Proposed Renewal of the Harvard/MIT DOE GTL Systems Biology Center 2007-2012

CONTENTS

Introduction.

This GTL-SysBio renewal proposal is configured to either stand alone with existing collaborations or to potentially act synergistically with DOE BioEnergy Research Centers (BRCs) soon to be funded. In particular, we have coordinated with the proposed VIBRANT Center (Vertically Integrated Bioenergy Research Applying Novel Technologies); see figure 1.

Our May 2002 proposal (http://arep.med.harvard.edu/DOEGTL) lead to the only funded GTL Center to propose integration of all five overall goals of the GTL program (below), and one of the few focused on microorganisms relevant to bioenergy production (as contrasted with other centers focused on remediation or sequestration), specifically, we work on *Prochlorococcus*,

Vibrio, Escherichia, Saccharomyces, and *Pseudomonas*
(1) Global proteomics & molecular machines
(2) RNA regulatory network measures
(3) Microbial communities
(4) Modeling of optimality of metabolic fluxes and 3D organization.
(5) Synthetic Biology (based on a supplement funded in 2004)

Overall progress Report for our GTL Systems Biology Center 2003-2006 in context of the original goals

Additional Progress information is distributed into each of the Specific Aims S0-S7 below and the bibliography of 105 publications from the grant so far. Citations below [in square brakets] refer to that bibliography.

(1) Global proteomics & molecular machines

We have developed and just published MapQuant [Leptos et al 2006] which to together with our Proteogenomic mapping and Motif-X algorithm constitute an extensive suite of software tools developed to help integrate and visualize proteomic data in a genomic context. These allow full annotation, quantitation, and discovery of peptide motifs involved in protein folding, complex formation and function. These software tools are available from our web page (http://arep.med.harvard.edu/proteins) and have been applied in the Church, Kucherlapati and Sarracino groups to determining the proteome of the photosynthetic marine bacterium *Prochlorococcus* marinus strains MED4 (adapted to high light levels) and MIT9313 (low light level adapted). These bacteria are especially interesting because its cell division cycle is coupled directly to the circadian rhythm governing the biochemical response to rising and setting of the sun, and therefore the population can be made highly synchronous. This is one of the first proteomics projects where over 80% of the proteins have been monitored over a time course. We observe changes in cell cycle and metabolic enzymes consistent with an optimal utilization of light cycle. We have completed a full "natural" (i.e. graded light/dark cycle 48 hr time series at 2 hr sampling (done in duplicate). We have also done a phage infection time series and found that all phage examined so far encoded photosynthetic genes [Lindell, et al 2005; Sullivan et al. 2006]. In pursuit of understanding very large protein complexes (whose presence in observed in all DOE microbial species), we have established an in vitro synthesis and assembly project described in section 5 below.

(2) RNA regulatory network measures

We have co-designed Affymetrix arrays for two widely different genomes of *Prochlorococcus* MED4 and MIT9313. We have designed Affymetrix Array and have obtained RNA quantitations with it over a set of time-points for MED4. We are using the same sample time-points for RNA and protein measures (see section 1 proteomics above), hopefully improving the interpretation of the combined RNA and protein profiles. Conditions surveyed included diel (circadian & cell cycles), nitrogen, phosphate, and phage infection.

Whole genome expression in *Prochlorococcus* MED4 and MIT9313
Infection of *Prochlorococcus* MED4 by the podovirus P-SSP7. Whole genome expression of

both MED4 and the podovirus. Experiment run, microarray data collected, microarray analysis in progress, proteomic data collection underway (Chisholm & Church labs). Preliminary microarray findings include: Phage encoded photosynthesis genes were expressed; transcription of approximately 40 MED4 genes were enhanced within 2 hours of infection (many of which are of unknown function) whereas overall MED4 mRNA transcript levels declined with time after infection; phage genes were expressed sequentially based on their position in the genome, with 4 distinct temporal expression profiles identified so far (each consisting of a group of genes).

Whole genome expression of *Prochlorococcus* MED4 grown over a diel cycle with a gradual increase and decrease in light intensity to mimic sunlight and maximize light/dark synchrony. Experiment run with samples taken every 2 hours for 48 hours. Microarray data collected, and analysis underway. The same samples have been used in proteomics, photosynthetic activity, and cell cycle analyses.

Effect of light quality on whole genome expression in *Prochlorococcus* MED4 data collected and analysed. Exposure of *Prochlorococcus* MED4 (high-light adapted) and MIT9313 (low-light adapted strain) to nitrogen starvation/stress has been analyzed using the same Affymetrix arrays as the diel cycle. [Tolonen et al. 2006]. Exposure of *Prochlorococcus* to phosphorus starvation. Whole genome microarray expression analyses on MED4 and MIT9313 response to phosphorus stress have been completed [Martiny et al. 2006]

(3) Microbial communities

We [Branda et al. 2006; Chu et al. 2006, Kearns et al. 2006; Morikawa et al. 2006] have studied proteins involved in matrix formation in *Bacillus subtilis* biofilms. Another set of films was studied by binding of exopolymers to the surface of calcite [Perry et al. 2005]. We [Zinser et al. 2006] have continued studies of ecotypes using improved quantitative PCR and extended discoveries of how *Prochlorococcus* cyanophage genomes adapted for infection of open-ocean photosynthetic cells specifically by fine tuning photoreactions via phage RNAs and proteins [Lindell et al. 2006]. By optimizing the use of 23S rRNA [Hunt et al. 2006] and 16S rRNA [Acinas, et al. 2005], the Polz group have pushed down into low abundance genotypic diversity in a natural bacterioplankton and wood-boring [Luyten et al. 2006] populations and fine-scale phylogenetic architecture of a complex bacterial community [Marcelino et al. 2006] and especially stratified microbial assemblages [DeLong et al. 2006]

New methods for whole genome amplification of single microbial cells while initially challenging are very desirable since they mean that metagenomic sequencing can (1) retain information relating multiple nucleic acid species within a cell (lost in ecosystem shotgun methods) and (2) prioritize allowing more comprehensive sampling of both the abundant and the diverse species within an ecosystem. Using a DNA phage polymerase (phi29) we [Zhang et al 2006] have developed tools for whole genome amplification from single cells to facilitate complete sequencing of *Prochlorococcus* from field populations (Chisholm & Church labs) as well as species with no close relatives sequenced like SAR86. This required several improvements in both the amplification and library steps. We are adapting this method to polony sequencing too. We have continued development and optimization of polony-bead (11 micron scale) methods since last year showing that they can be immobilized without gels showing that

ten times higher density (60 million per 1 cm flow cell). We have improved the Sequencing-by-Ligation (SbL) by capping the 3' ends of the template strand with reduced background fluorescence. This allows sequencing of 130 Mbp of raw data in 60 hours using a basic digital microscope at 1/20 the cost of conventional electrophoretic sequencing and at fewer than 0.3 errors per million in regions with 3X or higher coverage. We have partially automated the paired-end tag libraries reducing time from 4 weeks to 3 days.

(4) Modeling of optimality of metabolic fluxes, replication and 3D organization.
We have largely automated the task of going from raw genome sequence annotation to detailed metabolic optimality models in an SBML compatible form (in collaboration with our separately funded BioSpice efforts). We have developed related tools for studying expression dynamics of a cellular metabolic network. [Kharchenko, et al. 2006] Our metabolic engineering and flux tools have been extended to photosynthetic systems [Zucker et al unpublished].

We have developed a computational pipeline to go from annotated genomes to metabolic flux models and kinetic parameter fitting. Modeling optimality of a variety of living processes and sub-processes will be an increasingly crucial (and productive) part of microbial systems measures and synthetic biology. John Aach and Xiaoxia Lin have modeled the interaction of two co-dependent bacterial types in collaboration with Nick Reppas, who has designed and constructed with homologous recombination, a model system involved very well-defined metabolic cross-feeding. [Shendure et al. 2005]
We are trying to automate this design, synthesis and genome-alteration process so that many hypotheses that arise from the modeling segment of any DOE GtL project can be cost-effectively testing. This includes new methods for synthesizing dozens of genes from oligonucleotides.CADPAM (http://arep.med.harvard.edu/cgi-bin/cadpam/worktest/cadpamworks.pl) We (Matthew Wright) have developed tools for analysis of constraints and optimality of 4D chromosome structure during the replication of circular bacterial genomes. We are now beginning to develop new high-throughput experimental chromosome crosslinking methods to provide a solid foundation for such studies in *Caulobacter* [Umbarger et al unpublished].

(5) Synthetic Biology.
We have developed an integrated set of methods for multiplex amplification, error-correction and assembly of oligonucleotides synthesized on high-density photo-, ink-jet-, or electrically-programmable DNA chips (in collaboration with Cerrina, Agilent, and Combimatrix, respectively). In collaboration with Joe Jacobsen's group at MIT, we have improved the error rate using MutS mismatch selection by a factor of ten and have made the overall gene synthesis and assembly automated enough that it is now a stand-alone company called Codon Devices in Cambridge MA. They have delivered the largest commercial construct (35kbp) and came in first in a large gene synthesis competition. We have laid out in detail the methods, advantages of an in vitro synthetic biology "chassis" based on 152 genes (113kbp) involved in central dogma biopolymer synthesis, including key post-synthetic RNA and protein modifications and chaperones [Forster and Church, 2006a;b]. This system is useful for expressing membrane-toxic (or -philic) proteins, ribosome display selection, evolving new codes, new polymers (e.g. mirror-peptides, and poly-esters)

We have designed and are constructing genomes having new genetic codes in vivo. Initially all UAG stop codons remapped to UAA, then all AGG&AGA Arg codons remapped to CGX codons. This will give access to two or more novel amino acids in vivo and enhanced safety of synthetic biology due to a barrier to functioning of any imported or exported nucleic acids. We have developed a semi-automated homologous recombination process with a 6 hr cycle which does not require selectable markers to create many changes in the genome simultaneously [Isaacs, Bang, et al. unpublished].

We are specifically employing synthetic biology as a key part of a vertically integrated effort to engineer plants (especially C4 monocot grasses) and microbial ecosystems (*E.coli* and *Saccharomyces*) to improve productivity with respect to solar and nutrient efficiency in producing and converting lignocellulose into biofuels (alcohols and alkanes). Plant degrading pests and pathogens are being studied for tools to engineer cellulose pretreatment and degradation [Luyten et al. 2006; Bais et al. 2006; Cui et al. 2005]
Design, synthetic and evolutionary strategies to improve production and secretion of large hydrophobic molecules [Shendure et al. 2005; Reppas et al in preparation] (mentioned in section 4 above) is being expanded to include functional scanning of metagenomic libraries for new metabolic and resistances genes [Sommer, Dantas et al. unpublished] .

Proposed activities for the next 5 years. Specific Aims S0-S7.

S0: Potential Synergies with other Centers

This renewal has the same five goals as above, but with two fresh perspectives. First we expect to leverage synthetic biology for all five goals in contrast to its "add-on" status mid-way through our first five year GTL-SysBio grant. For example synthetic biology permits the readout of internal states of complex systems of cells "in situ" and non-destructively (goals 1 & 2). Second

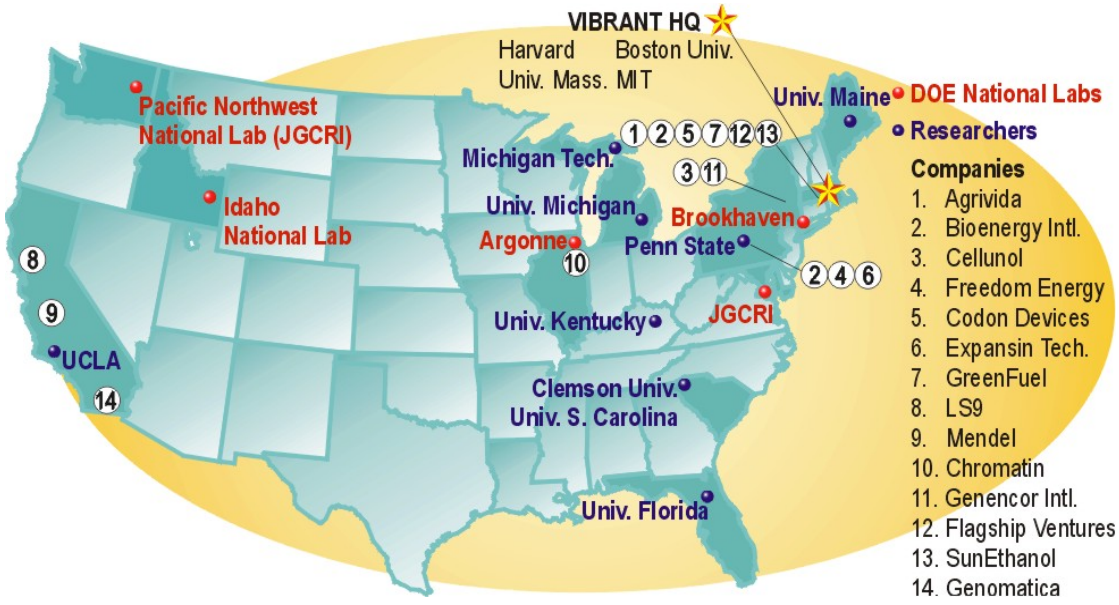


Figure 1. Locations of major VIBRANT research sites, including 19 universities and research facilities, four DOE National Labs, and 14 companies (four of which will receive funding from the Center).

we intend to focus more on possible biofuel applications (whether or not the BRC version of our VIBRANT Center is funded). We begin this with a reflection on the way that the enabling technologies and basic ecosystem analyses that we propose here would interface with one or more BRCs.

Primary Productivity aims to improve the yield of bioenergy-relevant feedstocks that have potential to help meet the estimated billion ton annual biomass resources needed to meet the 30x30 goal. Examples include maize as a food crop biofeedstock and poplar and switchgrass as non-food biofeedstocks and *Crambe* and algae as alternative biofeedstocks. GTL will be used to identify natural variants and construct engineered plants that exhibit improved productivity and stress tolerance. We will also be used to design trait delivery and containment systems that will abet safe and efficient deployment of strains. Potential collaborators include Danny Schnell (U. Mass), C. Joshi (Mich. Tech), Gary Peter (U. Florida), Tom Ulrich (Idaho National Lab), James Zhang (Mendel Biotechnology,) and Daphne Preuss (Chromatin, Inc.).

Cell Wall Synthesis aims to understand and optimize plant cell walls so that biofuel plants can be engineered with a polymer composition that is easier and more efficient to deconstruct and ferment, thereby yielding more economical biofuel production, understand and alter carbon partitioning, to explore and exploit cell wall structure-modifying enzymes such as extensins, and to study and develop enzymes and other techniques for controlling the cross-linking of lignin and polysaccharide polymers. Potential collaborators include Dan Cosgrove, Mark Guiltanan, and Ming Tien (Penn State), C. Joshi (Mich. Tech), Karen Koch (U. Florida), and C-J. Liu (Brookhaven National Lab)

Cell Wall Deconstruction aims to identify and optimize processes that efficiently deconstruct cell wall polymers to component soluble sugars for downstream fermentation. We will bioprospect for new CWD enzymes in insect gut flora, shipworms, and plant pathogens; improve existing enzymes by *in vitro* selection and rational redesign; use novel DNA-display methods for optimizing cellulase enzyme mixtures; and develop "consolidated bioprocesssing" (CBP) by optimizing natively cellulolytic and ethanologenic *C. phytofermentans* and by expressing exogenous glycohydrolases in industrial ethanologens. We also explore the synergies of using plant cell wall-modifying enzymes along with cellulases, develop the breakthrough area of expressing these enzymes in harvest stage plants, and optimize deconstruction in wet silage. Potential collaborators include Lonnie Ingram (U. Florida) and Susan Leschine (U. Mass), Dan Cosgrove, Costas Maranas, Steve Benkovic, and Tom Richard (Penn State), Mike Raab (Agrivida), and Corey Radtke (Idaho National Lab).

Alkanes and Esters will explore several potential breakthrough areas for non-alcohol biofuel generation, including generation of biodiesel from microbes and C30+ hydrocarbons from land plants, and use of algae and cyanobacteria for biodiesel generation and as biofuel biofeedstocks (avoiding the difficulty of lignin Potential collaborators include Wayne Curtis, Don Bryant and John Golbeck (Penn State); Joe Chappelle (U. Kentucky), Stephen Cardayre (LS9)

Alcohol Production aims to optimize fermentation of cell wall sugars, principally to ethanol but also to butanol. In addition to using metabolic engineering and *in vitro* evolution to increase tolerance to ethanol and biomass hydrolysate inhibitors and to increase cofermentation of pentoses and hexoses, we use low-cost genome resequencing and identify mutations in strains optimized for specific improvements, and then use genome reengineering to generate and evaluate combinations of mutations. We also develop new microbes for fermentation that operate efficiently with existing cellulases, in addition to the CBP developments noted above. Potential collaborators include Lonnie Ingram and Susan Leschine, George Church (HMS), Jim Collins (BU), and Costas Maranas (Penn State).

Novel Enabling Technology develops core and new basic enabling technologies applicable to all threads, including single cell and next generation DNA sequencing; accurate and low-cost multiplex DNA synthesis; genome re-engineering; RNA and protein aptamers as cell-internal allosteric metabolite sensors; novel membrane protein production and structure determination methods; and optimization of artificial selections that include consortia and ultra-high complexity *in vitro* selections. Potential collaborators include William Shih (HMS), and Philip Laible and Debbie Hanson (Argonne National Lab).

Systems Biology integrates omics, biochemical, and structural data to develop computational models of regulatory and metabolic networks and characterizations of gene function from a comprehensive whole genome or organism physiology perspective vs. selected pathways that

will be the principal focus of many horizontal thread projects. Our SYSB thread is largely organized as a set of core capabilities (87, 220), including proteomics, cell wall imaging and analysis, metabolomics, low cost DNA sequencing, plant transformation, plant gene expression, and a computational modeling core that will develop large-scale regulatory and metabolic models of microbes and plants. Potential collaborators include include Jim Collins, Tim Gardner, and Daniel Segre (BU), Costas Maranas, Pat Cirino, John Carlson, and Wayne Curtis (Penn State), Jeff Blanchard (U. Mass), K.T. Shanmugam (U. Florida), Vicki Vance (U. S. Carolina), and Bruce Kristal (Brigham and Women's Hospital). Our low-cost DNA sequencing capability represents a collaboration between Joint Genome Institute (see Letter of Commitment).

