Title of Project: Methanogenic archaea and the global carbon cycle: a systems biology approach to the study of *Methanosarcina* species

Institution: University of Illinois

Date Received: 6-14-2010

Principal Investigator: William W. Metcalf

Co-PIs: Nathan Price, Rachel Whitaker, Ping Ma

Project Results:

The long-term goal of this multi-investigator project was the creation of integrated, multiscale models that accurately and quantitatively predict the role of *Methanosarcina* species in the global carbon cycle under dynamic environmental conditions. To achieve these goals we pursed four specific aims: (1) genome sequencing of numerous members of the Order *Methanosarcinales*, (2) identification of genomic sources of phenotypic variation through *in silico* comparative genomics, (3) elucidation of the transcriptional networks of two *Methanosarcina* species, and (4) development of comprehensive metabolic network models for characterized strains to address the question of how metabolic models scale with genetic distance.

Genome sequencing of members of the order Methanosarcinales. During the course of this project, we sequenced and annotated the genomes of 154 members of the order *Methanosarcinales*. These genomic data provide a degree of coverage currently unrivaled amongst Archaea, creating an invaluable resource that enables detailed studies of the evolution, adaptation, metabolism, biochemistry and physiology of these versatile methanogens. Initially, we fully sequenced the genomes of twenty-three members of the genus *Methanosarcina,* including *M. sp.* MTP4, *M. thermophila* TM1, *M. thermophila* CHTI55, *M. thermophila* MSTA1, *M. vacuolata* Z761, *M. sp.* Kolskee, *M. barkeri* Weismoor, *M. barkeri* MS, *M. barkeri* 227, *M. barkeri* 3, *M. siciliae* C2J, *M. siciliae* HI350, *M. siciliae* T4M, *M. sp.* WH1, *M. sp.* WWM586, *M. lacustris* Z7289*, M. horonobensis* HB1, *M. mazei* SarPi, *M. mazei* LYC, *M. mazei* S6, *M. mazei* TMA, *M. mazei* WWM610, and *M. mazei* C16. We also completed high coverage draft sequences of four additional *Methanosarcina* species: *M. baltica* GS1, *M. sp.* Naples 100, *M. lacustris* ZS and *M. calensis* Cali. All of these strains are publically available from the DSMZ strain collection (http://www.dsmz.de). All genome sequences have been deposited in GenBank (see NCBI BioProjects 230935-230962).

To provide a measure of the genome variability in natural populations, we also isolated and determined draft genome sequences from a large number of methanogenic archaea from the Columbia River Estuary (Oregon, USA). In all, we isolated nearly 200 strains using a variety of selective growth substrates from three sub-sites of varying salinity, pH and nutrient levels. Importantly, the chemical and metagenomic context of the samples used for our isolation have been characterized by our collaborator, Dr. Holly Simon from the Oregon Health and Sciences University. High coverage, draft genome sequences were then determined for 127 isolates. Phylogenetic analyses of these genomes show that the sequenced isolates encompass the full diversity of the genus *Methanosarcina* (102 strains), as well more distantly relates members of the order *Methanosarcinales* (25 strains). Roughly half of these sequences have been deposited in GenBank (accession numbers JJOR00000000-JJRB00000000), the other are currently being processed).

Identification of genomic sources of phenotypic variation. To further our understanding of the evolution and adaptation of *Methanosarcina* species we analyzed subsets of our genome sequence dataset to reveal the footprints of adaptive change in response to natural selection, which can be seen within the patterns of genome variation. This high-resolution comparative analysis of close relative is the key to deciphering these patterns, because the mutational events that create the observed genomic changes occurred on recent time scales and are not obscured by the accumulation of many evolutionary events.

To aid in this analysis, we developed the Integrated Toolkit for the Exploration of Pan-genomes (ITEP), a software package that allows facile curation, analysis, and visualization of gene families. ITEP was designed to allow users great flexibility in analysis of protein families. The software includes tools for finding and annotating uncalled genes, integrating phylogenetic and synteny information, computing and visualizing protein families, and storing the results in a user database. Among its' many applications, ITEP greatly simplifies analysis of gene conservation, loss and gain. For example, analysis of clade specific gene conservation suggests a compelling story for the evolution of the *Methanosarcinales* in which sequential acquisition of genes needed for electron transport, C-1 metabolism and aceticlastic methanogenesis led to a new form of “respiratory” methanogenesis that employs a membrane-bound electron transport chain and growth substrates not observed in other methanogenic groups. Similar analyses have been performed to categorize the gain, loss and transfer of genes involved in hydrogen and nitrogen fixation, revealing discrete patterns of traits directly related to ecosystem function and adaptation.

