Microbial Genomics, Comparative and Functional Analyses at JGI-LLNL

Stephanie Malfatti, Lisa Vergez, Maria Shin, Mari Christensen, Jeff Elliott, Dorothy Lang and Patrick Chain
Lawrence Livermore National Laboratory, Joint Genome Institute, CA

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

Since the initiation of the Department of Energy's Joint Genome Institute, and as part of the DOE's Microbial Genome Program, the LLNL microbial genomics group has, as part of the JGI, been involved in various aspects of delivering finished genomes and performing detailed analyses, including comparative genomics, for publication purposes. Though we have only published 18 of our ~50 finished bacterial genomes, we are currently in the final phases of analysis for an additional 10 and are in various stages of working on annotation and comparative analyses for another 16 microbial genomes. Here, we outline our data management and finishing processes, and also present a few of the recently completed microbial genomes, including *Pseudomonas putida* F1, *Psychrobacter* sp. PRwf-1 and *Sinorhizobium medicae* WSM419.

JGI-LLNL Miicrobial Genome Projects

Taxa	Organism	GC %	Size	Status
Betaproteobacteria	*Burkholderia ambifaria AMMD*	67%	7.53	Man. In Prep
Betaproteobacteria	*Burkholderia ambifaria MC40-6*	66%	7.7	Active
Betaproteobacteria	*Burkholderia cenocepacia AU1054 Rg2/BSL2*	64%	7.28	Man. In Prep
Betaproteobacteria	*Burkholderia cenocepacia hi2424*	66%	7.76	Man. In Prep
Betaproteobacteria	*Burkholderia cenocepacia MC0-3*	66%	7.9	Active
Betaproteobacteria	*Burkholderia vietnamiensis G4*	66%	8.4	Man. In Prep
Betaproteobacteria	*Burkholderia sp. 383*	61%	8.8	Man. In Prep
Betaproteobacteria	*Burkholderia multivorans ATT17616*	66%	7.0	Active
Betaproteobacteria	*Burkholderia phymatum STM 815*	62%	8.6	Active
Betaproteobacteria	*Burkholderia phytofirmans PsJN*	62%	8.1	Active
Betaproteobacteria	*Burkholderia xenovorans LB400*	62%	9.77	Published
Firmicutes	*Clostridium sp. OhiLAs*	36%	3.0	Active
Deltaproteobacteria	*Desulfovibrio vulgaris DePue*	63%	3.6	Finished
Alphaproteobacteria	*Ehrlichia canis*	29%	1.32	Published
Alphaproteobacteria	*Ehrlichia chaffeensis*	29%	1.8	Finished
Euryarchaeota	*Methanosaeta thermophila PT*	53%	1.9	Finished
Betaproteobacteria	*Methylibium petroleiphilum PM1*	69%	4.6	Published
Actinobacteria	*Mycobacterium gilvum PYR-GCK*	68%	5.9	Finished
Actinobacteria	*Mycobacterium sp.JLS*	68%	6.0	Finished
Alphaproteobacteria	*Nitrobacter hamburgensis*	62%	5.01	Man. In Prep
Alphaproteobacteria	*Nitrobacter winogradskyi Nb-255 (ATCC 25391)*	62%	3.4	Published
Betaproteobacteria	*Nitrosomonas europaea*	51%	2.8	Published
Betaproteobacteria	*Nitrosomonas eutropha C91*	49%	2.82	Man. In Prep
Gammaproteobacteria	*Nitrosomonas oceani C-107*	50%	3.5	Published
Betaproteobacteria	*Nitrosospira multiformis Surinam*	54%	3.23	Man. In Prep
Cyanobacteria	*Nostoc punctiforme*	41%	9.2	Published
Cyanobacteria	*Prochlorococcus marinus MED4*	31%	1.66	Published
Cyanobacteria	*Prochlorococcus marinus MIT9313*	51%	2.4	Published
Gammaproteobacteria	*Pseudomonas putida F1*	59%	5.9	Finished
Gammaproteobacteria	*Pseudomonas putida W619*	61%	5.7	Active
Gammaproteobacteria	*Psychrobacter sp. PRwf-1*	44%	2.9	Finished
Crenarchaeota	*Pyrobaculum arsenatuicum DSM 13514*	55%	2.1	Finished
Crenarchaeota	*Pyrobaculum calidifontis JCM 11548*	57%	2.0	Finished
Alphaproteobacteria	*Rhodopseudomonas palustris CGA009*	65%	5.47	Published
Alphaproteobacteria	*Rhodopseudomonas palustris BisA53*	64%	5.5	Man. In Prep
Alphaproteobacteria	*Rhodopseudomonas palustris BisB18*	65%	5.51	Man. In Prep
Alphaproteobacteria	*Rhodopseudomonas palustris BisB5*	65%	4.89	Man. In Prep
Alphaproteobacteria	*Rhodopseudomonas palustris HaA2*	66%	5.33	Man. In Prep
Gammaproteobacteria	*Shewanella baltica OS195*	46%	5.3	Active
Gammaproteobacteria	*Shewanella putrefaciens 200*	45%	4.7	Active
Gammaproteobacteria	*Shewanella putrefaciens CN-32*	44%	4.5	Finished
Gammaproteobacteria	*Shewanella sp. PV-4*	54%	4.6	Finished
Gammaproteobacteria	*Shewanella sp.* W3-18-1	45%	4.7	Finished
Alphaproteobacteria	*Sinorhizobium medicae*	61%	6.7	Finished
Firmicutes	*Staphylococcus aureus JH1*	34%	2.9	Finished
Firmicutes	*Staphylococcus aureus JH9*	34%	2.9	Finished
Cyanobacteria	*Synechococcus WH8102*	59%	2.4	Published
Betaproteobacteria	*Thiobacillus denitrificans, ATCC 23644*	66%	2.91	Published
Gammaproteobacteria	*Thiomicrospira crunogena*	43%	2.4	Published
Epsilonproteobacteria	*Thiomicrospira denitrificans*	35%	2.2	Man. In Prep
Verrucomicrobia	*Verrucomicrobium TAV2*	59%	5.7	Active
Verrucomicrobia	*Victivallis vadensis*	59%	4.6	Active