Synthetic Biology builds on methods from Systems Biology and NET to plan and develop inter-operable, well-characterized parts and devices with a focus on organism-dependent, specific applications that enable flexible, large-scale pathway and genome re-engineering. A key component of our SYNB thread is VIBRANT's participation in a Synthetic Biology Parts and Devices Foundry that integrates synthetic biology resources within VIBRANT (based on a special relationship with Codon Devices, Inc) with those of several other institutions, including Joint Genome Institute. We also use SYNB techniques to optimize the regulatory networks of fermenting microbes, and to develop artificial chromosomes for plants. Potential collaborators include Pam Silver (HMS), Jim Collins (BU), Joe Jacobson and Drew Endy (MIT), and Daphne Preuss (Chromatin, Inc).

Process Integration aims to model, evaluate, and optimize the complete biomass to biofuel production pathway as it proceeds from agriculture through the biorefinery to its output products, including physico-chemical components such as pretreatment, product extraction and modification, and waste management technologies. Mathematical models can then be tested over the entire processing pathway at pilot scales in our "Fields-to-Wheels" (F2W) testbed, which will support a wide range of standard and alternative processing pathways. PI will also develop standard measures and supply standard biomass hydrolysates to functional threads to enable gathering of consistent and realistic data for process modeling and VIBRANT economic estimates. PI also develops new non-biological treatment and product extraction technologies that must integrate with biological components in an optimized biofuel production pipeline. Potential collaborators include Tom Richard (Penn State), Richard Hess, Tom Ulrich and Corey Radtke (Idaho National Lab), and Mike Henson and Scott Auerbach (U. Mass).

Systems Integration explores the systematic impacts of wide-scale biofuel production and usage on national and international economics, the impacts of policies, and Ethical, Legal, and Social Implications (ELSI). It goes beyond Process Integration by considering market forces and the environmental and social effects of a biofuel economy, and evaluates the VIBRANT research portfolio on the basis of expected economic impacts. Potential collaborators include Tom Richard and Clare Hinrichs (Penn State), Erin Baker and Mike Henson (U. Mass), and Richard Hess (Idaho National Lab).

<u>Project S1</u>: Improve and Multiplex Single Cell Genome Sequencing on Environmental Samples and other sequencing applications (HMS: Church). We seek to develop a highly efficient polymerase cloning (ploning) method to perform genome sequencing of single bacterial cells that improves our already demonstrated method for ploning and sequencing single cells without laboratory culturing. Improvements will include methods to remove contaminating cell-free DNA in environmental samples, experimental and computational methods to fill in sequencing gaps and to close the genome, and developing a massively parallel microfluidic platform for ploning 100s to 10000s single cells simultaneously.

<u>Aim S1.1</u> Multiplex selective amplification of genomic regions

We have published a method to sequence over 70% of the genome of a single cell [Zhang et al 2006] using random hexamers priming of strand-displacement polymerase followed by shotgun sequencing. Even the most under-represented regions of this procedure proved to be amenable to targeted sequencing to fill-in gaps left in the shotgun phase. The advantage of this approach relative to shotgun sequencing of BAC (or smaller) clones is that it enables correlation of genes far apart on a chromosome or on completely separate chromosomes in the same cell (or cluster of cells, e.g. predator-prey relations). It also allows us to prioritize rarer cells for sequencing without the high overhead of sequencing the most abundance genomes far more than needed. In aims S1.1-S1.3 we propose adapt some of these procedures which are partially working on human DNA and to automate these and enable multiple paths to selective sequencing of other genomes and metagenomes.

We are developing two strategies for selective genomic subset capture and amplification. One is a scale-up of the "selective circularization" (SeleCirc) approach of Dahl et al (2005, NAR 33:e7). The second is an extension of the molecular inversion probe (MIP) technique to capture extended sequences (i.e. fill-in greater than 1bp, generally 50-150bp; *eMIP*). On an exonome-wide scale, both require tens of thousands of oligonucleotide selectors. We are thus developing technology to allow chip-based oligonucleotide synthesis to replace the cumbersome, expensive column-based approach. This requires one to amplify a complex pool of oligonucleotides from the chip while minimizing the introduction of bias.

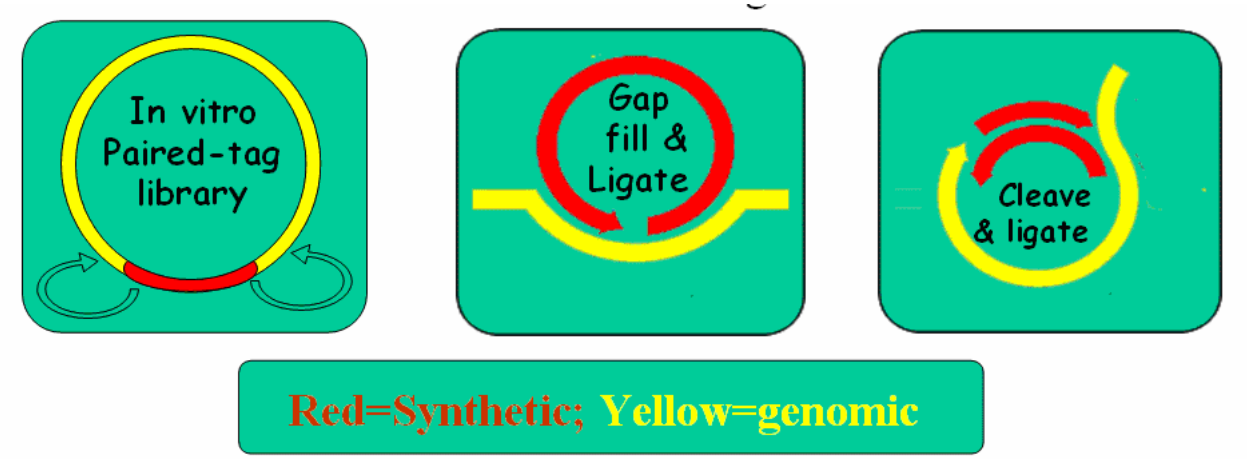


Figure S1.1 Three Polony library strategies: Left to right: paired-end-tags for bridging 1 kbp to 100 kbp shotgun, eMIP and SeleCirc for 50 to 200 bp targeting.

As proof-of-concept, we sought to:

1) Measure the 'purity' of the SeleCirc reaction when performed with a 17,000-plex chip-synthesized oligo pool on human genomic DNA
2) Measure the distribution of target abundances in this reaction
3) Measure the distribution of target abundances in an eMIP reaction performed with a 480-plex column-synthesized oligo pool on genomic DNA

In terms of SeleCirc reaction product purity, out of 96 reads, 76 were placeable to a unique location in the genome. Of these, 76 were from the set of selected sequences.
We constructed a Polony tag library from each reaction (SeleCirc and eMIP), and performed one run of Polony sequencing per library. Shown below is the dynamic range of product species in each selection reaction. We observe at least 5 orders of magnitude variation in both cases.

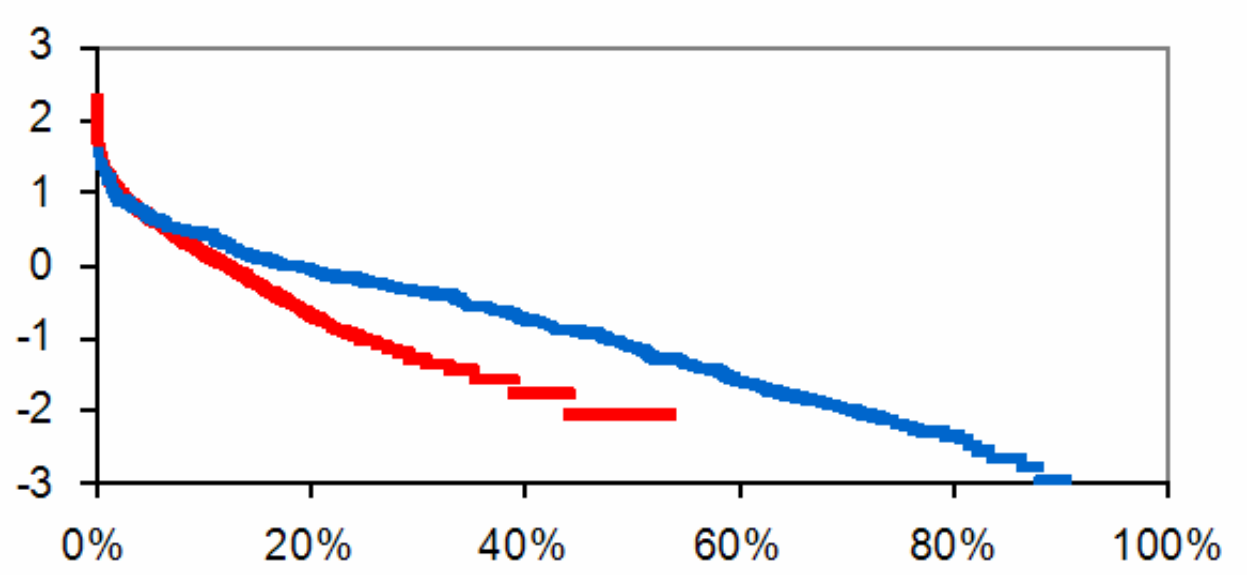


Figure S1.2. Relative abundances of SeleCirc (red) and eMIP (blue) targets. X axis is sorted percentile of exon targets, and Y axis is log10(relative abundance).

Figure S1.3. Our Nikon/Hamatsu multiplex polony sequencing device.

<u>Aim S1.2</u> Improve Polony (aka "next generation") sequencing . We have 7 instruments similar to the one in Figure S1.3 running nearly full time. This is probably the largest number of next-generation machines in use in one location. These processes which are cost-limiting are the

reagents and data acquisition rate (reflected in equipment amortization). We are working with a Massachusetts startup company, Enzymatics to bring down the price of the enzymes, as well as exploring the options of making our own oligonucleotides.

We propose to improve the speed of movement 6-fold over the next year with new X-Y stages and autofocusing. The optics are undergoing a major change to increase data collection rate by 20-fold. The latter involves a very small company still in stealth mode, but our intention and theirs is to get the full description into the public domain as soon as possible, as is already the case for all of the rest of the polony hardware, software, and wetware. We are very excited about this project, but if this level of disclosure for this proposal is unacceptable, we would be happy to drop this sub-aim, and redirect the budget to other aspects of Aims S1.1-4.

<u>Aim S1.3</u> Two-dimensional SNP genotyping by polony sequencing.

Recent advances in SNP genotyping have led to significant improvements in throughput and cost. For example, Illumina's Infinium and Affymetrix's GeneChip technologies can simultaneously assess 500,000-600,000 genotypes for ~$1000. However, it is unclear whether the throughput and cost of genotyping via these platforms can be further improved by the orders of magnitude that will be necessary to keep these platforms as cost-effective options relative to low cost DNA sequencing. We have recently started to develop a next generation SNP genotyping method that is seamlessly integrated with ultra-low cost DNA sequencing and will keep the cost of one million genotypes at least 100-fold below the cost of genome sequencing. This technology will enable the parallel assessment of many SNPs (dimension one) in many samples (dimension two) and hence is termed two dimensional (2D) genotyping. In the short-term, two-dimensional genotyping will reduce the cost of genotyping to approximately $100-200 per million genotypes, while in the 5 year time-frame the cost will be reduced to <$50 per million genotypes.

The key concept of the method is that a large number of padlock probes can be used to specifically capture SNPs from genomic DNA, and the associated SNP identities and genotypes can be subsequently assessed by massively parallel DNA sequencing (Figure 1). This is very similar to the MIP genotyping method (Hardenbol et al. 2003; Hardenbol et al. 2005) but can be implemented with a larger throughput in a more flexible manner. Combining padlock probes with DNA sequencing creates a distinct feature not possible with any of the current array-based methods: multiplexing on a large number of samples (Syvanen 2005). To achieve two-dimensional genotyping, padlock probes circularized on different samples are tagged with unique sample barcodes and pooled for DNA sequencing. The genotype at a given SNP locus of a certain sample will then be decoded by the combinations of three barcodes -- allele barcode, locus barcode and sample barcode – all obtained in a single sequencing run. This provides a tremendous advantage over existing technologies in that a single technology platform can be used for projects with a wide spectrum of SNP number and sample size combinations.

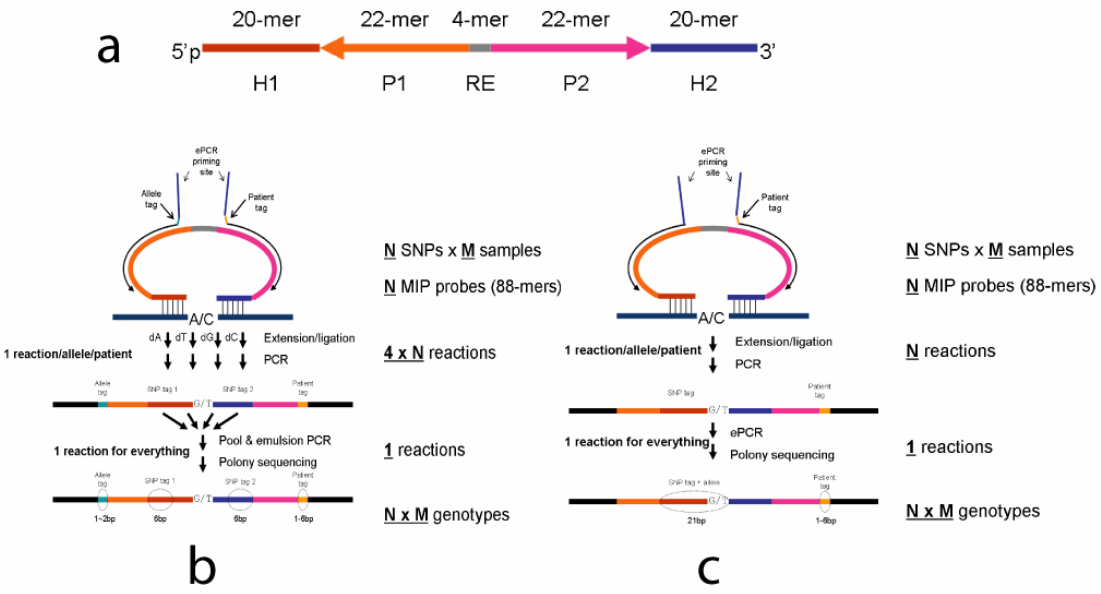


Figure 1. Two-dimensional genotyping by polony sequencing on padlock probes. (a) Design of padlock probes. Similar to MIP probes, each padlock probe has two locus-specific capturing sequences (H1 and H2) at the 3'- and 5'-ends. All probes share two common priming sites (P1 and P2) and a restriction endonuclease recognition site in the middle. **(b)** Polony sequencing on padlock probes. Padlock probes for **N** SNPs are annealed to genomic DNA and circularized similar to MIPs. This part of procedures is carried out one reaction for each allele and each sample separately. After circularization and release of padlocks, PCR is performed with primers carrying allele tags and sample tags. As a result, the amplicons for **M** samples are tagged with unique barcodes and pooled for polony sequencing. SNPs are identified based on the 6+7bp within the capturing sequences H1 and H2, which can accommodate up to 67 million (4^{13}) loci. For a small faction of SNPs with the same 13bp barcodes, additional barcodes can be added between H1 and P1 or P2 and H2. **(c)** An alternative strategy based on sequencing methods with longer read length. For sequencing methods with sufficiently long read lengths padlock probes are used to capture the SNPs themselves for sequencing. Since no allele-specific extension is involved, the circularization can be carried out in one tube instead of four, thereby potentially increasing throughput

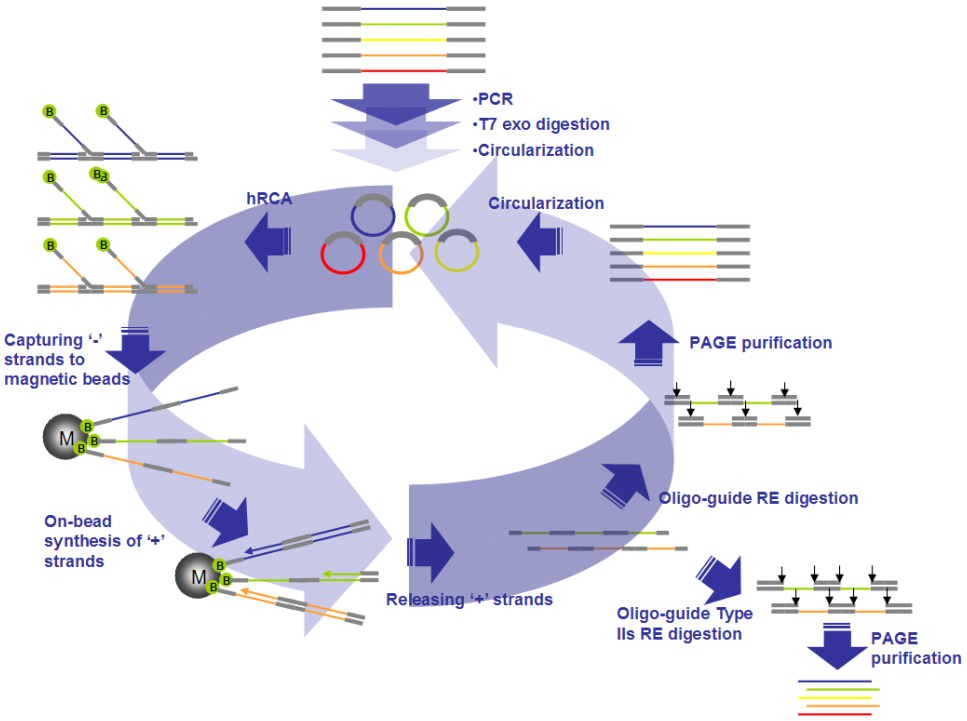


Figure 2. Large-scale production of padlock probes from oligonucleotides synthesized on programmable DNA chips. On-chip synthesized oligonucleotides have common 3' and 5'-adaptors flanking the padlock probes, which allows PCR amplification of all speies in a pool. After size-selection, linear oligonucleotides are circularized using a 'helper oligo" and amplified by hyperbranched rolling circle amplification (hRCA). The amplified DNAs are captured by magnetic beads, and used as templates for the synthesis of the target strands. The adaptors also have Type IIs restriction recognition site, so that padlock probes can be released from the linear concatemers by oligo-guided restriction endonuclease digestion on single-stranded concatemers. In addition, a large amount of monomers can be regenerated by digestion on another restriction enzyme recognition site within the 'helper oligo', and re-circularized. Using this renewal procedure, an unlimited amount of padlock probes can be produced from one set of chip-synthesized oligos.

Development of 2D genotyping relies on recent advances in two closely related fields: DNA synthesis and DNA sequencing. Making millions of padlock probes is non-trivial given the current capability of solid-phase DNA synthesis. Padlock probes are approximately 100bp in length and thus genotyping one million SNPs requires the synthesis of roughly ~100Mb of DNA in a large quantity. Such large scale synthesis would be prohibitively expensive under conventional DNA synthesis methods. With column-based solid phase DNA synthesis, the cost is ~$0.1/base, which translates to a total cost of ~$10,000,000 for probe synthesis alone. Moreover, oligonucleotides longer than 70bp generally require additional PAGE purifications because of the presence of a high percentage of truncated sequences. Thus, it is impractical to produce millions of padlock probes using conventional DNA synthesis method.