In general, these gene presence/absence profiles are consistent with our previously published results showing two discrete types of energy conserving electron transport in *Methanosarcina*: one typified by the hydrogen-dependent electron transport chain of *M. barkeri* and the other by the hydrogen-independent electron transport chain of *M. acetivorans*. However, we were surprised to observe a third pattern, typified by *M. lacustris* species, which apparently represents a hybrid of the two types. Two important caveats arise from these comparative genomic studies. First, hydrogen metabolism is clearly a defining trait among the *Methanosarcinales*. This has significant implications for competition with other H_2 consuming organisms because the thermodynamics of hydrogen oxidation are superior for organisms that can use electron acceptors such as sulfate or Fe(III). Second, conclusions based on the presence/absence of hydrogenase genes must made with caution. In particular, we have previously shown that *M. acetivorans* possess hydrogenase genes, but that they are not expressed under any known growth conditions.

We used a similar approach to establish how genomic and ecological diversity is partitioned within and between 56 *Methanosarcina mazei* isolates from our Columbia River Estuary collection. Whole-genome analysis revealed two distinct, apparently co-existing clades, which we refer to as the ‘mazei-T’ and ‘mazei-WC’ clades. Genomic analyses showed that these clades differ in gene content and fixation of allelic variants, which point to potential differences in primary metabolism and also interactions with foreign genetic elements. Laboratory growth experiments revealed significant differences in trimethylamine utilization, supporting ecological differentiation of the two clades. The results suggest that environmental pressures in the Columbia River Estuary are selecting for a diverse repertoire of metabolic types within the *Methanosarcina*.

Elucidation of the transcriptional networks of two Methanosarcina species. A major goal of this project was to create a comprehensive dataset that describes the transcriptional regulation of *M. barkeri* and *M. acetivorans.* Integration of regulatory constraints into our metabolic models is expected to dramatically improve their performance. To achieve this goal, we completed an exhaustive characterization of transcriptional landscape of both *M. acetivorans* and *M. barkeri* using RNA-seq analysis. Transcript levels were determined after growth on all known substrates and several combinations of multiple substrates. Data were also collected at multiple points during batch culture to identify expression levels during various growth phases (lag, early-, mid-, late-exponential and stationary). These data provide a window into the global gene regulation and reveal discrete gene regulons involved in the use of methanol, methyl-amines and acetate. Interestingly, the data also suggest the presence of a novel respiratory pathway for oxidation of acetate using methanol as an electron acceptor. We also conducted RNA-seq experiments to identify the 5'-ends and to establish the in vivo half-lives of all mRNA transcripts in *M. acetivorans*. These data are currently being prepped for submission to NCBI's *Gene Expression Omnibus* (*GEO*).

During the early stages of this work, we developed a tool set to aid in the analysis of RNA-seq datasets. Accurate quantification of gene expression via RNA-seq relies on these read counts. During the first eighteen months of the project we developed effective statistical methods to accurately quantify gene expression based on RNA-Seq read counts. Several proposed methods were tested using public available datasets and their accuracies are compared. We also developed a method for bias correction of RNA-Seq by a dinucleotide expansion model with penalized regression for model-fitting. The resulting estimate outperforms the current best available estimates. For multiple gene expression data, an integrated nonparametric approach was developed for modeling the complex nonlinear interactions and a novel model free variable selection approach was developed for overcoming the bias induced for mis-specifying models.

Development of comprehensive metabolic network models and examination of how metabolic models scale with genetic distance. Genome-scale metabolic modeling is a powerful way to consolidate large amounts of biological information in a way that enables phenotype predictions. Accurate genome-scale models depend on accurate underlying metabolic networks, which are built by carefully curating any genomic and biochemical data that is available for the modeled species or related species.

One of the most significant achievements of this project is the construction of the first genome-scale metabolic network model for *Methanosarcina acetivorans*. To achieve this, we integrated the knowledge available in a wide variety of literature sources (over 150) to produce the most accurate model possible. The reconstruction predicts knockout lethality with an accuracy of 96% and also accurately predicts growth and product secretion rates. The model was used to make predictions about the mechanism for regeneration of the cofactor F_{420} during growth on carbon monoxide, hypothetical mechanisms for generation of byproducts previously un-observed in methanogens (such as formate and methylsulfides), and the implications of a hypothetical electron bifurcation in the soluble heterodisulfide reductase (HdrABC) on the growth phenotype. In addition to the new reconstruction of *M. acetivorans*, we also completed a substantial update to the previously published *M. barkeri* reconstruction. The updated model is significantly more accurate than the previous one, incorporates new gene annotations and metabolic functions, and provides additional insights into *M. barkeri* biology.

