ENIGMA Microbiology: Accurate, High-Throughput Identification of Stable Protein Complexes in *Desulfovibrio vulgaris*

John-Marc Chandonia[1,3*] **(JMChandonia@lbl.gov)**, Maxim Shatsky[1,3], Ming Dong[1], Haichuan Liu[2], Lee Yang[1], Jil T. Geller[1], Megan Choi[1], Barbara Gold[1], Nancy L. Liu[1], Marjon Khairy[1], Sonia Reveco[1], Tom R. Juba[5], Bonita R. Lam[1], Evelin D. Szakal[2], Simon Allen[2], Sunil Kumar[4], Farris L. Poole[4], **Steven E. Brenner**[1,3], **Steven C. Hall**[2], **Susan J. Fisher**[1,2], **Michael Adams**[4], **Terry C. Hazen**[6], **Judy D. Wall**[5], **Swapnil Chhabra**[1], **Jian Jin**[1], **H. Ewa Witkowska**[2], **Adam P. Arkin**[1,3], **Gareth P. Butland**[1], **Mark D. Biggin**[1], and **Paul D. Adams**[1]

[1]Lawrence Berkeley National Laboratory [2]University of California, San Francisco [3]University of California, Berkeley [4]University of Georgia, [5]University of Missouri, Columbia, [6]Oak Ridge National Laboratory

http://enigma.lbl.gov

Project Goals: *Desulfovibrio vulgaris* has been selected as a model bacterium for intensive study by ENIGMA because it can reduce heavy metals and radionuclide contaminants present in the soil at many DOE sites, rendering the contaminants insoluble. ENIGMA seeks to model, at a molecular systems level, how this and similar bacteria respond to natural and human induced changes in their environment and how this alters their ability to stabilize contaminants in the soil. A component of our strategy is to develop and use high throughput pipelines to purify and characterize soluble protein complexes. We expect that these interaction data will improve our ability to produce accurate metabolic and regulatory models of key members of microbial communities.

The group led by Mark Biggin has developed a novel method for identification of stable, soluble protein complexes in microbes. In a small-scale pilot study, we showed that many protein complexes survive intact through a series of orthogonal chromatographic methods, with complex components having correlated elution profiles. These profiles were measured with the aid of mass spectrometry (MS) and iTRAQ reagents (Dong et al., 2008). We developed statistical and machine learning methods to analyze a full-scale data set, which were required in order to obtain biologically meaningful results due to the high potential for false positives (FP) caused by co-elution of proteins that are not part of a complex. Our methods were tuned using a manually curated gold standard (GS) set. As a first high-throughput study, we demonstrated this technique in identifying a high-precision subset of stable protein complexes in *Desulfovibrio vulgaris*. We have shown that our predicted network of interactions is significantly enriched in pairs with similar functional annotations. The quantitative information from elution profiles allowed us to develop a statistical model to estimate the false discovery rate in our predictions; because this varies according to how “crowded” the eluted fractions are, we are able to identify a subset of hundreds of highly reliable (i.e., with very low false discovery rate) interactions, as well as a much larger set of interactions that can be predicted with known false discovery rates. Advantages of the tagless approach include not requiring a mutant library (needed for alternative tag-based approaches such as TAP), and a false discovery rate comparable to TAP.

The group led by Gareth Butland has identified a number of protein complexes using TAP. We have developed an automated pipeline for synthesis of tagged gene constructs in collaboration with Swapnil Chhabra (Chhabra et al., 2011). To date, over 700 pulldowns (comprising more than 600 unique *D. vulgaris* strains) have been subject to analysis. In these experiments, more than 10,000 interactions were detected with over 1,000 distinct prey proteins. Using curated GS datasets (as in the tagless analysis), we filtered out ubiquitous proteins and other likely FP, resulting in a set of high-confidence interactions. A number of these interactions have been reciprocally confirmed, using strains in which the original prey protein was tagged and used as bait. Preliminary analysis of the data have identified several novel complexes, including multiple paralogous versions of the DnaJK-GrpE chaperone complex, each of which is bound to a small protein that may act as an allosteric regulator.

References

Chhabra SR, Butland GP, Elias D, Chandonia JM, Fok OY, Juba T, Gorur A, Allen S, Leung CM, Keller K, Reveco SA, Zane GM, Semkiw ES, Prathapam R, Gold B, Singer M, Ouellet M, Szakal ED, Jorgens D, Price MN, Witkowska HE, Beller HR, Arkin AP, Hazen TC, Biggin MD, Auer M, Wall JD, Keasling JD. 2011. Generalized Schemes for High Throughput Manipulation of the *Desulfovibrio vulgaris Hildenborough* Genome. *Applied and Environmental Microbiology 77*:7595-7604.

Dong M, Yang LL, Williams K, Fisher SJ, Hall SC, Biggin MD, Jin J, Witkowska HE. 2008. A "tagless" strategy for identification of stable protein complexes genome-wide by multidimensional orthogonal chromatographic separation and iTRAQ reagent tracking. *J Proteome Res 7*:1836-1849.

Funding: ENIGMA is a Lawrence Berkeley National Laboratory Scientific Focus Area Program supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Foundational Science through contract DE-AC02-05CH11231.