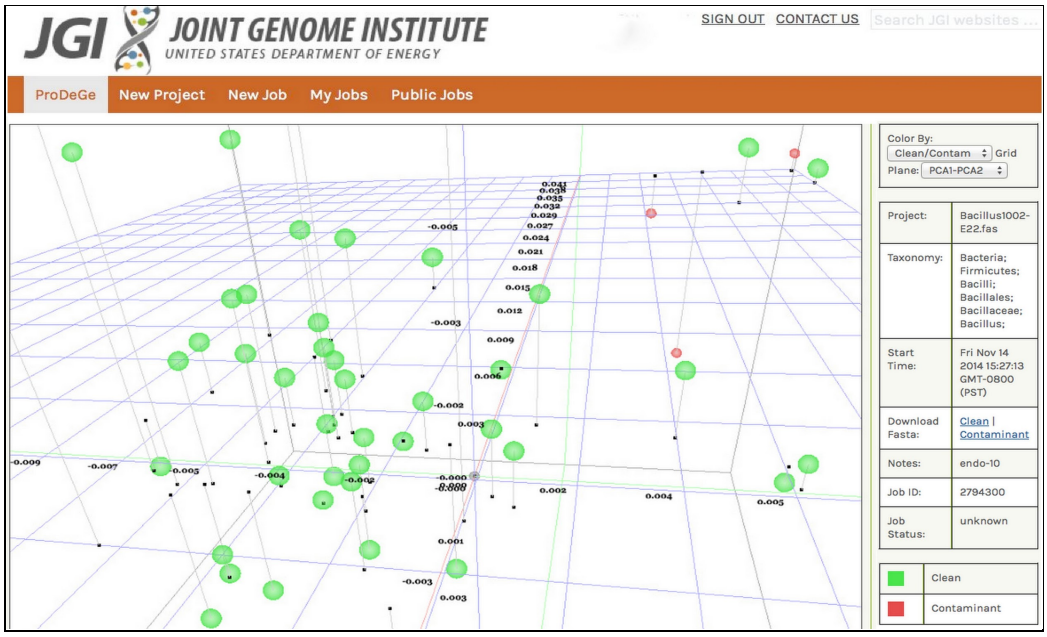

Figure 2: Visualization of the SAG dataset Bacillus sp. JGI 0001002-E22 (IMG Taxon OID 2528768030) shown using ProDeGe's web-based version at http://prodege.jgi-psf.org.

The performance of ProDeGe was evaluated using 182 manually screened SAGs (Figure 3) from two studies whose data sets are publicly available within the Integrated Microbial Genomes system (IMG; Markowitz, et al., 2014): genomes of 107 SAGs from an Arabidopsis endophyte sequencing project and 75 SAGs from the Microbial Dark Matter (MDM) project (Rinke et al., 2013). Manual curation of these SAGs demonstrated that the use of ProDeGe prevented 5311

potentially contaminated contigs in these data sets from entering public databases. Figure 4a demonstrates the sensitivity vs specificity plot of ProDeGe results for the above data sets. Most of the data points in Figure 4a cluster in the top right of the box reflecting a median retention of 89% of the clean sequence (sensitivity) and a median rejection of 100% of the sequence of contaminant origin (specificity). In addition, on average, 84% of the bases of a data set are accurately classified. ProDeGe performs best when the target organism has sequenced homologs at the class level or deeper in its high-quality prokaryotic nucleotide reference database. If the target organism's taxonomy is unknown or not deeper than domain level, or there are few contigs with taxonomic assignments, a target bin cannot be assessed and thus ProDeGe removes contaminant contigs using sequence composition only. The few samples in Figure 4a that demonstrate a higher rate of false positives (lower specificity) and/or reduced sensitivity typically occur when the data set contains few contaminant contigs or ProDeGe incorrectly assumes that the largest bin is the target bin. Some data sets contain a higher proportion of contamination than target sequence and ProDeGe's performance can suffer under this condition. However, under all other conditions, ProDeGe demonstrates high speed, specificity and sensitivity (Figure 4).

	Endophyte SAGs	MDM SAGs	Public nonJGI SAGs	GMGs
Number of genomes	107	75	188	13
Median number of contigs	89	60	75	274
Median contig length (nts)	6,215	4,502	3,119	3,102
Median GC %	59.3	38.2	39.4	56.5
Median contam % (manual curation)	7.15	8.27	NA	NA
Median contam % (ProDeGe)	28.7	22.8	30.5	27.1
Validation?	Yes	Yes	No	No

Figure 3: Features of data sets used to validate ProDeGe: SAGs from the Arabidopsis endophyte sequencing project, MDM project, public data sets found in IMG but not sequenced at the JGI, as well as genomes from metagenomes.