Acknowledgements

We wish to express thanks to our countless collaborators who have allowed us to participate in these studies, and who have contributed a great deal to our understanding of the organisms we sequence. We also wish to thank summer students, sponsored scholars and fellows, visiting graduate students, post-docs and scientists, as well as former members of our group, for their contributions to our past projects; Hector Ayala-del-Rio, Andrew Dailey, Vincent Denef, Jane Lamerdin, Par Larsson, William Marrs, Warren Regala, Chad Smith, Shawn Starkenburg, Stacia Thompson, and Cale Whitworth.
In addition, we thank our JGI colleagues at LANL, ORNL, PGF, Stanford for their input and help in improving our process.

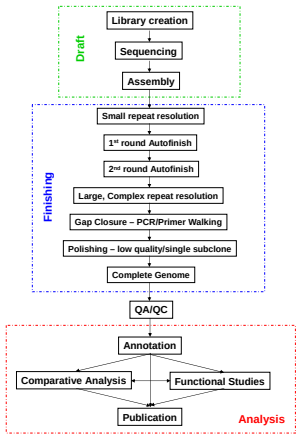
JG-LLNL Finishing Process

The JGI-LLNL finishing process starts with an automated round of repeat resolution using in-house designed software to resolve mis-assemblies caused by short repetitive elements (<3.5kb in length). This is followed by two rounds of automated primer design using Consed's Autofinish program which designs experiments for gap closure and ambiguity resolution. Next, we manually tackle large, complex repetitive elements (>5 kb in length), and are also employing PCR and primer walking to close the remaining captured and/or un-captured gaps. Additionally, we go through a "polishing" phase to resolve any low quality and/or single subclone regions to ensure the final error rate for each replicon is <1 error per 50 kb with a minimum of 2X coverage across the genome. The integration of pyrosequencing data into our pipeline has been another area of active research.

Once a genome is "completed", the genome is sent for final QA/QC by the independent Stanford group. After passing QA/QC, the complete set of replicons is sent for final Annotation at ORNL and then to the PGF for final incorporation into the IMG Database.

We coordinate directly with our genome collaborators in all cases, and have played an integral role in the analysis, annotation and comparative analysis of completed genomes and strive for a descriptive publication and sometimes, further functional studies.