We have recently developed a novel method to perform large-scale probe synthesis using programmable DNA chips (Figure 2). We have produced padlock probes targeting 29,908 SNPs using this method. Preliminary results showed that genotyping reads on "A" or "C" are more

accurate than those on “T” or “G” (Table 1). Making genotyping calls based on multiple sequencing reads can effectively eliminate such errors. In addition, direct Sanger sequencing on circularized padlock probes suggested that genotyping errors are often due to cross-contaminations of nucleotides in the reactions. We are optimizing the protocols to reduce nucleotide contaminations. More recently, we are developing another set of padlock probes to target 132,000 exonic SNPs. This probe set will allow us to perform both genotyping on genomic DNA and allele-specific quantification on transcriptomes applicable to a variety of diploid organisms of DOE relevance.

Table 1. Genoyping errors on single sequencing read.

Call\Actual	**A**	**T**	**C**	**G**
A	**95.62%**	1.83%	3.55%	2.61%
T	1.47%	**70.67%**	3.27%	10.44%
C	0.57%	8.41%	**89.29%**	5.95%
G	2.35%	19.09%	3.89%	**81.00%**

Selective amplification of megabase genomic regions.

Existing DNA amplification methods either have very high specificity towards relatively short DNA sequences (e.g. PCR, generally limited to amplifying region s less than 50kb) or have very little specificity (e.g. DOP-PCR, PEP/IPEP, ligation-mediated PCR, MDA). The only way to specifically amplify large genomic regions in the range of hundreds of kilobases to several megabases is to use *in vivo* cloning methods such as BAC/YAC cloning or TAR cloning. The major disadvantage of such *in vivo* methods is that they generally involve screening of hundreds (TAR cloning) to hundreds of thousands (BAC/YAC cloning) of clones to identify the ones containing the target inserts. In the genetic mapping of human diseases, there is a growing need to efficiently amplify/capture megabases genome regions. In performing positional cloning of genes involved in Mendelian diseases, researchers often reduce the search of candidate genes to a chromosomal region several megabases in size by linkage analysis. To identify the causative mutations, PCR-based re-sequencing of coding regions is commonly used but has a very limited power, because causative mutations do not necessarily locate within coding sequences. Complete re-sequencing of the entire candidate region is extraordinary costly and labor intensive because of the difficult of selecting a specific genomic region. We developed a selective hyperbranched amplification method to meet this challenge.

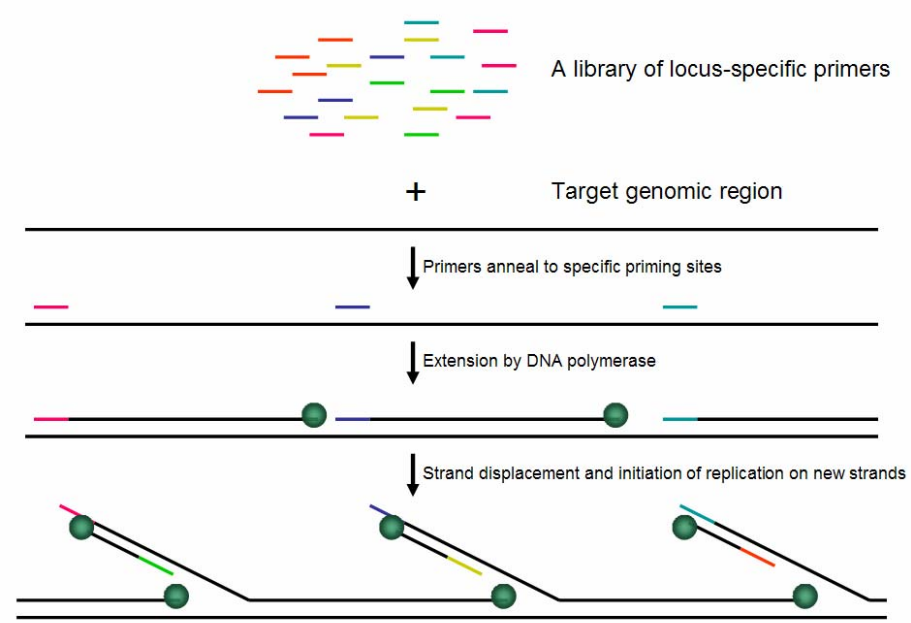


Figure 4. Selective hyperbranched amplification.

The selective hyperbranched amplification method relies on specific annealing of a complex library (~10,000 species per megabases) of single-stranded oligonucleotides (primers) to genomic DNA, and selective amplification of the target region by hyperbranched amplification through a strand displacement mechanism (Figure 4). To ensure high-specificity towards the target genomic region, the each primer has a unique binding site in the target genome. Complex libraries of target specific primers are produced using the chip-based oligo synthesized method developed for 2D genotyping (Figure 2).

We have designed primer libraries for selective amplification of two candidate regions (2.3Mb & 4.1Mb). A total number of 107,264 oligos have been synthesized on a NimbleGen chip, and primer libraries targeting the three candidate regions have been produced. We plan to test the amplification method on normal genomic DNA, then perform amplification and sequencing on variant DNAs. Since re-sequencing a megabase-size region requires only a very small fraction of the sequencing capacity of one polony sequencing run, we will tag each amplicon with a unique sample barcode, which could be decode by hybridization of fluorescently labeled probes. One polony sequencing run could accommodate a pool of amplicons from all the three candidate regions in four patients (~40Mb) at 7.5x sequencing coverage. Such a high degree of over-sampling will reduce the false-positive rate, also could buffer the potential bias during amplification.

<u>Aim S1.4</u> Gene expression profiling by Polony sequencing

In collaboration with Jae Kim in the Seidman Laboratory at HMS, we sought to extend the Polony sequencing technology to enable digital measurement of gene expression. The technology, termed Polony Multiplex Analysis of Gene Expression ("PMAGE"), employs an amplification-free library construction protocol and improved ligation sequencing biochemistry to deliver digital gene expression data at approximately 100-fold lower cost per tag relative to

conventional dideoxy SAGE. We additionally developed gel-less Polony bead arrays which decrease our cost (~4x, $0.02/kb Jan 2007 vs $0.08/kb Aug 2005) and increase our throughput by increasing feature density per unit area. We validated PMAGE by sequencing over 2.2e6 cDNA tags from murine cardiac left ventricles, resulting in the identification of more than 53,000 unique transcripts from approximately 16,000 genes. This sampling depth provided not only a robust gene expression profile, but also the first comprehensive characterization of low abundance mRNAs and transcription factors in the adult mouse ventricular myocardium.

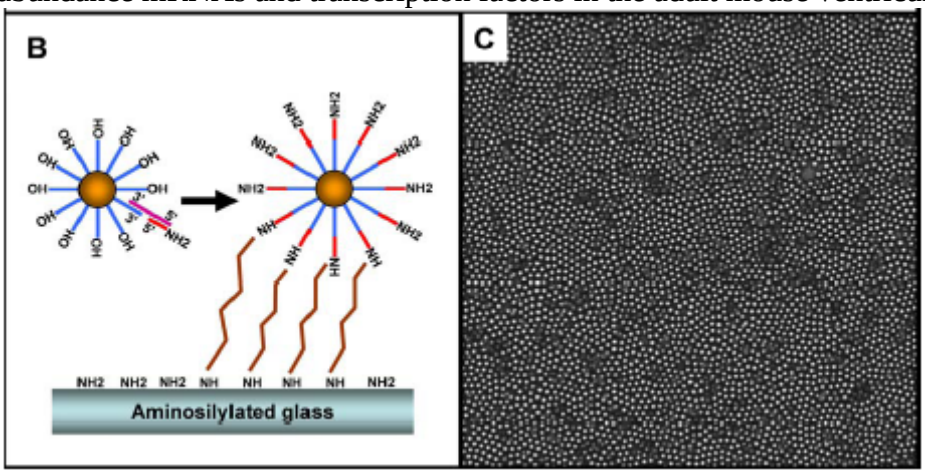


Figure 1. Gel-less arrays are formed by covalent attachment of amine-terminated bead-tethered DNA strands to aminosilanated glass. The resulting array is a densely-packed perfect monolayer which maximizes the number of features per unit area at the resolution limit of our optics.

References for Aim S1

Hardenbol P, et al. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res. 2005 15:269-75.

Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, Davis RW. Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol. 2003 Jun;21(6):673-8.

Kim JB, Porreca GJ, Song L, Greenway S, Gorham JM, Church GM, Seidman CE, Seidman JG (2007) Deep sequencing analysis of gene expression in disease pathogenesis. Science (in revision).

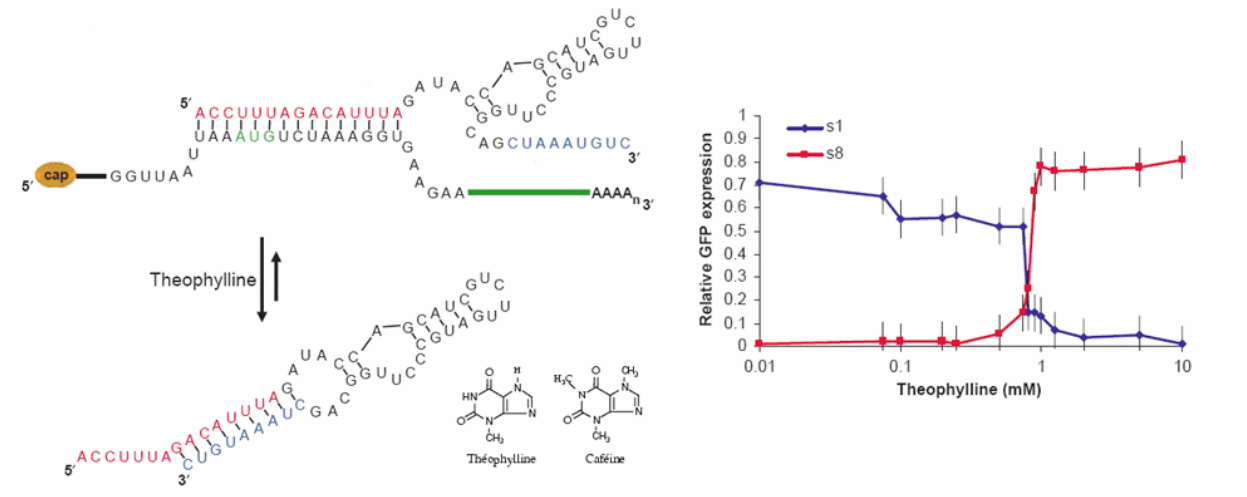
Shendure J, Porreca GJ, Lin X, McCutcheon JP, Rosenbaum AM, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-32

Syvanen AC. Toward genome-wide SNP genotyping. Nat Genet. 2005 Jun;37 Suppl:S5-10.

Zhang, K, Martiny, AC, Reppas, NB, Barry, KW, Malek, J, Chisholm, SW, Church, GM (2006) Sequencing genomes from single cells via polymerase clones. Nature Biotech. Jun;24(6):680-6.

<u>Project S2</u>: Development of engineered RNA molecules to control gene expression, sense metabolites, and optimize the properties of metabolic pathways (HMS: Church). Using our riboregulator technology (Isaacs et al. 2006, Isaacs et al. 2004), we will construct

functional RNAs that can serve as *in vivo* sensors of small molecules (Bayer & Smolke 2005), making translation of target genes responsive to these molecules (Isaacs et al. 2006, Isaacs et al. 2004). Specifically, we will (i) design riboregulators that enable precise tuning of expression of metabolic genes, (ii) design metabolite-responsive riboregulators that enable real-time, *in vivo* measurement of metabolite concentrations and fluxes, (iii) use multiplex DNA synthesis to combinatorially synthesize the genetic elements and (iv) employ these riboregulators to select strains exhibiting optimal flux through biofuel-generation pathways. Because ethanol and biodiesel synthesis depend on total cell reducing power, a first target for intra-cellular metabolite reporting will be NAD+ and NADH, for which RNA aptamers have already been developed (Lauhon & Szostak 1995). We will then develop riboregulators that cover metabolites along broader segments of the glycolytic and fermentation pathways.



<u>Aim S2.1</u> Automate production of allosteric riboregulators adapted to in vivo environments.

Two remarkable successes in this field have recently occurred. The first is the ease of programming structures for cis- and trans- ribo-regulators from first-principles -- i.e. without in vitro selection (Isaacs et al 2004, Isaacs et al 2006). The second is the adapting of aptamers selected only for binding *in vitro* to the sort of trans-regulatory pattern above to create a highly switch-like and tuneable allosteric regulator of mRNA function *in vivo* (Bayer & Smolke 2005) . So thereafter Farren Isaacs shifted to our group

<u>Aim S2.2</u> Integrate the riboregulators from Aim S2.1 with protein reporters and pathway selections.

The need for non-destructive readout of the internal cell states for individual cells is clear.

References for Aim S2

Bayer TS, Smolke CD. 2005. Programmable ligand-controlled riboregulators of eukaryotic gene expression. *Nat Biotechnol* 23: 337-43

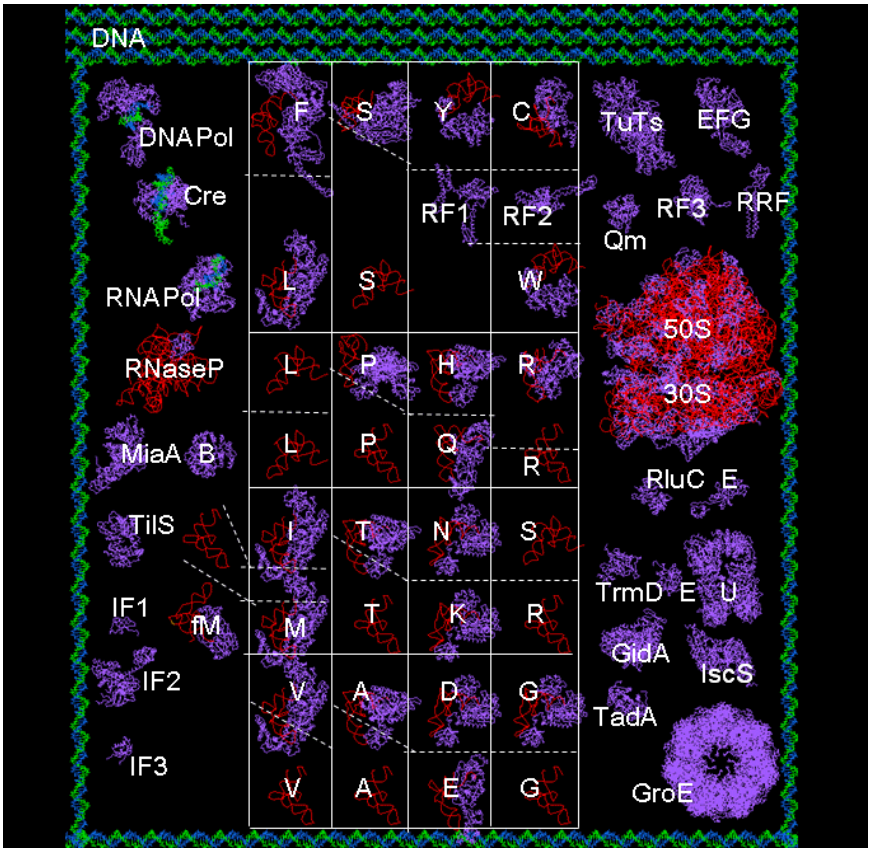
Isaacs FJ, Dwyer DJ, Collins JJ. 2006. RNA synthetic biology. *Nat Biotechnol* 24: 545-54

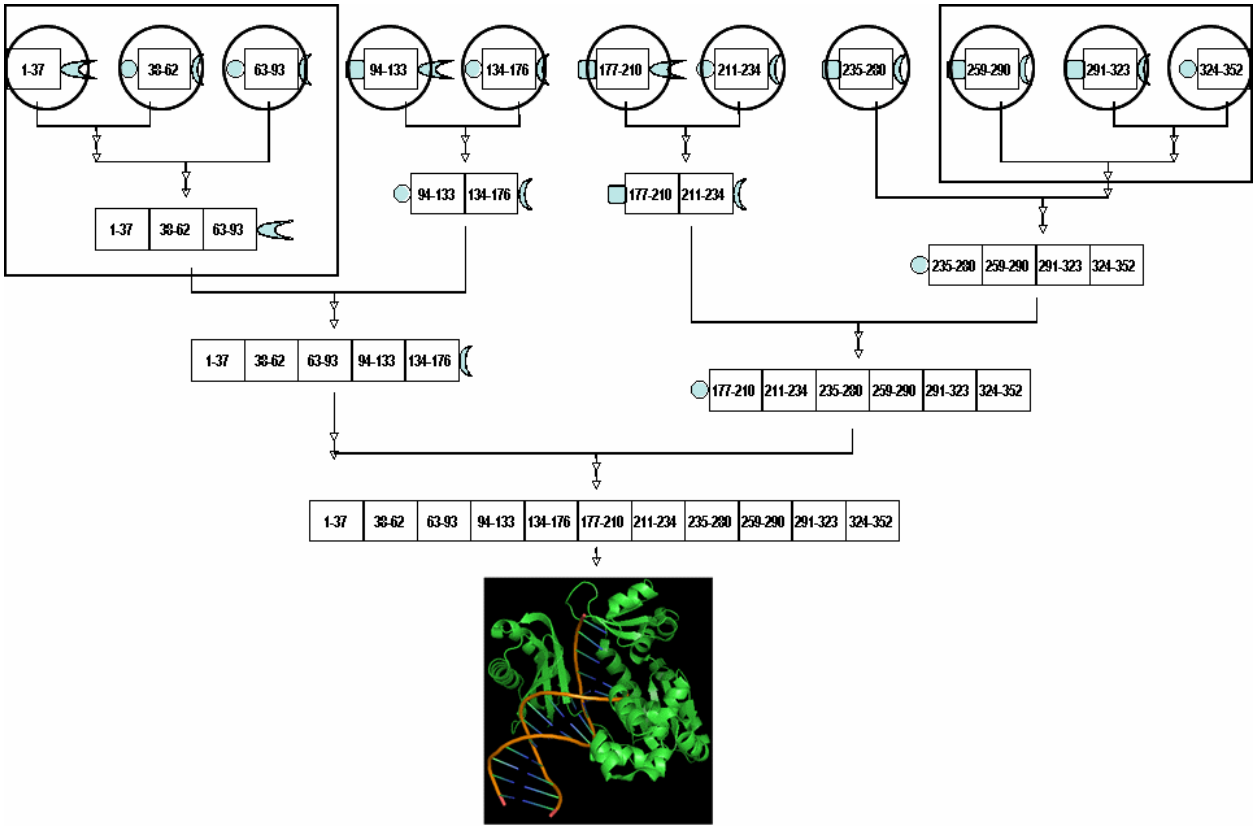
Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ. 2004. Engineered riboregulators enable post-transcriptional control of gene expression. *Nat Biotechnol* 22: 841-7

Lauhon CT, Szostak JW. 1995. RNA aptamers that bind flavin and nicotinamide redox cofactors. *J Am Chem Soc* 117: 1246-57

Project S3: Proteomics in vitro synthesis, in vivo quantitation, and structural studies (HMS: Church lab in collaboration with Forster, Shih, Gygi, Kucherlapati, Sarracino). This project is a core capability with our existing GTL Systems Biology Center that will be made available to VIBRANT projects. It is anticipated that Year 1 VIBRANT demand for proteomics will be within the capacity of this Systems Biology Center service.