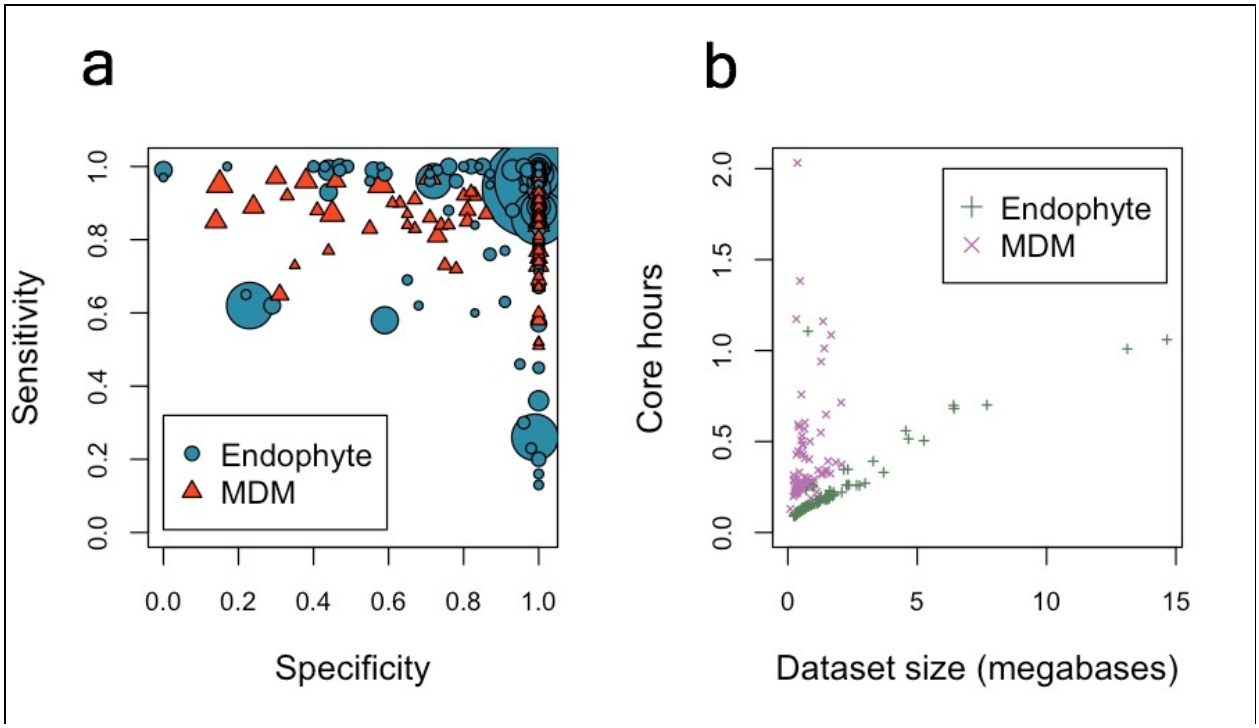

Figure 4: ProDeGe accuracy and performance scatterplots of 182 manually curated single amplified genomes (SAGs), where each symbol represents one SAG data set. (a) Accuracy shown by sensitivity (proportion of bases confirmed 'Clean') vs specificity (proportion of bases confirmed 'Contaminant') from the Endophyte and Microbial Dark Matter (MDM) data sets. Symbol size reflects input data set size in megabases. Most points cluster in the top right of the plot, showing ProDeGe's high accuracy. Median and average overall results are shown in Supplementary Table S1. (b) ProDeGe completion time in central processing unit (CPU) core hours for the 182 SAGs. ProDeGe operates successfully at an average rate of 0.30 CPU core hours per megabase of sequence. Principal components analysis (PCA) of a 9-mer frequency matrix costs more computationally than PCA of a 5-mer frequency matrix used with blast-binning. The lack of known taxonomy for the MDM data sets prevents blast-binning, thus showing longer finishing times than the endophyte data sets, which have known taxonomy for use in blast-binning.

The web application for ProDeGe allows users to export clean and contaminant contigs, examine contig gene calls with their corresponding taxonomies, and discover contig clusters in the first three components of their k-dimensional space.

Between April 25th and June 17th, The ProDeGe web-service hosted 60 users, and the stand-alone tool was downloaded 88 times. The ProDeGe manuscript hit number 8 on ISME's "Top Ten most downloaded articles in the last 15 days".

Conclusion

ProDeGe is the first step towards establishing a standard for quality control of genomes from both cultured and uncultured microorganisms. It is valuable for preventing the dissemination of contaminated sequence data into public databases, avoiding resulting misleading analyses. The

fully automated nature of the pipeline relieves scientists of hours of manual screening, producing reliably clean data sets and enabling the high-throughput screening of data sets for the first time. ProDeGe, therefore, represents a critical component in our toolkit during an era of next-generation DNA sequencing and cultivation-independent microbial genomics.

Copyright and License

DOE JGI team

Kristin Tennessen, Evan Andersen, Scott Clingenpeel, Natalia Ivanova, Tanja Woyke, Nikos Kyrpides

References

Cuvelier ML, Allen AE, McCrow JP, Messié M, Tringe SG, Woyke T *et al*. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci USA* **107**: 14679–14684.

Engel P, Stepanauskas R, Moran NA. (2014). Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* **10**: e1004596.

Field EK, Sczyrba A, Lyman AE, Harris CC, Woyke T, Stepanauskas R *et al*. (2014). Genomic insights into the uncultivated marine Zetaproteobacteria at Loihi Seamount. *ISME J* **9**: 857–870.

Kalisky T, Quake SR. (2011). Single-cell genomics. *Nat Methods* **8**: 311–314.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M *et al*. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* **42**: D560–D567.

Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* **10**: 18.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S *et al*. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**: 822–828.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2014). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *PeerJ PrePrints* **2**: e554v1.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al*. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Sharon I, Banfield JF. (2013). Genomes from metagenomics. *Science***6162**: 1057–1058.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM*et al*. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Tennessen, K., Andersen, E., Clingenpeel, S., Rinke, C., Lundberg, D.S., Han, J., Dangl, J.L., Ivanova, N., Woyke, T., Kyrpides, N., Pati, A. (2015). ProDeGe: a computational protocol for fully automated decontamination of genomes. The ISME Journal advance online publication 9 June 2015 doi: 10.1038/ismej.2015.100.

Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S *et al*. (2011). Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One* **6**: e26161.