Microbial Genome Strategy



Selected Genome Analysis Publications

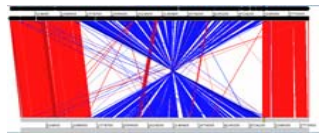
1: Kane SR, et al. Whole-Genome Analysis of the Methyl tert-Butyl Ether-Degrading Beta-Proteobacterium *Methylibium petroleiphilum* PM1. J Bacteriol. 2007 Mar;189(5):1931-45. Epub 2006 Dec 8.
2: Scott KM, et al. The genome of deep-sea vent chemolithoautotroph *Thiomicrospira crunogena* XCL-2. PLoS Biol. 2006 Nov;4(12):e383.
3: Chain PS, et al. *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A. 2006 Oct 17;103(42):15280-7. Epub 2006 Oct 9.
4: Brotcke A, et al. Identification of MglA-regulated genes reveals novel virulence factors in *Francisella tularensis*. Infect Immun. 2006 Dec;74(12):6642-55. Epub 2006 Sep 25.
5: Klotz MG, et al.. Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium *Nitrosococcus oceani* ATCC 19707. Appl Environ Microbiol. 2006 Sep;72(9):6299-315.
6: Loots GG, et al. Array2BIO: from microarray expression data to functional annotation of co-regulated genes. BMC Bioinformatics. 2006 Jun 16;7:307.
7: Chain PS, et al. Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol. 2006 Jun;188(12):4453-63.
8: Starkenburg SR, et al. Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. Appl Environ Microbiol. 2006 Mar;72(3):2050-63.
9: Beller HR, et al. The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium *Thiobacillus denitrificans*. J Bacteriol. 2006 Feb;188(4):1473-88.
10: Chain PS, et al. Whole-genome analyses of speciation events in pathogenic Brucellae. Infect Immun. 2005 Dec;73(12):8353-61.
11: Feil H, et al. Comparison of the complete genome sequences of *Pseudomonas syringae pv.* syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A. 2005 Aug 2;102(31):11064-9. Epub 2005 Jul 25.
12: Larsson P, et al.. The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. Nat Genet. 2005 Feb;37(2):153-9. Epub 2005 Jan 9.
13: Larimer FW, et al. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nat Biotechnol. 2004 Jan;22(1):55-61. Epub 2003 Dec 14.
14: Garcia E, et al. The genome sequence of *Yersinia pestis* bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J Bacteriol. 2003 Sep;185(17):5248-62.
15: Rocap G, et al. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003 Aug 28;424(6952):1042-7. Epub 2003 Aug 13.
16: Chain P, et al An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. Brief Bioinform. 2003 Jun;4(2):105-23. Review.
17: Chain P, et al. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. J Bacteriol. 2003 May;185(9):2759-73. Erratum in: J Bacteriol. 2003 Nov;185(21):6496.
18: Mavromatis, C., et al. The Genome of the Obligately Intracellular Bacterium *Ehrlichia canis* Reveals Themes of Complex Membrane Structure and Immune Evasion Strategies J. Bacteriol. 2006 188: 4015-4023.
19: Meeks JC, et al. An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. Photosynth Res. 2001;70(1):85-106.
20: Han CS, Chain PS. Finishing Repetitive Regions Automatically with Dupfinisher. Conference on Bioinformatics & Cmp. Biology, BIOCOMP 2006.

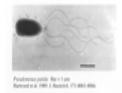
Comparative Genome Analysis

Pseudomonas putida

P. putida sp. KT2440

P. putida sp. F1





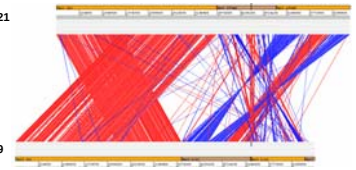
	P.putida KT2440	P.putida F1
	Chromosome	Chromosome
Size	6.18 Mb	5.96 Mb
G+C%	61.5%	62%
Genes	5350	5251

Pseudomonas putida is a metabolically versatile saprophytic soil bacterium that is known for its diverse metabolism and potential for development of biopesticides and plant growth promoters because of its ability to colonize rhizosphere of crop plants. *P. putida* F1 is one of the most well-studied aromatic hydrocarbon degrading bacterial strains. We have performed preliminary comparisons between *P. putida* strains F1 and KT2440, which is a strain incapable of aromatic hydrocarbon degradation. The major inversion (blue) is flanked by two ISPpu9 insertion sequences within the KT2440 strain. The KT2440 genome has 3 large phage insertions and 1 bacteriocin, which looks to be unique to this strain as they are not found in the F1 strain and may explain its smaller size (by about 220 Kb). Of interest will be to look for gene regions that are unique to the F1 strain, which may contribute to its exceptional biodegrading versatility.