Aim 3.1 Enhanced production and selection of proteins *in vitro* (in collaboration with Tony Forster, Vanderbilt)





Dpo4 (*Sulfolobus solfataricus* P2 DNA polymerase IV) 352 AA

Mirror aptamer for drug (by chemical synthesis)

Aim 3.2 Use of RNA aptamers in vivo to measure protein levels real-time, non-destructively and correlate with FT-MS MapQuant data.

Aim 3.3 Enhanced technologies for 3D NMR structures of membrane proteins (in collaboration with William Shih and James Chou, HMS)

References for Aim S3

Forster, AC & Church, GM (2007) Synthetic Biology Projects In Vitro Genome Research 17(1):1-6.

Forster, AC & Church, GM (2006) Toward Synthesis of a Minimal Cell. Nature-EMBO-Molecular Systems Biology 2:45.

Project S4: Functional Genomics Analysis of the Soil Bacterium *P. aeruginosa* and its Interactions with Arabidopsis and *C. elegans* (MGH: Ausubel)

PROGRESS REPORT

The goals of the proposed project were to develop genomic tools for the ubiquitous Gram-negative soil microorganism *Pseudomonas aeruginosa* (strain PA14) and then utilize these tools to study the mechanism by which PA14 develops into biofilm communities. It is thought that biofilms, highly structured communities of a microbial species or species in an extracellular polysaccharide matrix, are the major way that populations of many microbes live in the wild. Eliana Drenkard in the Ausubel lab observed that phenotypic variation, a stable change in phenotype of genetic or epigenetic origin, plays a key role in PA14 biofilm formation (Drenkard & Ausubel 2002). Major goals were to utilize a non-redundant transposon mutation library in *P. aeruginosa* strain PA14 to identify mutants that affect the rate at which spontaneous phase variants appear or form biofilms, carry out the sequencing and annotation of the *P. aeruginosa* PA14 genome, initiate analysis of the PA14 proteome, and study the role of environmental predation in *P. aeruginosa* genome evolution.

Sequence analysis of the *Pseudomonas aeruginosa* PA14 genome. We sequenced the PA14 genome (6.5 MB; (Lee et al. 2006); annotation available at http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/view_gene.cgi) and compared it to the genome sequence of the previously sequenced *P. aeruginosa* strain PAO1 (6.3 MB; (Stover et al. 2000)). PA14 has 5978 predicted ORFs (327 more than PAO1), but the genomes are largely colinear. A list of gene clusters present in PA14 strain but absent in PAO1 showed that many of these clusters have hallmarks of horizontally acquired DNA and contain a high proportion of genes of unknown function.

Role of phenotypic variation in the formation of *P. aeruginosa* biofilms. Previously, we reported that *P. aeruginosa* populations grown in liquid cultures contain a relatively high frequency of phenotypic variants (RSCVs for rough small colony variants) (Drenkard & Ausubel 2002), which undergo

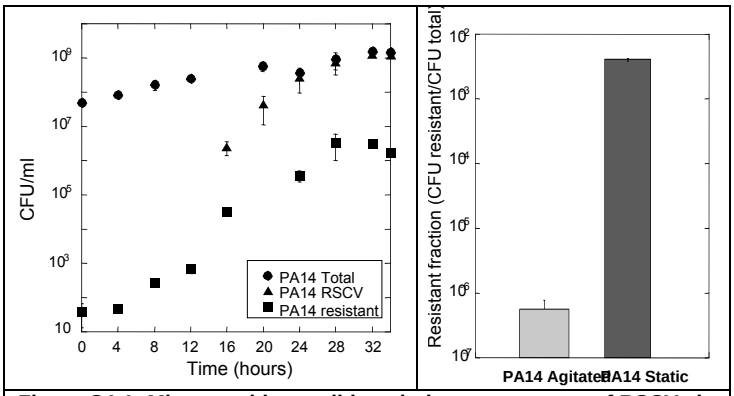


Figure S4-1: Microaerobic conditions induce emergence of RSCVs in PA14 wild-type cultures. (a) Time course showing emergence of RSCV in wild-type population under microaerobic conditions. (**b)** Comparison of population fraction resistant to antibiotics (kanamycin) under agitated (aerobic) and static (microaerobic) conditions.

transient phenotypic changes associated with biofilm formation. We hypothesized that subpopulations of bacteria characterized by increased ability to form biofilms were selected under the low oxygen and nutrient conditions found in biofilms (Drenkard 2003, Drenkard & Ausubel 2002). Consistently, a subgroup of auto-aggregative and highly adherent phenotypic variants are selected in *P. aeruginosa* biofilms (Boles et al 2004, Haussler et al 2003, Kirisits et al. 2005, Webb et al. 2004). Moreover, variant emergence in *P. aeruginosa* biofilms has been linked to an increase in biofilm resistance to environmental stress (Boles et al. 2004).

We tested the hypothesis that oxygen limitation selects/induces RSCVs in biofilms (Werner et al. 2004). RSCVs emerged in the population after approximately 16 hours of incubation under static conditions (Fig. S4-1a), and the RSCV number increased steadily reaching up to 75-76% of the total population after 28-32 hours. Additionally, cultures grown micro-aerobically showed a significant increase in the resistant fraction of the population compared to cultures grown under aerobic (aerated) conditions (Fig. S4-1b). Based on these results, the PA14 non-redundant transposon mutation library (see above) was used to identify *P. aeruginosa* genes involved in RSCV emergence. 24-hour mutant cultures grown under static conditions were used to inoculate media containing kanamycin. Mutants that showed decreased growth compared to the wild-type were selected for further analysis. Five particular mutants were examined in depth (Fig. S4-2). Three mutants contained transposon insertions in *hflC*, *recA* and GID6457 respectively. The protein encoded by the GID6457 gene contains an INT_P4 domain (bacteriophage P4 integrase), which is found in temperate bacteriophages, integrative plasmids, and pathogenicity islands (http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi). These mutants suggest that RSCV induction is associated with activation of genes involved in phage production. In agreement, Boles et al. (2004) have shown that inactivation of *recA* dramatically reduces PAO1 biofilm-induced colony variation (Boles et al. 2004). The other two mutants contain transposon insertions in the *mutS* and *mutL* genes that are involved in the mistmatch repair (MMR) system. In summary, our data suggest that more than one specific mechanism is involved in RSCV formation.

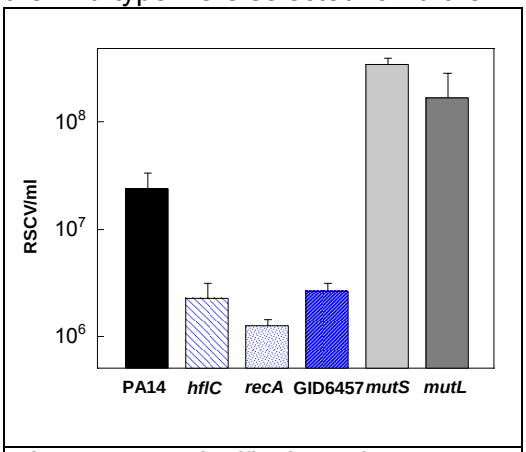


Figure S4-2: Identification of transposon mutants affected in RSCV formation. Comparison of number of RSCV variants formed after 20 hours incubation under microaerobic conditions.

Analysis of expression profiles of RSCV and wild-type PA14 strains. In collaboration with Andrew Goodman and Steve Lory (Harvard Medical School), the expression profiles of RSCV and wild-type PA14 strains were analyzed using *P. aeruginosa* strain PAO1 full genome arrays. Genes that encode proteins involved in energy/metabolism, membrane

transport/secretion and gene regulation were expressed differentially in RSCV. Interestingly, 8 genes contained in a phage-related operon were strongly induced in RSCV and appear to be involved in pyocin (*P. aeruginosa* bacteriocins*)* production. It appears that pyocins are not simply defective phages, but are phage tails that have been evolutionarily specialized as bacteriocins (Nakayama et al. 2000). Importantly, all the pyocin-related genes induced in RSCV were also shown to be up-regulated by oxidative stress in *P. aeruginosa* PAO1 (Chang et al. 2005) supporting the idea that RSCVs are involved in biofilm-mediated resistance to environmental stress. Mutants from the PA14NR set that contain insertions in some of the genes identified in the microarrays are currently being analyzed.

P. aeruginosa **proteomic analysis.** In collaboration with the Church lab and the Proteomics core of the Harvard Partners Center for Genetics and Genomics (HPCGG), we have initiated proteomic analysis of *P. aeruginosa* strain PA14 to validate gene identifications from the PA14 Genome Sequencing project, as well as to pinpoint genes that were missed by gene-prediction methods. PA14 was grown under four conditions including two minimal medium conditions and 2 rich medium conditions. Samples were fractionated using PAGE, digested with trypsin, and analyzed by LC-tandem MS. Data analyses were performed using the X!Tandem and SEQUEST engines. Hits from the reverse translation were used to define expectation value cutoffs for high-confidence peptide identification. So far using X!Tandem data, 2389 genes are represented by two or more high confidence peptide identifications and an additional 572 genes are represented by a single high-confidence peptide. Additionally, peptides corresponding to 18 potential new genes have been identified. Eight of these genes overlap predicted genes in alternative frames, two of which are antiparallel to the previously-predicted genes. The remaining ten genes are in regions previously identified as intergenic.

Project S4: Functional Genomics Analysis of the Soil Bacterium *Pseudomonas aeruginosa* and its Interactions with Arabidopsis and *C. elegans* (MGH: Ausubel)

PROPOSED EXPERIMENTS:

Background

P. aeruginosa is a common soil microorganism that is likely to be subjected to a significant amount of predation by bacterial-grazing soil dwelling organisms including nematodes and protozoa. Defensive strategies necessary to survive predation are an important requirement for the persistence of many bacteria in soil environments (Matz & Kjelleberg. 2005, Matz et al. 2005). Bacterial adaptations developed to avoid predation may include increased bacterial motility, changes in surface characteristics (i.e. O-antigen variability) or biofilm or microcolony formation. The fact that biofilms constitute the predominant life-style of many bacteria in natural environments suggests that biofilms have a high antipredator fitness. Grazing by protozoa stimulates formation of bacterial microcolonies of *Pseudomonas* sp., which increases prey size (Hahn et al. 2000). In addition, once mature biofilms are established, *P. aeruginosa* biofilms display acute toxicity to protozoan predators by up-regulating lethal toxins (Matz et al. 2003). Therefore, *P. aeruginosa* avoids predation by formation of inedible biofilm microcolonies and production of anti-predator compounds by existing biofilms. The overall goal of the proposed experiments is to determine the role of environmental predation in *P. aeruginosa* genome evolution.

Aim S4.1: Screen the *P. aeruginosa* PA14 non-redundant library for anti-predation-related genes using a *C. elegans* feeding assay

Preliminary data. To examine the possibility that *P. aeruginosa* evolution may be driven in part by its ability to avoid predation, an experimental system developed in the Ausubel lab will be used in which an age-synchronized population of *C. elegans* is fed *P. aeruginosa* in place of its normal *E. coli* food (Mahajan-Miklos et al. 1999,Tan et al. 1999). When *C. elegans* are fed a lawn of *E. coli* OP50, they develop, produce progeny and completely consume the lawn. In contrast, when fed PA14, the nematodes die, produce limited progeny and are unable to consume the bacterial lawn. *P. aeruginosa* strains show a wide range of ability to avoid predation as measured by their ability to kill *C. elegans* (Fig. S4-3) and by the accumulation of *P. aeruginosa* in the *C. elegans* intestine. Strain PA14 is highly effective at killing *C. elegans* whereas PAO1 is not (Fig. S4-3). A microarray-based analysis was performed to test the hypothesis that the ability of *P. aeruginosa* strains to avoid predation correlates with the presence of particular genes (Kim et al. 2002)). 285 synthetic oligonucleotides (70 mers) were arrayed corresponding to PA14 genes absent in PAO1. However, no strong correlation was observed between the presence of specific PA14 genes and predation avoidance (Lee at 2006).

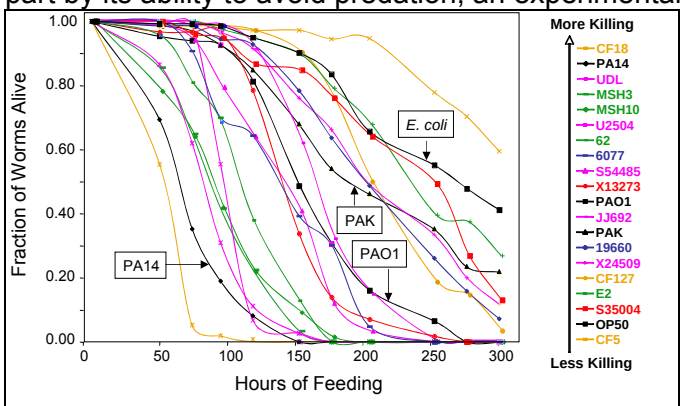


Figure S4-3. *C. elegans* survival curves in the presence of *P. aeruginosa* strains and an *E. coli* control. The names of each strain tested are sorted according to the rank order of their ability to avoid predation by *C. elegans*.

In the microarray experiment described above, is possible that there was insufficient statistical power to identify specific predation-avoidance genes in PA14 among the relatively large set of genes present in PA14 but absent in PAO1. We therefore used a functional approach to first identify PA14 genes required for predation avoidance. We utilized a genome-wide, non-redundant PA14 transposon insertion mutant library in which approximately 80% of

the ORFS are represented by a single transposon insertion (Liberati et al. 2006). As an initial test set, we chose the 332 mutants in this library that correspond to genes specific for PA14 (absent in PAO1) and examined the ability of these mutants to kill wild type *C. elegans*. Although 11 PA14 specific genes were identified that reduced the ability of *P. aeruginosa* to kill *C. elegans*, none were predictive of the ability of the other 18 *P. aeruginosa* strains' ability to kill *C. elegans*. Thus, in contrast to the expectation that the ability to kill *C. elegans* would correlate with specific horizontally transferred blocks of genes, PA14 genes important in avoiding *C. elegans* predation appear to be randomly distributed in other *P. aeruginosa* strains.

Proposed Experiments for Aim S4.1: To expand upon the preliminary work described above and to identify a comprehensive set of anti-predation genes important for survival of *P. aeruginosa*, we are currently screening the 5280 mutants in the PA14 non-redundant mutant library for their ability to kill *C. elegans*. To date, 398 of the 5,280 PA14 mutants have been identified in a primary screen and among these, 83 are either auxotrophs or exhibit poor growth on minimal media. The following additional experiments need to be completed:

1. *Directly ascertain whether the 315 non-auxotrophic PA14 mutants that support growth of C. elegans are defective in killing C. elegans.* Each PA14 mutant will be tested in the standard *C. elegans* killing assay where the death of a population of L4 wild type nematodes exposed to PA14 examined over time. We have already begun this screen and 35% of the PA14 mutants picked up in the primary screen are strongly attenuated in *C. elegans* killing.
2. *Verify the identity of the genes required for predator killing.* We will verify the identity of the PA14 genes required for *C. elegans* killing by examining the phenotype of other mutations in the same gene and by making non-polar deletions in the identified locus. We have available a set of over 27,000 sequenced mutants from which the non-redundant set was picked.
3. *Determine whether the identified genes required for killing of the predator C. elegans are also involved in killing of the predator Dictyostelium discoideum.* *D. discoideum* lives part of its life cycle as a free-living amoebae and is a predator of bacteria. PA14 mutants identified in the *C. elegans* killing assay will be tested using a previously developed *D. discoideum* plate assay for the ability of PA15 to kill and/or be consumed by the amoebae. Work from our lab and the Mekalanos lab (Pukatzki et al. 2002, Pukatzki et al. 2006, Tan et al. 1999) has shown that mutations in *P. aeruginosa* quorum sensing attenuate the killing of both *C. elegans* and *D. discoideum*.
4. *Determine whether in a set of 20 diverse P. aeruginosa strains the presence of the PA14 genes required for predator killing in C. elegans (or C. elegans and D. discoideum) correlates with the ability of a particular P. aeruginosa strain to avoid predation.* The identified set of PA14 genes required for killing of *C. elegans* and D. discoideum predators will be examined in 19 diverse *P. aeruginosa* isolates (see Fig. FMA-3) using custom microarrays as described above or bioinformatics analysis if genome sequences are available.
5. *Develop methods for determining whether predator killing correlates with the ability to avoid being consumed.* We propose to develop methods to directly ascertain whether or not the survival of PA14 is diminished under circumstances where the predator is not killed as effectively. A lawn of bacteria will be exposed to a set number of sterile worms (probably in the 100-500 range) and the consumption of bacteria after a specific time period will be determined.

Aim S4.2: Determine the role of *P. aeruginosa* RSCVs and biofilm in predation avoidance

As mentioned above, biofilms are protective structures that protect bacteria against grazing by predators. While previous reports have suggested that phenotypic variants play an important role in biofilm formation (Drenkard 2003, Drenkard & Ausubel 2002), recent data show that protozoan grazing strongly selects for the emergence of variant phenotypes (Matz et al 2002, Matz et al. 2005). Based on these data and on the "insurance hypothesis", an ecological model

that predicts that diverse communities are better able to resist external stresses, it has been proposed that the functional diversity generated by phenotypic variation provides selective advantage by increasing persistence in variable environments (Matz et al 2005).