Given the great effort required to build high-quality metabolic networks, there is a great need to assess the extent to which they can be applied to related organisms. Using existing genome-scale metabolic models of *M. acetivorans* C2A and *M. barkeri* Fusaro as references, we built models for each newly-sequenced *Methanosarcina* and found that the propagated models were missing critical functionality that was present in the reference models. Closer examination revealed clear problems in the published models and in the genomic data that were not apparent without a comparative approach. Many of these problems could be fixed using genome curation tools or through targeted literature searches, while others suggest promising candidates for further pathway discovery. Thus, high-throughput sequencing of multiple diverse organisms in a genus is valuable not only to study phenotypic diversity but also to improve the consistency of metabolic networks and genomes and to guide further research in less-studied phylogenetic clades.

Finally, we collaborated with the group of Dr. Zaida Luthey-Schulten (Dept. of Chemistry, University of Illinois at Urbana-Champaign) to develop a complete, whole-cell model of *Methanosarcina acetivorans*. We characterized size distribution of the cells using differential interference contrast microscopy, finding them to be ellipsoidal with mean length and width of 2.9 μ m and 2.3 μ m, respectively, when grown on methanol and 30% smaller when grown on acetate. We used the single molecule pull down (SiMPull) technique to measure average copy number of the Mcr complex and ribosomes. A kinetic model for the methanogenesis pathways based on biochemical studies and recent metabolic reconstructions for several related methanogens is presented. In this model, 26 reactions in the methanogenesis pathways are coupled to a cell mass production reaction that updates enzyme concentrations. RNA expression data (RNA-seq) measured for cell cultures grown on acetate and methanol was used to estimate relative protein production per mole of ATP consumed. The model captures the experimentally observed methane production rates for cells growing on methanol and is most sensitive to the number of methyl-coenzyme-M reductase (Mcr) and methyl-tetrahydromethanopterin:coenzyme-M methyltransferase (Mtr) proteins. A draft transcriptional regulation network based on known interactions is proposed which we intend to integrate with the kinetic model to allow dynamic regulation.

Products Delivered:

Gu, C. and Ma, P. 2011. Nonparametric regression with cross-classified responses, Canadian Journal of Statistics,39: 591–609. (The article is selected as the best paper in the 2011 volume of the Canadian Journal of Statistics)

Benedict, M., M. Gonnerman, W.W. Metcalf and N.D. Price. 2012. Genome Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon *Methanosarcina acetivorans* C2A. J. Bacteriol. **194:** 855-65. (Highlighted in *Microbe*, March 2012)

Dalpiaz, D., He, X., and Ma, P. 2012. Bias correction in RNA-Seq short-read counts using penalized regression , Statistics in Biosciences , DOI: 10.1007/s12561-012-9057-6.

Gonnerman, M.C., M.N. Benedict, A.M. Feist, W.W. Metcalf. 2013. Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. Biotechnol. J. **8:**1070-9.

Bendict, M.N., J.R. Henriksen, W.W. Metcalf, R.J. Whitaker and N.D. Price. 2014 ITEP: An integrated toolkit for exploration of microbial pan-genomes. BMC Genomics **15:**8.

Peterson, J., P. Labhsetwar, J.R. Ellermeier, P.R.A. Kohler, A. Jain, T. Ha, W.W. Metcalf and Z.A. Luthey-Schulten. 2014. Towards A Computational Model of a Methane Producing Archaeum. Archaea. **2014:**898453.

Youngblut, N.D., J.S. Wirth, J.R. Henriksen, W.W. Metcalf, R.J. Whitaker. 2014. Genomic and phenotypic differentiation among *Methanosarcina mazei* populations from Columbia River sediment. ISME J. *in revision.*

Benedict, M.N., J.R. Henriksen, J. Luke, M.E.M. Metcalf, S.J. Stevens, N.D. Youngblut, N.D. Price, R.J. Whitaker, W.W. Metcalf. 2014. Mapping metabolic models across closely related *Methanosarcina* genomes leads to mutually reinforcing error correction. *In preparation.*

Program Manager: Joseph R. Graber 301-903-1239