Sinorhizobium medicae

S.Meliloti 1021

S.Medicae WSM419



S. meliloti 1021

	Size	G+C %
Chromosome	3.65 Mb	62.7%
pSymA	1.35 Mb	60.4%
pSymB	1.68 Mb	62.4%

S. medicae WSM419

	Size	G+C %
Chromosome	3.78 Mb	62%
Plasmid1	1.57 Mb	61%
Plasmid 2	1.23 Mb	60%
Plasmid 3	0.219 Mb	60%

We compared the genomes of two gram negative Alphaproteobacterial nitrogen-fixing facultative legume symbionts *S. medicae* strain WSM419 and the published *S. meliloti* 1021. *Sinorhizobium medicae* can be distinguished from *S. meliloti* by its unique capacity to fix nitrogen in association with annual and perennial *Medicago* hosts of world-wide agronomic value. *S. medicae* is also superior in its nitrogen fixation and acid tolerance characteristics. Preliminary whole genome alignments reveal the colinear nature of the chromosome and a largely colinear (one large rearrangement) pSymB (pExo) megaplasmid, while the main replicon responsible for nitrogen fixation and nodulation, pSymA, has undergone numerous changes since the two species diverged. It will be interesting to see if the various phenotypic differences can be explained solely on the basis of pSymA divergence or if smaller differences in the chromosome or other megaplasmid are responsible.

Psychrobacter sp. PRwf-1

P. arcticus 273-4

	Size	G+C%
Chromosome	2.65 Mb	43%

Psychrobacter sp. PRwf-1

	Size	G+C%
Chromosome	2.98 Mb	45.0%
Plasmid1	13.9 kb	38.0%
Plasmid 2	2.1 Kb	40.0%

P. cryohalolentis K5

	Size	G+C%
Chromosome	3.06 Mb	42.3%
Plasmid1	41.2 Kb	38.3%

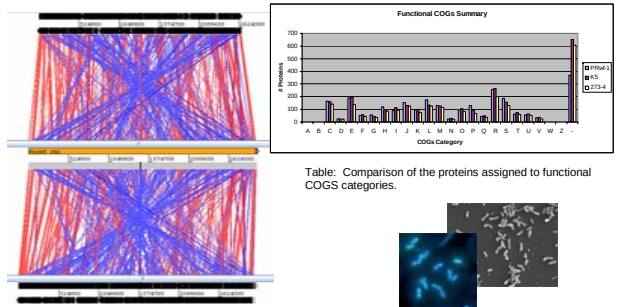


Table: Comparison of the proteins assigned to functional COGS categories.

Psychrobacter species are capable of reproducing at temperatures ranging from -10 to 40 degrees Celsius and have been isolated primarily from low temperature marine environments including Antarctic sea ice, ornithogenic soil, and sediments, the stomach contents of the Antarctic krill Euphausia, sea water (NW pacific ocean, 300 m depth), the deep sea, and the internal tissues of a marine ascidian. Yet, *Psychrobacter* species have also been isolated from a pigeon feces bioaerosol, a poultry carcass, fermented sea food, human blood and tissues, and the lungs of an infected lamb. The sequenced psychrophilic species, *P. arcticus* 273-4 and *P. cryohalolentis* K5 were isolated from 20-40 thousand-year-old Siberian permafrost core/cryopeg, where the in situ temperature is -9 to -11 degrees Celsius, while *Psychrobacter sp.* PRwf-1 was isolated from Puerto Rican waters where the temperature is significantly higher. These comparisons show a great deal of genome shuffling but little horizontal acquisition, thus it is likely that specific adaptations throughout the genome permit growth at different temperature ranges.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.