Proposed Experiments for Aim S4.2:

1. *Study the role that RSCVs and biofilms play in predation avoidance*. Mixtures of PA14 wild-type and RSCV bacteria will be incubated in liquid media contained in 24-well tissue culture plates along with either *C. elegans* or *D. discoideum* for 3-4 days. Subsequently, the frequency of RSCV versus wild-type phenotypes will be assessed after predation by plating serial dilutions of the mixed cultures. The same experiments will be designed using biofilms grown in tissue culture plates prior to predator inoculation. Controls will consist of planktonic unattached bacteria. Serial dilutions of the scraped biofilms and planktonic cultures will be used to assess survival of planktonic versus biofilm bacteria after comparing to plates that do not contain predators.
2. *Confirm the involvement of RSCVs in predation avoidance*. PA14 mutants defective in RSCV production (described in the Progress Report) will be tested for their ability to survive predation. Liquid cultures from the different mutants obtained after incubation under static conditions (that promote RSCV induction) will be incubated with *C. elegans* and *D. discoideum*. Population survival of the different mutants will be evaluated by plating serial dilutions of the cultures after short exposure to predators.
3. *Identify genes/proteins that are involved in the formation of RSCVs and that have a role in predation avoidance.* Proteomic analysis of *P. aeruginosa* PA14 RSCV variants will be performed using wild-type strain PA14 grown under static conditions in ISOGRO medium (^{15}N labeled, and unlabeled). Aliquots will be removed at 0,8,16,20, and 24 hours and analyzed using mass spectrometry (system LTQ-XL). Relative protein abundances will be analyzed by comparing the MS1-measured abundances of labeled and unlabeled protein using the SIEVE (Thermo Corporation) software package. Comparison of data from the proposed RSCV with data from previously published biofilm development time courses (Southey-Pillig, et al. 2005, Waite,et al. 2006) may reveal proteins specifically related to RSCV formation as those differentially expressed between RSCVs and biofilms.

Aim S4.3: Additional proteomic analysis of *P. aeruginosa* strain PA14

In addition to the proteomic analysis that we will carry out related to RSCV formation, we propose to survey protein expression in a variety of growth conditions including growth in the presence of homoserine lactone quorum sensing signalling molecules and growth in the presence of other organisms (such as plants). The latter condition is particularly relevant to the experiments proposed for the DOE BioEnergy Research Center (VIBRANT proposal) since preliminary data suggest that a variety of plant cell-wall degrading enzymes are excreted or expressed on the cell surface when Pseudomonads encounter plants. Proteomic analysis will be carried out as described in Aim 2.

PROPOSED DELIVERABLES

1. Full-genome elucidation of *P. aeruginosa* strain PA14 genes involved in predation avoidance identified by transcriptional and proteomic profiling and by functional genomic analysis.
2. Determination of the role that RSCV and biofilms play in predation avoidance.
3. Proteomic analysis of *P. aeruginosa* PA14 genes activated by interaction with Arabidopsis.

References for Aim S4

Boles, B. R., M. Thoendel, and P. K. Singh. 2004. Self-generated diversity produces "insurance effects" in biofilm communities. Proc Natl Acad Sci U S A 101:16630-5.

Chang, W., D. A. Small, F. Toghrol, and W. E. Bentley. 2005. Microarray analysis of Pseudomonas aeruginosa reveals induction of pyocin genes in response to hydrogen peroxide. BMC Genomics 6:115.

Drenkard, E. 2003. Antimicrobial resistance of *Pseudomonas aeruginosa* biofilms. Microbes Infect 5:1213-9.

Drenkard, E., and F. M. Ausubel. 2002. *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. Nature 416:740-3.

Hahn, M. W., E. R. Moore, and M. G. Hofle. 2000. Role of Microcolony Formation in the Protistan Grazing Defense of the Aquatic Bacterium Pseudomonas sp. MWH1. Microb Ecol 39:175-185.

Haussler, S., I. Ziegler, A. Lottel, F. von Gotz, M. Rohde, D. Wehmhohner, S. Saravanamuthu, B. Tummler, and I. Steinmetz. 2003. Highly adherent small-colony variants of *Pseudomonas aeruginosa* in cystic fibrosis lung infection. J Med Microbiol 52:295-301.

Kim, C. C., E. A. Joyce, K. Chan, and S. Falkow. 2002. Improved analytical methods for microarray-based genome-composition analysis. Genome Biol 3:RESEARCH0065.

Kirisits, M. J., L. Prost, M. Starkey, and M. R. Parsek. 2005. Characterization of colony morphology variants isolated from Pseudomonas aeruginosa biofilms. Appl Environ Microbiol 71:4809-21.

Lee, D. G., J. M. Urbach, G. Wu, N. T. Liberati, R. L. Feinbaum, S. Miyata, L. T. Diggins, J. He, M. Saucier, E. Deziel, L. Friedman, L. Li, G. Grills, K. Montgomery, R. Kucherlapati, L. G. Rahme, and F. M. Ausubel. 2006. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol 7:R90.

Liberati, N. T., J. M. Urbach, S. Miyata, D. G. Lee, E. Drenkard, G. Wu, J. Villanueva, T. Wei, and F. M. Ausubel. 2006. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103:2833-8.

Mahajan-Miklos, S., M. W. Tan, L. G. Rahme, and F. M. Ausubel. 1999. Molecular mechanisms of bacterial virulence elucidated using a Pseudomonas aeruginosa-Caenorhabditis elegans pathogenesis model. Cell 96:47-56.

Matz, C., P. Deines, and K. Jurgens. 2002. Phenotypic variation in *Pseudomonas* sp. CM10 determines microcolony formation and survival under protozoan grazing. FEMS Microbiology Ecology 39:57-65.

Matz, C., and S. Kjelleberg. 2005. Off the hook--how bacteria survive protozoan grazing. Trends Microbiol 13:302-7.

Matz, C., D. McDougald, A. M. Moreno, P. Y. Yung, F. H. Yildiz, and S. Kjelleberg. 2005. Biofilm formation and phenotypic variation enhance predation-driven persistence of *Vibrio cholerae.* Proc Natl Acad Sci U S A 102:16819-24.

Matz, C. B. T., Rice SA, and K. S. 2004. Microcolonies, quorum sensing and cytotoxicity determine the survival of Pseudomonas aeruginosa biofilms exposed to protozoan grazing. Environmental Microbiology 6:218-226.

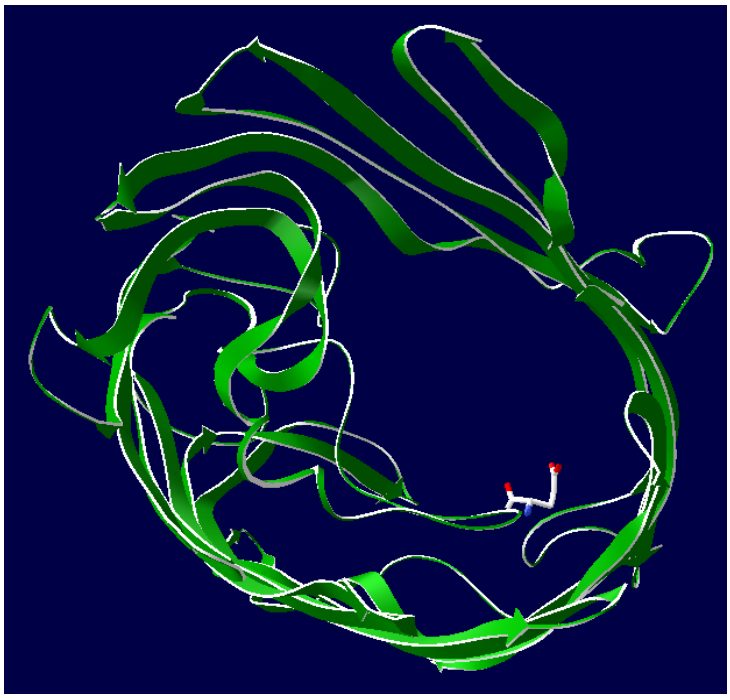
Nakayama, K., K. Takashima, H. Ishihara, T. Shinomiya, M. Kageyama, S. Kanaya, M. Ohnishi, T. Murata, H. Mori, and T. Hayashi. 2000. The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol 38:213-31.

Pukatzki, S., R. H. Kessin, and J. J. Mekalanos. 2002. The human pathogen Pseudomonas aeruginosa utilizes conserved virulence pathways to infect the social amoeba Dictyostelium discoideum. Proc Natl Acad Sci U S A 99:3159-64.
Pukatzki, S., A. T. Ma, D. Sturtevant, B. Krastins, D. Sarracino, W. C. Nelson, J. F. Heidelberg, and J. J. Mekalanos. 2006. Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci U S A 103:1528-33.
Southey-Pillig, C. J., D. G. Davies, and K. Sauer. 2005. Characterization of temporal protein production in Pseudomonas aeruginosa biofilms. J Bacteriol 187:8114-26.
Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson. 2000. Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature 406:959-64.
Tan, M. W., S. Mahajan-Miklos, and F. M. Ausubel. 1999. Killing of Caenorhabditis elegans by Pseudomonas aeruginosa used to model mammalian bacterial pathogenesis. Proc Natl Acad Sci U S A 96:715-20.
Tan, M. W., L. G. Rahme, J. A. Sternberg, R. G. Tompkins, and F. M. Ausubel. 1999. Pseudomonas aeruginosa killing of Caenorhabditis elegans used to identify P. aeruginosa virulence factors. Proc Natl Acad Sci U S A 96:2408-13.
Waite, R. D., A. Paccanaro, A. Papakonstantinopoulou, J. M. Hurst, M. Saqi, E. Littler, and M. A. Curtis. 2006. Clustering of Pseudomonas aeruginosa transcriptomes from planktonic cultures, developing and mature biofilms reveals distinct expression profiles. BMC Genomics 7:162.
Webb, J. S., M. Lau, and S. Kjelleberg. 2004. Bacteriophage and phenotypic variation in Pseudomonas aeruginosa biofilm development. J Bacteriol 186:8066-73.
Werner, E., F. Roe, A. Bugnicourt, M. J. Franklin, A. Heydorn, S. Molin, B. Pitts, and P. S. Stewart. 2004. Stratified growth in Pseudomonas aeruginosa biofilms. Appl Environ Microbiol 70:6188-96.

Project S5: Generation of Metabolic Analysis Models for Center Organisms and Evolutionary Optimization (HMS: Church). We have developed automated methods (294) for generating metabolic models for organisms of interest to the GTL Systems Biology Center (Feist et al. 2007), a capacity that we will make available to VIBRANT and/or other GTL Centers. We have used Pathway Tools (Karp et al. 2002) to generate numerous models (including 14 strains of cyanobacteria), and also employ Mixed Integer Linear Programming to fill gaps in models. We plan to incorporate use of the Atomic Reconstruction of Metabolism (Arita 2003) database, and novel algorithms under development, to improve reaction balancing and function assignment to enzymes with non-specific EC numbers. (Segrè et al. 2002, 2003, 2005)

<u>Aim S5.1</u>: Metabolic Analysis Models for Center Organisms and Evolutionary Optimization We have developed automated methods

<u>Aim S5.2</u>: Metabolic Evolutionary Optimization We have developed automated methods



Co-evolution of mutual biosensors/biosynthesis sequenced across time & within each time-point -- Independent lines of TrpD & TyrD co-culture.

We observe:

5 OmpF mutations: pore protein coding region, mainly large, hydrophilic AA to small and/or hydrophobic AA substitutions, e.g. 42R-> G,L,C, 113 D->V, 117 E->A

2 OmpF promoter cis-regulatory up mutations -12A->C, -35 C->A

5 Lrp trans-regulator knock-outs: 1b del, 9b del, 8b del, IS2 insert, R->L in DNA-binding domain.

We see heterogeneity within each time-point and multiple “solutions”

References for Aim S5

Arita M. 2003. Representing Metabolic Networks by the Substrate-Product Relationships *Genome Informatics* 14: 300-1
Feist AM, Henry CS, Reed JL, Krummenacker M, Zucker JD, et al. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1261 ORFs and thermodynamic information (submitted). *Molecular Systems Biology*
Karp PD, Paley S, Romero P. 2002. The Pathway Tools software. *Bioinformatics* 18 Suppl 1: S225-32
Segrè D, Deluna A, Church GM, Kishony R. 2005. Modular epistasis in yeast metabolism. *Nat Genet* 37: 77-83
Segrè D, Vitkup D, Church GM. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 99: 15112-7
Segrè D, Zucker J, Katz J, Lin X, D'Haeseleer P, et al. 2003. From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* 7: 301-16
Shendure J, Porreca GJ, Lin X, McCutcheon JP, Rosenbaum AM, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-32

Project S6: Genomic structure, metagenomics, horizontal gene transfer, and natural diversity of *Prochlorococcus* and Vibrio

(MIT: Sallie Chisholm, Martin Polz, Eric Alm; HMS: George Church)

INTRODUCTION

Elucidating the design principles of microbial systems is critical not only for exploiting them to serve human needs, but also for understanding their natural roles in regulating biospheric processes. To achieve this mission will require the development of relevant model microbial systems that can be studied at all levels of biological organization—from the information encoded in genomes to global population phenomena, and the processes that link the two.

Over the past 3 years, as part of the Harvard-MIT Genomics:GTL Center, the Chisholm and Polz laboratories have been working on developing *Prochlorococcus* and *Vibrio* as models for studying microbial systems biology (see Appendix 1 for a detailed 'progress to date" section). We propose to continue this work, viewing these organisms as molecular machines that could be useful for synthetic biology and bioenergy production. Our overarching goal is to develop a deep understanding of the design of these cells, the variations in their designs, and the constraints that have shaped this variation at the cell-environment interface. We have also added a new member to our team, Eric Alm (Assist. Prof. MIT), who brings his expertise in metabolic and evolutionary modeling, and systems biology.

THE MODEL ORGANISMS AND THEIR DOE RELEVANCE

Cost-effective bioenergy production will likely require the design of open, large-scale systems, which are therefore subject to challenges, such as internal fluctuations, predation and invasions by natural populations. Thus ideally, such systems would achieve stability and resilience by mimicking natural design principles and/or self-organize according to the required outputs. This is the framework motivating our work on the model systems, *Prochlorococcus* and *Vibrio*. These organisms represent fundamentally different habitats and lifestyles. The former is an open-ocean autotroph that thrives with a minimal genome on minimal resources in hyper-oligotrophic environments, while the latter is a cosmopolitan coastal heterotroph that exploits diverse, high-nutrient habitats. They have already been the focus of

intense study under the auspices of our existing Harvard-MIT Genomics:GTL Center, and our progress to date is reviewed in Appendix 1).

Our model organisms are not only useful for studying microbial diversity and self-assembly, they also have central relevance to the production of bioenergy. *Prochlorococcus* represents the most efficient photosynthetic machine nature has presented us with (BOX 1), while, collectively, *Vibrio* represents one of the most versatile heterotroph groups with metabolic potential covering alcohol and light oil production (BOX 2).

PROCHLOROCOCCUS

The most efficient photosynthetic cell

Prochlorococcus is uniquely suited as a model organism for the study of photosynthesis as a basis of bioenergy systems and synthetic biology.

- It is the most efficient light absorbing cell and carbon fixing machine (Box 1);
- It has highly efficient nutrient utilization (low ambient nutrient concentrations);

Box 1. *Prochlorococcus*: The Most Efficient Photosynthetic Cell

Oxygenic photosynthesis occupies broad evolutionary design space. *Prochlorococcus* occupies a uniquely useful portion of that space, and thus is a useful model for the development of bioenergy. It is the smallest cell, and has the smallest genome, of any oxygenic phototroph, and requires only sunlight and inorganic compounds for growth.

Light Absorption: *Prochlorococcus* is the most efficient light absorber of any micro-alga. Its small size and high pigment content place this cell "in a singular domain, where the probability for photons to be absorbed by a cell can exceed that of being scattered" (Morel et al., 1993). More quantitatively, *Prochlorococcus* has the highest Q_a/Q_b Ratio of any known cell, where Q_a is the efficiency factor for absorption and Q_b that for scattering. (Fig. B1) Only in *Prochlorococcus* does this ratio exceed one, resulting in extremely efficient use of photons. This high efficiency is necessary for this cell to maintain populations at the bottom of the ocean euphotic zone (>200m) where it is known to thrive.

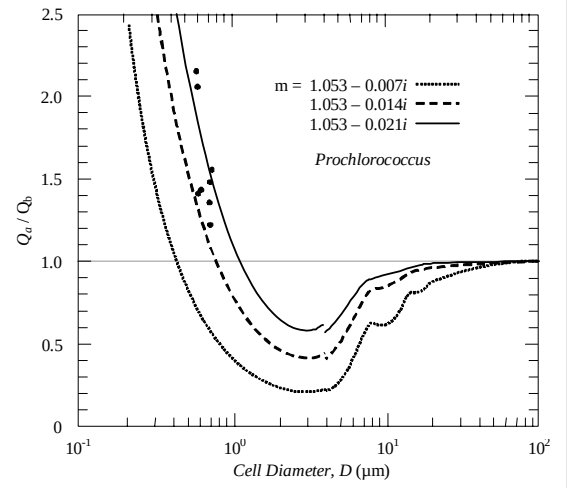


Figure B1. Ratio of efficiency factor for absorption to the efficiency factor for scattering, plotted as a function of the diameter *D*. The theoretical curves are computed (van de Hulst approximation for $D > 4$ μm and exact Mie solution for $D < 4$ μm) for the three values of the complex index of refraction as indicated. Superimposed are the values of Q_a/Q_b for *Prochlorococcus* strains (black dots) , which are the only algal strains that fall above a value of 1.0. (Morel et al., 1993)

Carbon fixation: *Prochlorococcus* is also a very efficient carbon fixing machine: chlorophyll-specific carbon fixation rates (P^B values) in excess of 10 fg C/(fg Chl·hr) have been recorded (Bruyant et al., 2005). This is at the upper extreme of phytoplankton production rates reported by (Behrenfeld and Falkowski, 1997). The physiological basis for this is not clear, but it is known that *Prochlorococcus* has a Form IA Rubisco, phylogenetically distinct from the Form IB found in other cyanobacteria and green plants (Hess et al., 2001). It is possible that this has a high inherent k_{cat} . Allometric considerations predict that P^B values will be inversely correlated with cell size (Montecino and Quiroz, 2000). Alternatively, iIt may be that the small size of *Prochlorococcus* cells enables an exceptionally efficient carbon concentrating mechanism, resulting in very high pCO_2 in the carboxysomes. Aspects of this carbon fixation efficiency are addressable experimentally, by *in vitro* enzymological assays of Rubisco and *in vivo* studies of carbon metabolism.

- It has a minimal genome (~2,000 genes), relatively few regulators and simple regulatory networks, which can be an interesting chassis for building other features;
- Its natural habitat can easily be reproduced in the laboratory;
- Its relatively slow growth rate and tight synchrony on light-dark cycles make fine scale resolution of metabolic processes possible;
- The complete transcriptome over the light-dark cycle has been characterized;
- Many genomic variants (ecotypes) exist in culture and have been physiologically characterized, and 12 have been sequenced;
- Hundreds of *Prochlorococcus*-specific phage have been isolated, and 18 genomes have been sequenced so far;

- The complete transcriptomes of host and phage during infection have been characterized
- We can recognize and count 'wild' *Prochlorococcus* cells using flow cytometry, and thus can enumerate them, and flow sort them, for subsequent meta-genome and transcriptome analysis;
- We can study the dynamics of ecotypes of *Prochlorococcus* in their natural environment.

Oceanic dominance as a carbon converter

Prochlorococcus is the single most important contributor to gross photosynthesis on a global scale (Goericke and Welschmeyer, 1993; Liu et al., 1997; Partensky et al., 1999). Moreover, this group can be easily and rigorously studied in situ, and as such it serves as a sentinel of open ocean ecosystem state, and a 'miner's canary' for global change (Fig. 1). In order to understand the mechanisms underlying these global changes, we must understand the regulation of cellular processes and their diversity.

Studies of the ocean metagenome are also revealing that *Prochlorococcus* is the most well represented microbial group in the tropical and sub-tropical Atlantic and Pacific (DeLong et al., 2006; Venter et al., 2004), the highest representation found for any single taxon.

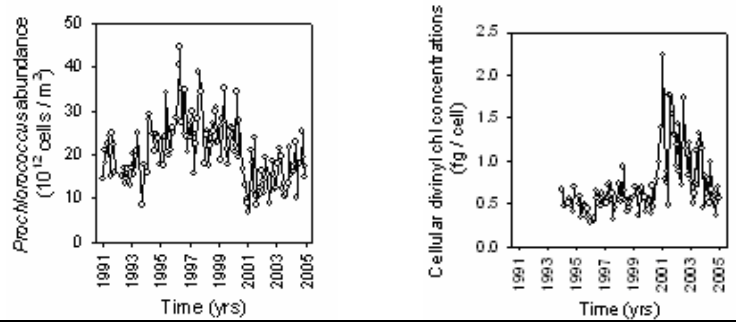


Figure 1. Response of *Prochlorococcus* populations to large scale climate induced changes. (A) abundance (B) mean chlorophyll per cell, in the North Pacific. The dramatic change between 1999 and 2001 result from a late 90's El Nino event. Data from (HOT-DOGS©).

VIBRIO

Versatile carbon converting machines

Vibrios display very high metabolic diversity and physiological adaptability, and are longstanding models for heterotrophic processes.

- Many vibrios degrade cellulose and chitin, the most abundant polymers;
- Some strains produce substantial quantities of aliphatic hydrocarbons using a (to date) unique pathway (BOX 2);
- They can degrade aliphatic and aromatic hydrocarbons;
- Vibrios can carry out numerous fermentations, including ethanol production.

The distribution of these metabolic properties among different *Vibrio* taxa (from groups of strains to species) indicates that many of the pathways are modular and mobile (Fig. 2). Because *Vibrio*s possess many means of genetic exchange, they provide an excellent platform with which to explore the mixing and matching of genetically encoded metabolic properties in populations of cells under natural and artificial selection.

Fast growth rate, and genetic diversity and malleability

The *Vibrio*s are a good model for exploring the biotechnological potential of naturally existing genetic diversity for biofuel production and strain engineering. They possess a diversity of metabolisms, the fastest recorded growth rates, and are genetically malleable in the laboratory. We will develop and test new approaches to experimental evolution and strain engineering that exploit the genetic diversity of co-existing *Vibrio* populations in a way that mimics the adaptive processes occurring in complex natural systems.

Figure 2. Metabolic versatility of microdiverse *Vibrio* strains. Strains are organized into columns according to phylogenetic relationships (Hsp60 gene); rows show positive (red), negative (white) and ambiguous (pink) ability to metabolize different carbon substrates shown on the left. Data compiled from replicate BIOLOG plates; strains are primarily *V. splendidus* except outgroup species in 6 columns on the right.

BOX 2. Oil-producing *Vibrio*s

Lipid production from higher plants or microalgae may be a feasible route for alternative fuel production. Liquid fuels can be produced from algal strains with high lipid content by various physical and/or chemical processes, such as direct lipid extraction to produce a diesel-oil substitute, transesterification of algal lipids to form ester fuels, and hydrogenation of algal biomass to yield hydrocarbons. Oily substances are also formed via liquefaction of microalgal biomass by thermochemical reactions under high-pressure and high-temperature conditions. Oil-producing bacteria would be an ideal complement to algae and would have several advantages over algae. They can grow much faster, sustain higher density populations, and have the potential to convert metabolic byproducts of other organisms or waste biomass to hydrocarbon. However, their lipid has been believed to be generally too low (<1% of dry cell mass) for commercially viable production. Recently, a *Vibrio furnissii* strain was isolated, which had an extremely high yield of lipids and hydrocarbons (Park et al., 2001). The strain was isolated from activated sludge, which is diluted with seawater for cooling purposes, on media containing acetate, propionate, glucose, glycerol and yeast extract (with trace minerals and vitamins) (Park et al., 2001). In pure culture, a lipid layer on the surface of the media was noticed. When analyzed, these hydrocarbons were $C_{15}H_{32}$, $C_{18}H_{38}$, $C_{21}H_{44}$, $C_{22}H_{46}$ and $C_{24}H_{50}$ corresponding to kerosene and light oil. Maximum hydrocarbon accumulation was achieved by early stationary phase. The total amount of lipid reached 124% of dry weight and hydrocarbon content was 48% of total lipid. Thus carbon yield was ~11% for hydrocarbon production on the basis of the consumed amount of substrates (Park et al., 2001). These values are comparable to the best oil producing algal strains.

Further studies showed that hydrocabons are only produced by growing cells and are not extensively metabolized after growth ceases (Park, 2005). A wide variety of sugars and organic acids (but not amino acids) led to hydrocarbon accumulation with pentanoic acid producing the highest yield (per gram dry weight) but glucose the highest amount in media standardized to 60 mM C (glucose resulted in 26 mg hydrocarbon per 50 ml culture). The hydrocarbon characteristics differed with different carbon substrates but variation in C:N ratio and oxygen concentration had no apparent effect on hydrocarbon yield. A striking difference in alkane production of *V. furnissii* to most other organisms is the production of even numbered C-backbones. Odd numbered alkanes are believed to result from reductive decarbonylation of fatty aldehydes. *V. furnissii*, on the other hand, produces alkanes via reduction of 1-alcohol. Using radiolabeled hexadecanoic acid, the production of all reduction intermediates was confirmed: acid, alcohol, aldehyde, alkane (Park, 2005). This pathway is to date unique; only the reverse, alcohol production from alkane, has been documented.

Overall, *V. furnissii* M1 is unique among bacteria in that no other strains have been documented to have such high hydrocarbon yields. Moreover, it is significant that *Vibrio*s can reach extremely high growth rates and are ‘algal’ associated organisms so that one can imagine a stable, sun light driven oil producing consortium.

* * *

Collectively, these features make ***Prochlorococcus*** and ***Vibrio*** excellent model systems for "Cross-scale Systems Biology," and excellent cellular chassis or power supplies for bioenergy production and conversion. Fundamentally, we must understand how an organism adapts to its natural environment, across all scales of influence, so that we can determine how it should be modified, for maximal productivity in a man-made environment.

PROPOSED WORK

Our objective is to gain a deep understanding of the biology of *Prochlorococcus* and *Vibrio* at all scales of biological organization, from individual cell design to the dynamics of large populations.

Specifically, we have the following four over-arching aims:

1. **Identify the patterns of genome diversity within and among natural *Prochlorococcus* and *Vibrio* populations, and relate these patterns to the cellular metabolism, population genetics, and the ecology of these groups;**

2. **Characterize the design of the cellular machineries of *Prochlorococcus* and *Vibrio*;**

3. **Characterize the role of phage and horizontal gene transfer in shaping genome diversity, population structure, and metabolism in *Prochlorococcus* and *Vibrio*;**

4. **Harnessing natural genetic diversity and processes for strain engineering.**

Aim S6.1: Identify the patterns of genome diversity within and among natural *Prochlorococcus* and *Vibrio* populations, and relate these patterns to cell metabolism, population properties, and the ecology of these groups

Our first goal is to explore the patterns (genes and genetic elements) of diversity among genomes within a phylogenetic framework (*i.e.*, from very closely to distantly related groups of genomes) and in an explicit environmental context (*i.e.*, micro-scale to global). Recent research in *Prochlorococcus* and *Vibrio* has shown that genomes co-existing within an environmental compartment can be diverse (Coleman et al., 2006; Thompson et al., 2005). It appears that they contain a core genome shared by all individuals, but that there are also significant numbers of genomic islands containing genes that occur at low frequency within the population (Appendix 1). Both *Prochlorococcus* and *Vibrio* are excellent models in which to study population-wide diversity since considerable information on environmental distribution and genomic diversity already exists.

Sub-aim S6,1.1: The peripheral genes in *Prochlorococcus*: Patterns, function and ecological correlates of genome diversity from cultured isolates and wild populations

Project summary: To understand how novel genes, together with the core genome, interact to produce a wide variety of physiological variants, we propose a combination of whole genome sequencing, physiological characterization, functional genomics, and metabolic modeling using our extensive culture collection of diverse *Prochlorococcus* strains. In addition we will assess the patterns of genomic diversity in wild *Prochlorococcus* populations, to understand how variability in the peripheral genome correlates to environmental factors.

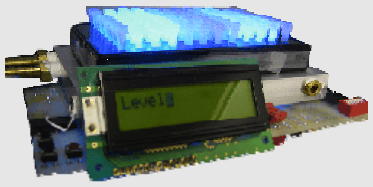
Proposed deliverables: Novel gene interactions will be investigated through analyses of the co-occurrence, conserved operon structure, and phylogenetic distribution of peripheral genes in our cultured isolates. We propose to characterize the fundamental 'niche space' of each strain using our custom-designed high-throughput culturing system, which allows us to grow cells along gradients of light, temperature, and nutrients all in the small-scale format of 96-well plates (Box 3). To determine which types of genome variation contribute most to physiological differences (e.g. gene content in genomic islands vs. regulatory sequence changes), we will expand our recent work using Affymetrix microarrays for two *Prochlorococcus* strains using NimbleGen custom microarrays for an expanded list of strains. We will compare the global transcriptional responses of several strains across the phylogenetic spectrum under standard conditions of nutrient starvation, different light levels, and phage infection. This work will likely reveal genes or genome features important for adaptation, suggesting testable hypotheses about their distribution in the oceans.

Box 3. Measuring cell fitness in multi-parameter space

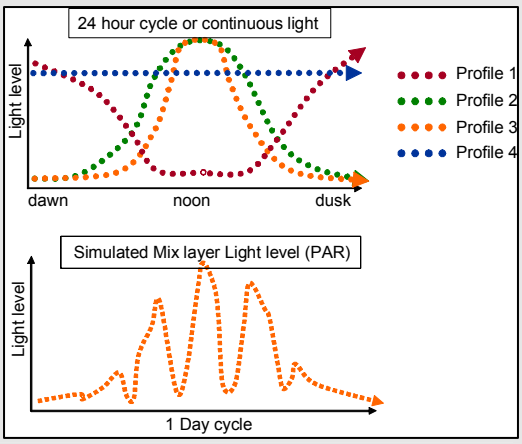
Hutchinson defined the fundamental niche of an organism as an n-dimensional hypervolume, every point in which a species can survive and reproduce (Hutchinson, 1957). The dimensions of the hypervolume are physical and chemical parameters. The niche concept is very useful for ordering our knowledge about microbes, particularly in the genomic era: The information encoded in the genome of an organism defines the n-dimensional "space" (figurative space) within which a cell lineage can persist. Defining this space is the foundation for the biology of a cell.

To begin to define the fundamental niche of different *Prochlorococcus* strains we have developed a high-throughput culturing system, based on a 96-well plate foundation, that allows us to grow cells along gradients of light, temperature and nutrients (see below). Nutrient gradients (e.g., P and N limitation) can be superimposed on the light and temperature limitation by adjusting the ratio of the elements in the media so one is limiting, and running the system as a semi-continuous culture, transferring daily. The mini-cultures will reach a quasi steady-state, growing at a rate that is equal to the supply rate of the limiting nutrient.

Custom designed mini-illuminator system for the study of relative fitness of *Prochlorococcus* as a function of light, temperature, and nutrients.



Mini-Illuminator System. LEDs are used as a light source in 96-well plates, and the system is cooled with a water bath, allowing simultaneous gradients in light and temperature. Relative growth rate of the cells is measured by measuring bulk chlorophyll fluorescence with a 96-well fluorometer. (Design by Hilton, Drakare, and Coe, Chisholm Lab).



Light intensity in the wells is easily programmed using an excel spread sheet to represent the diel solar cycle (top), or any oscillation frequency to simulate the light experience of a cell in the mixed layer of the ocean, or in a heavily stirred bio-reactor (bottom). low light, oxygen minimum zones, and unusual nutrient regimes. If indeed the peripheral genome is involved in local adaptation, then we expect to find novel genes in these environments enabling physiological diversity. We will culture those strains in the laboratory and characterize their physiology. We propose to sequence the genomes from 20 of these new isolates to understand the mechanisms by which *Prochlorococcus* thrives throughout such a wide geographic and depth range in the oceans.

Whole-community metagenomics approaches have begun to reveal the coexisting diversity within an entire microbial community (DeLong et al., 2006; Venter et al., 2004). To understand the within- and between- population diversity of *Prochlorococcus*, we propose an analogous population genomics approach. We will use our Cytopeia InFlux high-speed cell sorter to separate *Prochlorococcus* from the rest of the microbial community, and then sequence metagenomic libraries from these sorted populations. This approach allows us to examine the distribution of peripheral genes (e.g. those found in genomic islands) which might not be identifiably "*Prochlorococcus*" in whole-community libraries while obtaining a greater coverage of *Prochlorococcus* than what we can obtain from whole-community libraries. We are currently pioneering this approach using samples from the Hawaii Ocean Time Series -a relatively stable stratified environment- and the Bermuda Atlantic Time-series (BATS) -a strongly seasonal dynamic environment (in collaboration with Ed Delong's group). Here we propose to follow the population genomic structure of *Prochlorococcus* at BATS, at two depths, over a seasonal cycle (two more time points in July and January (200Mb total), to add to our existing October sample). We will examine patterns both in the seasaonal succession of different phylogenetic clusters, and in the abundance of different peripheral genes. From these data, combined with time-series physicochemical parameters from the BATS program, we can begin to understand whether peripheral genes are largely selectionally neutral or adaptive in *Prochlorococcus*, and we can better delineate the between- and within-population diversity of this organism.

To complement the metagenomic approach, we will use whole genome amplification of individual sorted cells (Zhang et al., 2006) to sequence ten isolates from the surface waters of the Pacific and ten from the Atlantic. This approach reveals the entire peripheral gene content of a single cell, and can thus inform our knowledge of gene interactions. These 20 genomes will be crucial for describing within-population and between-population diversity. We will also use these genomes to look for signatures of HGT in genomic islands (Coleman et al., 2006). Our goal is to understand the mechanisms for genetic exchange and the source of peripheral genes. As discussed in Aim 3, we propose to sequence the vectors (phage) for HGT as well, enabling a more complete picture of genome dynamics in this system.

Finally, to explore the outer bounds of the peripheral genome, we propose a directed isolation approach to obtain new isolates of *Prochlorococcus* from diverse environments, such as extremely such as extremely low light, oxygen minimum zones, and unusual nutrient regimes. If indeed the peripheral genome is involved in local adaptation, then we expect to find novel genes in these environments enabling physiological diversity. We will culture those strains in the laboratory and characterize their physiology. We propose to sequence the genomes from 20 of these new isolates to understand the mechanisms by which *Prochlorococcus* thrives throughout such a wide geographic and depth range in the oceans.

Sub-aim S6.1.2: Fine-tuning of core genes in *Prochlorococcus*: Adaptation along environmental gradients

Project summary: Adaptation occurs not only through gene loss and gain, but also through fine-tuning of protein sequences and regulation in the core genome. Few studies have been able to detect such tuning (e.g. (Bielawski et al., 2004)) because it requires a large number of sequences and knowledge of relevant environmental parameters. Given our extensive culture collection, the abundance of *Prochlorococcus* genes in metagenomic databases (DeLong et al., 2006; Venter et al., 2004), and the increasing availability of environmental sequences from a variety of oceanic regimes, we are presented with an unparalleled opportunity to detect fine-tuning of protein sequences along environmental gradients in the oceans.

We will focus on key genes in photosynthesis and nutrient uptake, two processes that have evolved towards extreme efficiency in *Prochlorococcus* and which likely vary along gradients of light, temperature, and nutrient concentration in the oceans. As described in our detailed progress to date section (Appendix 1), we have shown that two *Prochlorococcus* HL-adapted ecotypes have strikingly different distributions in the oceans, despite sharing over 99% 16S rRNA identity, and that these distributions may be driven in

part by temperature adaptations. This adaptation is likely encoded by small sequence changes in core genes. Recent biochemical work has demonstrated that certain sequence changes in PSII reaction center proteins affect the temperature kinetics of photosynthesis (Shlyk-Kerner et al., 2006). We predict that such patterns will be evident in *Prochlorococcus* sequences and may help explain the environmental distributions of different ecotypes.

Phosphorus availability is thought to be an important driver of genome evolution in *Prochlorococcus* and we have suggested that gene content varies depending on phosphorus availability (Martiny et al., 2006). Our goal is to extend this paradigm further, from gene presence/absence to gene sequence. We predict that sequences of key genes such as *pstS*, the periplasmic binding protein for phosphate, and *phoE*, the outer membrane porin for phosphate, will show signatures of optimization to different phosphate regimes in the oceans. Eventually, given a sequence of unknown origin, we may be able to predict some features of the environment from which it came.

Proposed deliverables: Using whole genome sequences and metagenomic datasets, we will reconstruct the phylogenetic history of key metabolic genes (and putatively neutral loci for comparison) and test whether sequences cluster by environmental parameters. Signatures of environmental adaptation may only be evident in select residues of each protein, and thus we will use branch-site models implemented in PAML (Yang, 1997) to test whether a subset of residues are under positive selection, and correlate sequence variation in these residues to environmental parameters. Finally, these molecular evolution results will guide experiments that test whether sequence variants have different kinetics or binding efficiencies that could contribute to local adaptation. Elucidating structure-activity relationships is a crucial step in understanding how cells adapt their core metabolic processes to their environment, and will enable us to design sequence variants with desirable properties. The *Prochlorococcus* system is uniquely suited to this challenge due to the sheer number of wild sequences available, and the extensive knowledge of environmental parameters such as light, temperature, and nutrient concentrations in the oceans.

Sub-aim S6.1.3: Estimate structure-function relationships and effects of biogeography in microdiverse *Vibrio* genomes

Project summary: Efficient large-scale bioenergy and biotechnology applications will likely use genetically diverse communities acting in relatively heterogeneous and fluctuating environments. To better understand the role of diversity in ensuring predictable function at the population level, we will conduct a model study that aims at determining (i) what type of genomic variation naturally co-exists under a specific set of temporal or spatial environmental conditions, (ii) what are relative rates of different modes of genome differentiation, and how does population structure influence these rates (e.g., is HGT significantly higher among co-occurring strains), and (iii) how does selection arising in different environments map onto specific cellular sub-systems?

We propose to explore the fine scale genome differentiation and distribution of *Vibrio* strains in the water column as a realistic yet tractable model. In our preliminary work, we have already shown that groups of closely related *Vibrio* strains are differentially distributed in the water column on microscales, *i.e.*, they either are predominantly free-living, attached to particles or associated with zooplankton (Fig. 3). Moreover, for some groups of strains these environmental preferences appear to change over the season indicating niche flexibility.

Proposed deliverables: We will build on our already existing extensive library of ~2,000 strains, which are of known environmental origin and have been ordered into genetic clusters by sequencing of housekeeping genes (see Appendix 1). Among those, we will select ~100 genomes for sequencing (see below) phenotypic characterization (metabolic diversity) to represent groups, which display different environmental preference and a range of phylogenetic relationships. Moreover, gene expression analysis done in Aim 2.3 will aid in the interpretation of phenotypic differences among genomes. Our goal is to quantify relative rates of recombination, gene loss and import of foreign DNA among genomes as well as

genome regions whose evolution suggests the action of natural selection. Importantly, whole-genome data allow independent characterization of population demographic structure, which can otherwise confound the statistical analysis of selection.

Our genome sequencing strategy is to pick 10 genomes for full sequencing and ~85 additional, closely related genomes for polony re-sequencing to identify SNP variation. We note that we also have funding from the Moore foundation to sequence several *Vibrio* genomes and will be able to sequence an additional 5 of the water column isolates to give us a number of diverse genomes, which can be used as templates for re-sequencing. Further, we will use high-throughput hybridization, individual gene-targeted PCR and sequencing to investigate the distribution of specific loci that the genome analysis suggest to be associated with ecological clusters from the broader collection of environment specific isolates. We also note that we have recently submitted a proposal to NSF as a specific outgrowth of the current GTL funding to expand the search for ecologically differentiated *Vibrio* clusters to other environmental compartments (e.g., animal guts, surfaces) so that numerous synergies may arise.

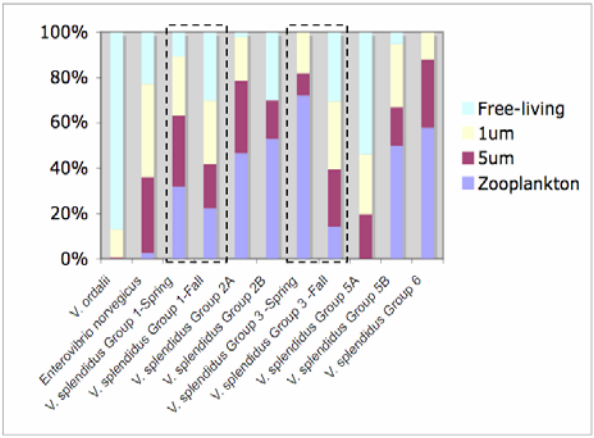


Figure 3. Example of distribution differences of *Vibrio* among free-living, particle and zooplankton associated compartments of the water column community. Water samples were successively partitioned into > 63 > 5 > 1 > 0.22 µm fractions, respectively (the 1 µm fraction is considered ambiguous since it can consist of cells attached to very small particles or large cells). Identification of strains was by sequencing of Hsp60 genes. Only clusters with significant associations (Fisher's Exact Test) are shown. The results demonstrate a trend toward genotypic partitioning into specific compartments on a fine-scale, though some fraction are also present in other compartments. This may be due to the high potential of gene flow either by HGT or migration, which cannot currently be discerned due to the sequencing of a single protein coding gene. Note the shifting preferences of different *V. splendidus* clusters and varying shifting preference from spring to fall within two clusters (dashed boxes).

Using these ~100 genome sequences, we will focus on genome-wide analysis of positive selection, generating a picture of the selective forces acting across all loci during ecological specialization across the comparative framework of closely related genomes. Because few such studies have been performed in bacteria, it is not clear which methods are best suited (or even applicable) to bacterial genomes. Thus, a key deliverable will be a benchmark of complementary approaches to identifying selection (relative evolutionary rates, dN/dS rates, McDonald-Kreitman tests, F_{ST}, *etc.*) (McDonald and Kreitman, 1991; Sarich and Wilson, 1967). Additional tests based on allele frequency (Fay and Wu, 2000; Tajima, 1989), may only be applicable if recombination rates within our sequenced populations are very high relative to the rate of 'clonal expansion' of beneficial mutations, thus our data set may help to resolve some of the basic features of bacterial population genetics.

We will annotate the genomes using the MicobesOnline pipeline (Alm et al., 2005) and use this information to look for concerted selective effects acting over functionally related, but physically unlinked, gene loci, such as an excess of positive selection on genes within the same metabolic pathway. These data will provide a context for our extensive analysis of metabolic phenotypes and statistical tests of environmental association of different groups of strains. Specifically, we can discriminate genes that are under unusual selective pressures in one particular lineage (environment-specific adaptations) from those under strong selective pressures across all strains/species (*e.g.*, genes involved in phage invasion). As a

preliminary step in this direction, we have developed a 'relative-rates' test that is normalized to both a species-specific molecular clock, and a gene-family specific evolutionary rate, and applied it to a set of 744 core genes across a set of 30 gamma-proteobacterial genomes. Using this approach, we have identified a number of pathways under unusual selective pressures in different lineages such as: the flagellar subsystem in enterobacteria, sulfur metabolism in *Buchnera* species, and sugar and amino acid metabolism in *Idiomarina loihiensis*. This combination of evolutionary inference with systems-biological analysis represents a powerful step toward harnessing the full potential of comparative genomics in an ecological context.

Aim S6.2: Characterize the design of the cellular machinery of *Prochlorococcus* and *Vibrio*

As the smallest – in terms of both cell and genome sizes – oxygenic photoautotroph, *Prochlorococcus* is a unique case study in cellular design, and one that might provide insights highly relevant to bioenergy engineering. *Vibrio*s make an excellent heterotrophic model system for studying the design of cellular networks in part because of their similarity to the well-characterized *Escherichia coli*, and also because of their potential biotechnological role in energy conversion. Our goal is to understand the global functioning of these model autotrophic and heterotrophic cells by deciphering gene regulatory networks, describing gene expression under various growth conditions, quantitatively measuring levels of all proteins, and building comparative models of *Prochlorococcus* strains. Central to this is uncovering how the cellular complement of gene products is regulated in response to environmental drivers, particularly the diel cycle, which the cells are continually subjected to in the surface ocean. Overall, a detailed and quantitative description of cellular design will contribute to understanding of the prominence of *Vibrio* and *Prochlorococcus* in ocean ecosystems, their susceptibility to changing environmental conditions and their lessons for photobiological and metabolic engineering.

Sub-aim S6.2.1: The Core Genome of *Prochlorococcus*: Is this the minimal oxygenic photoautotroph?

Project Summary: The core genome shared by all *Prochlorococcus* sequenced to date contains approximately 1200 genes (Appendix 1). If these genes constitute a viable cellular metabolism, this core genome may represent the minimal chassis for building a photosynthetic machine. To explore this concept we have constructed metabolic models for the 12 sequenced *Prochlorococcus* strains, and for the core genome shared among them using Pathway Tools (Karp et al., 2002). This first step shows that relatively few well-characterized metabolic pathways differ between the strains. Pathways that differ include nutrient degradation and salvage pathways, along with pathways involved in biosynthesis of cell surface components. This model, however, is incomplete and limited to well-characterized pathways. A more complete analysis based on shared metabolites will be used to answer the question: does this evolutionarily conserved core genome form a blueprint for minimal photosynthetic machines, and if so what added nutrients and additional transporters will be required for proper function?

Proposed Deliverables: To answer this question, we propose to build flux balanced models of metabolism for each of the 12 strains as well as the core genome of *Prochlorococcus*. Building a model of each strain from its sequenced genome can be done in a semi-automated fashion (Segre et al., 2003) this project will benefit from additional development in the Lin, Segre, and Church labs, such as detailed modeling of photosynthesis, which is being proposed in the concurrently submitted VIBRANT DOE Bioenergy Center proposal. This analysis will yield "missing" reactions, whose presence will be tested biochemically in whole-cell lysates using off-the-shelf biochemical methods. We will also grow cultures in parallel controlled environments to better understand the effect of factors such as light cycles and nutrient concentrations on different strains.

Sub-aim S6.2.2: Quantitative analysis of the *Prochlorococcus* proteome

Project Summary: The well-characterized, streamlined genome and minimal resource requirements of *Prochlorococcus* simplify quantitative proteomic analyses, and such measurements will be very valuable for understanding the organism's physiology and ecology. Quantitation of the protein machinery underlying key cellular processes will be essential to elucidate molecular mechanisms of environmental adaptation.

In the oceans, *Prochlorococcus* has had to adapt its cellular processes to function in this daily rhythm. Capitalizing on our study of the diel cycling of the *Prochlorococcus* transcriptome (Appendix 1), we will analyze the temporal patterns of mRNA-level and protein-level gene expression to better understand the relationship between the two for aid in cellular modeling. Two particular patterns of transcriptional/translational contrast might be anticipated: First, are there gene products that show significant time lags between transcription and translation? Second, are there products that show a diel rhythm of transcription, but a relatively constant protein inventory?

Proposed Deliverables: We propose to develop methods for the global quantitative analysis of the *Prochlorococcus* proteome. These will include isotope-labeling strategies for both relative and absolute (i.e., copies-per-cell) quantification. Protein-level quantification will extend our understanding of gene expression to the level where metabolism is affected and where *Prochlorococcus* interacts with its environment. We will determine the cellular complement of nutrient transport systems, which will elucidate the molecular basis of these cells' adaptation to oligotrophy. The diel protein measurements will, in conjunction with microarray transcript data, constitute the first full, time-resolved picture of gene expression over an environmentally driven cell cycle.

Sub-aim S6.2.3: The design of the *Prochlorococcus* regulatory system

Project Summary: *Prochlorococcus* genomes contain relatively few genes and a very low number of transcription regulators (only 28 transcription factors are predicted in *Prochlorococcus* MED4, an order of magnitude lower than most bacteria). As the most efficient photosynthesis machine (Box 1), this simple system represents an attractive chassis for synthetic biology and bioenergy applications. Here we propose to describe the complete regulatory network of *Prochlorococcus*, toward the goals of generating testable hypothesis about the evolution and physiology of each strain, as well as engineering new cellular processes in this minimal chassis.

Proposed Deliverables: We will use a combination of genomics, molecular biology and bioinformatics approaches to decipher the molecular mechanisms regulating the expression of all genes in *Prochlorococcus* MED4. We will adapt 5'- and 3'-RACE (Rapid Amplification of cDNA End) methods (Chenchik et al., 1996) in order to identify the start and end points of each mRNA produced by *Prochlorococcus* MED4 under a diel (light-dark) cycle by taking advantage of the polony sequencing technology developed in the Church laboratory. The procedure is akin to SAGE and Ditag techniques (Harbers and Carninci, 2005; Ng et al., 2006), and the construction of the initial sequencing-libraries is underway. We anticipate that we will be able to describe most, if not all, *Prochlorococcus* MED4 mRNA using this approach, including novel small regulatory RNA. Since transcription factor binding sites are typically located in close proximity to transcription initiation sites, we expect this analysis to greatly contribute to the identification of important DNA regulatory motifs. In order to associate these motifs with the relevant proteins that recongnize them, bacterial one-hybrid assays (Meng et al., 2005) and/or SELEX (Liu and Stormo, 2005) will be used to determine the DNA binding specificities of all 28 *Prochlorococcus* MED4 transcription regulators. Each transcription regulator has already been cloned in the appropriate vectors and library screening will begin shortly. Genome-wide chromatin immunoprecipitation assays (ChIP-chip) will also be performed for several transcription factors. Many custom polyclonal antibodies are currently being raised in order to proceed to the immunoprecipitation under native conditions. Sebastien Rodrigue, a postdoctoral associate in the Chisholm laboratory, has extensive experience with ChIP-chip assays and will be conducting these experiments (Rodrigue et al., 2007). The information from transcription units and regulatory protein binding motifs will be combined into computational models to predict the most probable role of every transcription regulator on the expression of each gene. As part of

the concurrently submitted VIBRANT DOE Bioenergy Center proposal, the data obtained from these experiments will also be used in conjunction with numerous *Prochlorococcus* gene expression profiles (over 200 are already available in the Chisholm lab) to refine gene expression modelling in *Prochlorococcus* through a collaborative effort with Jim Collins in the Bioengineering Department at Boston University (Faith et al., 2007). Since the protein domains involved in DNA binding are very well conserved between *Prochlorococcus* strains (Rodrigue, unpublished data), we will take advantage of the information obtained from the MED4 strain to predict regulatory network structures in all 12 sequenced *Prochlorococcus* strains in order to generate testable hypotheses about the adaptation of each strain to particular growth conditions. We are also interested in the respective roles of genome content and gene expression regulation in the evolution of closely-related *Prochlorococcus* strains (see also Aim 1). The information on regulatory binding sites and promoters will further constitute a set of new "parts" for synthetic biology applications.

Sub-aim S6.2.4: Identify constraints on the evolution of *Vibrio* cellular networks

Project Summary: A key challenge in bacterial genetics is to understand the extent to which the peripheral genome (genes specific to strain or group of strains) and gene expression differences among closely related genomes are relevant for population-level processes in nature. Thus, we will leverage the genomic and ecological data from Aim 1.3 to address fundamental population genomic questions: (i) to what extent and under what conditions is the 'peripheral' genome neutral or adaptive; (ii) is the fitness contribution of orthologous gene loci similar among related strains, or are there differences due to 'epistatic' interactions among gene loci; (iii) to what extent do changes in regulation contribute to ecological adaptation; and (iv) how much (presumably neutral) variation in regulatory strategies exists within populations?

While much of the genomic variation among ecologically specialized populations occurs in the peripheral genome, the function of these strain-specific genes remains poorly annotated when compared to the core genome. We will use high-throughput deletion mutagenesis techniques developed in the Church laboratory to map the contribution of these genes to fitness across a compendium of laboratory-controlled environments. In addition, gene expression is thought to be a major contributor to phenotypic variation both between closely-related species (Tirosh et al., 2006) and also within species (Brem et al., 2002; Townsend et al., 2003). On very short time-scales (100's to 1000's of generations of laboratory evolution) changes in gene regulation are a major contributor to adaptive genome evolution (Elena and Lenski, 2003). Indeed, this basic principle of optimizing global expression patterns has been successfully adopted as a strategy for strain engineering of industrial microorganisms (Alper and Stephanopoulos, 2007).

Proposed Deliverables: Natural populations of *Vibrio* differ widely in their gene content, as evidenced by megabase size differences between even closely related strains (Thompson et al., 2005), implying a significant role for strain-specific genes in environmental adaptation. The role of these genes, however, and, in particular, their relative contribution to organismal fitness under different conditions is poorly understood. Indeed, the relative contribution to fitness (even in laboratory-controlled environments) is unknown for nearly all genes, despite the fact that such data would be of enormous value from population genetics, ecological and engineering points of view. We will begin to characterize the cellular roles of these strain specific genes using transposon-based insertional mutagenesis. The Church laboratory and others have developed transposon library screening technology that makes use of an embedded T7 promoter to quantify the relative fitness of insertional mutants in parallel using DNA microarrays (Badarinarayana et al., 2001; Chan et al., 2005; Sassetti et al., 2003; Winterberg et al., 2005). Because expensive bar-coding and sequencing is avoided, libraries can be constructed and screened for multiple strains using NimbleGen strain-specific arrays, which were recently applied to transposon library screening in *E. coli* K12 (Winterberg et al., 2005).

We will test our reference strain, *V. splendidus* 12B01, under a wide range of conditions including different growth media, stress conditions (high/low temperature, acid/alkaline stress, high salinity, *etc*.), and where possible, other ecologically relevant factors (co-culture with algae, growth on zooplankton, *etc*.) as knowledge of ecologically important factors emerges from our studies in Aim 1.3. In addition, the

same conditions will be used (see below) to characterize gene expression patterns. For a limited, but phylogenetically diverse, set of additional *Vibrios*, we will repeat these measurements with a smaller compendium of primarily environmental stress conditions, to better understand how the fitness contributions of orthologous loci vary across lineages, and to address question (ii) above.

In addition to fitness data, we will use a combination of experimentation and computational analysis to reconstruct the 'core' *Vibrio* transcriptional regulatory network as well as variation within that network among strains. First, we will use 'polony' sequencing to investigate the genomic sequence of a large number strains (~100, as described in Aim 1.3). Genome sequencing of multiple yeast strains has shown the utility of this approach for identifying cis-regulatory sequences, but this large number of genomes has not been previously used for motif detection, largely because of the cost of traditional sequencing methodologies. In addition, we will make use of large databases of manually curated bacterial protein-DNA interactions culled from the scientific literature (RegulonDB, RegTransBase, etc.)(Salgado et al., 2006), especially those focusing on the model organism most closely related to the *Vibrio* genus, *Escherichia coli*. Both of the databases mentioned above are currently integrated into the MicrobesOnline database (Alm et al., 2005). In addition, we have considerable experience with manual reconstruction of regulatory networks, and have been actively developing new tools for manual motif detection that will soon be incorporated into the MicrobesOnline website.

In addition to the protein-DNA interactions predicted from genome sequence and literature on related organisms, we will probe the *Vibrio* regulatory network using gene expression microarrays. We will test vibrios under a compendium of environmental conditions (see above). This compendium will provide the raw data for a suite of network inference tools (cMonkey, Inferelator) that has been successfully demonstrated on *de novo* network inference in the archaeon *Halobacterium* sp. NRC-1 (Bonneau et al., 2006). Interactions identified from this step will be compared to those from literature-culled databases, manual inspection, and genome sequence analysis to produce a set of high-confidence regulatory motifs with multiple sources of evidence. These motifs will serve as the backbone of the *Vibrio* regulatory network. With a reference regulatory network in hand, we will estimate ecologically relevant variation by conducting a more limited study of expression patterns focusing on a more limited compendium of stress conditions, but extending our analysis to additional strains (~10) that span a range of phylogenetic distances from the previously sequenced reference strain, *V. splendidus*, 12B01. In particular, we will screen for regulatory components/strategies specific to ecologically-informative phylogenetic clusters, and survey the level of diversity in regulatory strategies co-existing within populations addressing questions (iii) and (iv) above.

Aim S6.3: Characterize the role of phage and horizontal gene transfer in shaping genome diversity, population structure, and metabolism in *Prochlorococcus* and *Vibrio*

A central goal in synthetic biology is the creation of new or optimized (metabolic) functions. In natural bacterial communities, this is largely achieved by horizontal gene transfer (HGT), including both homologous and illegitimate recombination. However, HGT also has the power to blur boundaries among populations because recombination is decoupled from sexual reproduction and can happen over wide phylogenetic distances. Thus it remains unclear how ecological specialization can stably differentiate co-existing genomes. For applications employing complex open systems, it is important to quantitatively understand the rates and bounds of HGT under different constraints.

Our goal is to explore the modes and mechanisms of HGT in a quantitative population genetic framework. *Prochlorococcus* and *Vibrio* are excellent model systems since they are subject to very different environmental constraints (e.g., single cell growth in plankton vs. biofilm formation), which may lead to predominance of different HGT mechanisms. While the *Vibrio*s possess all documented modes of HGT (e.g., plasmids, phages, integrons) and multiple different lifestyles, *Prochlorococcus* is highly optimized and may only have phage as an efficient HGT mechanism. The comparison of both types of organisms

will be enabled by well-characterized strain collections, metagenomic information and already existing phage collections.

Sub-aim S6.3.1: Identify shared and unshared genomic features of phage that infect *Prochlorococcus*, and develop a population genomic framework for phage-host interactions

Project Summary: The first three genomes of cyanophage infecting *Prochlorococcus* revealed novel gene combinations and gave hints of a dynamic phage-host coevolutionary process in the oceans (Lindell et al., 2004; Sullivan et al., 2005; Sullivan et al., 2006). Now, with further genome sequencing and ultrastructural studies, our goal is to characterize the bounds of cyanophage 'genome space' and define the major "types" – both morphological and genomic – that infect *Prochlorococcus*. Furthermore, by assessing phage diversity in different environments and comparing it with host diversity (Aim 1), we hope to develop a population genomic framework for understanding phage-host interactions.

Proposed Deliverables: We will describe *Prochlorococcus* phage genome variability through a combination of culture-independent metagenomics surveys targeting natural viral communities, in conjunction with analyses of culture-based cyanophage isolates. The naturally occurring oceanic cyanophage communities will be examined using end-sequencing from large (~40kb) and sequencing of small (~2-4kb) insert clone libraries made from DNA extracted from 3 photic zone depths at each the Bermuda Atlantic (BATS) and Hawaii Ocean (HOT) Time Series sites. These samples have been collected and are in the process of sequencing under DOE funding. End-sequences from these environmental samples will be used to determine which DNA fragments are "cyanophage" (*sensu* (DeLong et al., 2006)) for further analyses. For example, a selected subset of large-insert DNA fragments will be completely sequenced to examine genome variability, while all end-sequences will be used to estimate the number of "types" in these natural communities using the PHACCS system (Angly et al., 2005). Further, given that cyanophage genomes carry a number of host metabolic genes, including genes involved in photosynthesis and phosphate scavenging (Sullivan et al., 2005; Sullivan et al., 2006), we expect that gene content in cyanophage genomes will co-vary with host diversity along these depth gradients (parallel cell-fraction libraries from both BATS and HOT are also being sequenced through DOE funding), perhaps even informing our understanding of the host's physiological state.

To complement these environmental metagenomics analyses of phage genome fragments, extensive culture-based complete genome sequencing and analyses will inform our understanding of the genomic repertoire of *Prochlorococcus* cyanophages. We propose here to focus on deep isolations from a single depth of the HOT environmental samples described above using only the LL *Prochlorococcus* strain NATL2A, chosen because it is known to be infected by diverse phage morphologies (Sullivan et al., 2003). These data will be used to evaluate whether phage isolated on a particular host strain from a single water sample represent quasi-clonal progeny from recent lytic events, or whether they represent a diverse set of phage types; these two extremes have profound implications for predicting phage infection behavior in a natural or engineered setting.

Sub-aim S6.3.2 Properties and role of 'host' metabolic genes in phage genomes

Project Summary: As discussed in Appendix 1, cyanophage encode genes for host metabolic processes, including phosphate acquisition (*pstS*), photosynthesis (*psbA*, *psbD*, and *hli*), and carbon conversion (*talC*). Here we propose intensive studies of two genes, *pstS* and *talC*, to understand the interaction of phage and host metabolism and the evolution of host metabolic processes.

Each of three *Prochlorococcus* phage genomes and one of four Synechococcus phage genomes contains a *talC*-like gene, encoding an aldolase family protein (Mann et al., 2005; Sullivan et al., 2005), yet no *talC* ortholog is found in any other known viral genome. Preliminary biochemical evidence suggests that the *talC* gene in cyanophage encodes a functional transaldolase (L. Thompson, unpubl.).

Transaldolase transfers a three-carbon dihydroxyacetone moiety from fructose 6-phosphate to erythrose 4-phosphate (reversibly) as part of the pentose phosphate pathway (Wood, 1985). Although 'classical' transaldolases are ubiquitous in *Prochlorococcus* and *Synechococcus*, only one strain of *Prochlorococcus* has a *talC*-like gene (Dufresne et al., 2003; Palenik et al., 2003; Rocap et al., 2003). Gene expression analysis of cyanophage P-SSP7 infection of *Prochlorococcus* strain MED4 (see Appendix 1) indicates that phage *talC* mRNA is made during infection (Lindell et al., 2007).

We hypothesize that cyanophage *talC* encodes a transaldolase, that the transaldolase reaction is a limiting step in the pentose phosphate pathway during infection, and that phage-encoded transaldolase helps relieve this bottleneck, mobilizing stored carbon to produce ribose and NADPH, thus providing biosynthetic material (ribose) and reducing equivalents (NADPH) to help increase phage production and therefore phage fitness. We propose to investigate the respective roles of host and phage transaldolases by examining their kinetic properties *in vitro*. Importantly, as an enzyme in the pentose phosphate pathway, transaldolase could be important to biological means of hydrogen production. Products of the pentose phosphate pathway can be directed to bacterial or archaeal hydrogenase enzymes to produce hydrogen (Woodward et al., 2000), a source of energy that does not produce carbon dioxide as a waste product.

Cyanophage genomes also encode PstS, the periplasmic binding protein for phosphate (Sullivan et al., 2005). In phosphate-starved *Prochlorococcus* cells, *pstS* is strongly upregulated (Martiny et al., 2006). We hypothesize that the phage-encoded *pstS* helps cells acquire phosphate, which is needed for phage genome replication, during infection. This gene is sometimes found in multiple copies in host genomes (Martiny et al., 2006), and we hypothesize that phage are responsible for the horizontal transfer of *pstS*, which is then retained in multiple copies in phosphate-limited environments. We propose to study this host-phage metabolic interaction experimentally, and to trace the evolutionary history of this gene using cultured and wild sequences from both host and phage.

Proposed Deliverables: We have successfully cloned, expressed, and purified host and phage transaldolase (L. Thompson unpubl.). We will measure and compare key kinetic parameters across host and phage versions of transaldolase. Specifically, we plan to look at catalytic efficiency or 'turnover number' (k_{cat}) and substrate affinity (K_M) for the principal substrates. We predict that the phage-encoded transaldolase has robust activity, possibly with a turnover number higher than the host transaldolase. Similarly, we predict that substrate specificity (k_{cat}/K_M) for certain substrates may be higher for the phage version. We also want to see if the enzymes are inhibited by certain sugars, such as arabinose 5-phosphate, as has been observed for *E. coli* transaldolase (Sprenger et al., 1995), which would be manifested as decreased reaction velocity in the presence of inhibitor. We predict that, unlike the host ortholog, the phage transaldolase is not subject to inhibition, since there is little evolutionary advantage to regulate enzyme activity during infection in order to maintain a metabolic steady-state.

To examine the role of *pstS* during infection, we will infect phosphate-replete and phosphate-starved *Prochlorococcus* MED4 cells with a model cyanophage, P-SSM4. We will monitor production of new phage DNA and particles, and whole-genome expression of both host and phage using our Affymetrix microarrays. We will look for patterns in the expression of host phosphate-uptake genes and phage-encoded *pstS*, and look to see how phosphate starvation affects phage production.

We also propose a detailed evolutionary analysis of *pstS* to understand potential gene swapping and recombination between host and phage, and to understand the selective pressure on host and phage *pstS* in different environments. We will extract *pstS* sequences from *Prochlorococcus* and phage genomes (including those proposed for sequencing in Aims 1.1 and 3.1) and metagenomic datasets from a variety of oceanic regimes and use phylogenetic analyses to determine whether phage and host copies form separate clusters, implying a single transfer event, or whether they cluster among each other, implying repeated transfers over time. We will also look for intragenic recombination as we have done for *psbA* (Sullivan et al., 2006). Finally, we will test whether *pstS* clusters correspond to environment, given that the Sargasso Sea is often phosphate-limited and might impose stronger selective constraints on genes involved in phosphate uptake. These in-depth studies of *pstS*, and *talC*, together with our previous

work on photosynthesis genes, will be crucial for understanding the coevolutionary dynamic between host and phage in different environments.

Sub-aim S6.3.3: Develop Phage Genetics for *Prochlorococcus*

Project Summary: As described in Appendix 1, we have made advances in developing genetic systems for *Prochlorococcus* host strains and we continue to improve these systems. Our tool kit would be greatly enhanced if we could develop a way of modifying phage genomes. However, this process is significantly complicated by the fact that all *Prochlorococcus* phages isolated to date appear to be lytic and hence difficult to isolate in a state where they can be manipulated by homologous recombination. We propose to engineer a mechanism by which to create mutations in cyanophage genomes using a well-characterized phage-host system P-SSP7 (approximately 44 kb) and *Prochlorococcus* MED4.

Proposed Deliverables: We will clone the entire P-SSP7 genome into *E. coli* conjugative BAC vector pMBD14 (Martinez et al., 2004), which can carry up to 300kb of DNA stably in *E. coli*, and is also capable of conjugation into other host strains. Once the phage is cloned into this vector, we will test the ability of the recombinant vector to be transferred via conjugation into both MED4 and a strain of *Synechococcus*, and to cause a productive infection. If this conjugation followed by phage gene expression is successful, we can then generate specific mutants of the phage on the vector, such as knockouts of the host genes present on the cyanophage, and test the effects of these knockouts on infection.

Sub-aim S6.3.4: Mechanisms of HGT in *Vibrio* and their relative importance in creating and maintaining genome diversity in natural populations

Project Summary: We will explore the diversity and host range of phage and plasmids, which are known to be important drivers of genome dynamics. To this end, we will isolate and sequence phages and plasmids and characterize their host range using our existing library of 2,000 strains from the water column. Moreover, we will characterize conditions under which transformation will be induced as a potential mechanism for restriction of gene transfer to environmentally co-occurring strains.

Proposed deliverables: We have already developed methods for isolation and testing of cross-infectivity of phages. This has yielded a collection of 20 phages isolated on different 5 different *Vibrio* strains. We will continue this collection taking advantage of the well-characterized *Vibrio* strains from Aim 1. Phage isolates will be isolated on a specific strain, typed and characterized for host range using our extensive strain collection. These properties will be correlated with host characteristics, such as O-antigen diversity and other potential phage receptors (e.g., the OmpK protein). Several representative phages will be sequenced to determine their genome characteristics and carriage of host genes. We will also explore the possibility of developing 'marker' genes to determine (i) the relationships and diversity of phage isolates, and (ii) *Vibrio* phage population-structure directly in the environment by culture-independent methods (PCR-amplification, cloning and sequencing). In particular, we will determine to what extent the phages display genetic structure (clusters), and to what extent that structure corresponds to host population structure (e.g., in MLSA genes or other properties such as O-antigen diversity). Host genomes will also be screened for integrative phages and their host distribution determined by high-throughput hybridization of *Vibrio* strains.

We will also isolate and sequence plasmids to determine the types of genes they carry and their potential role in increasing host fitness. In preliminary work, we have determined that ~30% of our Vibrio strains carry at least one plasmid and that all of these carry *tra* gene signatures, which indicates that they are sexually transferred. We are also interested in integrative plasmids, which can play a role in generation of high frequency of recombination strains. Similar to the phages, we will characterize plasmids (by restriction analysis of all, and sequencing of select representatives) and determine their host range (either by incidence in different strain groups or by experimental mating). We will also search for potential marker genes, which can then be used to assay the presence of specific types of plasmids in strains or environmental samples either by PCR or hybridization.

To test whether transformation is adaptive and triggered by specific environmental conditions, we will assay strains for induction of competence under different environmental conditions. Candidate conditions are biofilm formation on different natural surfaces or other conditions, which trigger high population density (Meibom et al., 2005).

AimS6.4: Harnessing natural genetic diversity and processes for strain engineering

Classical strain improvement techniques represent a powerful methodology for producing commercially useful microorganisms, yet artificial selections using these techniques tend to focus on point mutation, limiting the efficiency with which the fitness landscape for a particular strain can be searched. In natural populations, horizontal gene transfer and recombination among closely related strains serve to increase the pool of genetic variation that can be used for adaptive evolution. In contrast to mutagen-induced variation, which is largely deleterious, they provide access to a pool of variation that has been culled by purifying selection. Indeed, our studies suggest that within closely related (>99% similarity in 16S rRNA) natural populations, thousands of distinct genotypes can co-exist, and that horizontal transfer and recombination among individuals are common (Thompson et al., 2005). Emerging technologies such as gene-trait mapping and genome-shuffling promise to make these alternative evolutionary mechanisms accessible within laboratory selections (Gill et al., 2002; Zhang et al., 2002), opening the door to selections that harness the diversity of entire populations rather than a single strain.

Sub-aim S6.4.1: Strain improvement strategies motivated by natural modes of gene exchange

Project Summary: In order to realize the potential of engineered biological systems, such as bacteria-based energy conversion systems, it will be necessary to isolate or engineer strains that can withstand the environmental extremes that facilitate these chemical processes. These include extremes of temperature, pH, salinity, as well as buildup of a desired (but often toxic at high concentration) end product such as ethanol or alkanes. Using the Vibrios as a model, we will seek to improve the range of environmental regimes that strains can tolerate using a directed evolution strategy that mimics natural adaptive processes.

Proposed Deliverables: We will adapt a series of genetic technologies that mimic natural evolutionary processes (intrinsic mutation, HGT, and homologous recombination) to both improve strains and better understand the roles of these processes in natural systems. To directly compare results from each of these technologies, we will test them against the same selective regime (extremes of temperature, pH, salinity, and possibly other factors such as ability to grow in co-culture with algae). Genomic changes will be determined by polony sequencing, and ultimately can be compared to adaptations detected in Aim 1.3.

To mimic intrinsic mutations, we will employ long-term adaptation (experimental evolution) of our reference strain, *V. splendidus* 12B01 (Elena and Lenski, 2003; Lenski, 1991; Shendure et al., 2005). We will carry out chemostat and batch culture experimental evolution under the selective regimes described above. In a preliminary study, we have evolved bacterial strains growing in chemostats towards doubling times of <4 minutes, well below widely believed minimum doubling time for bacteria of 9 minutes (Eagon, 1962).

As an analog to natural HGT events in the laboratory, we will adapt the parallel gene-trait mapping approach of (Gill et al., 2002) to incorporate DNA from multiple exogenous sources into a host strain; a donor DNA library is used to transform a host strain and the relative fitness of competitively grown transformants is measured by microarray hybridization. We will leverage our collection of sequenced *Vibrio* strains with high-density microarrays, which will allow us to screen exogenous DNA from multiple strains in a single selection experiment.

We will also explore strategies that mimic natural homologous recombination and leverage our existing strain collections. We will adapt genome-shuffling methods based on protoplast fusion, which is based on disrupting the cell wall (or outer membrane), fusion of cytoplasmic material, and regeneration of recombinant genotypes. This has been used in conjunction with rounds of mutagenesis and selection (genome-shuffling), and has improved the efficiency of strain engineering in bacteria by several orders of magnitude (Dai and Copley, 2004; Patnaik et al., 2002; Zhang et al., 2002). Protoplast fusion in *Escherichia coli* has been reported with efficiencies less than but comparable to well-established gram-positive models (Dai et al., 2005); we will further optimize methods for disrupting the outer membrane and adapt protocols to our *Vibrio* strains. We have recently generated and fused protoplasts of *V. splendidus* 12B01, generated a library of mutants with complementary antibiotic resistance markers, and we will begin quantifying the efficiency of genome-shuffling in *Vibrio*s within upcoming weeks. We will use polony sequencing to determine the hybrid genotypes resulting from protoplast fusion. Whereas most previous efforts were directed at mutagenized variants of a single reference strain, we will leverage the natural genetic diversity contained within our *Vibrio* strain collection. Co-existing natural strains contain a greater degree of sequence variation than mutagenized strains from a single ancestor, and the average fitness of these genotypes is much higher than that of heavily mutagenized strains, in which mutant alleles have not been extensively culled by natural selection. In addition, we will be able to directly test whether genes or alleles enriched in natural populations are selected for in laboratory experiments that mimic the conditions in which those populations are found.

References for Aim S6 Proposal

Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L., and Arkin, A.P. (2005). The MicrobesOnline Web site for comparative genomics. Genome Research *15*, 1015-1022.

Alper, H., and Stephanopoulos, G. (2007). Global transcription machinery engineering: A new approach for improving cellular phenotype. Metab Eng.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., and Rohwer, F. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics *6*, 41.

Badarinarayana, V., Estep, P.W., 3rd, Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G.M. (2001). Selection analyses of insertional mutants using subgenic-resolution arrays. Nat Biotechnol *19*, 1060-1065.

Behrenfeld, M.J., and Falkowski, P.G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnol Oceanogr *42*, 1-20.

Bielawski, J.P., Dunn, K.A., Sabehi, G., and Beja, O. (2004). Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. PNAS *101*, 14824-14829.

Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol *7*, R36.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752-755.

Bruyant, F., Babin, M., Genty, B., Prasil, O., Behrenfeld, M.J., Claustre, H., Bricaud, A., Garczareck, L., Holtzendorff, J., Koblizek, M*., et al.* (2005). Diel variations in the photosynthetic parameters of *Prochlorococcus* strain PCC 9511: Combined effects of light and cell cycle. Limnology and Oceanography *50*, 850-863.

Chan, K., Kim, C.C., and Falkow, S. (2005). Microarray-based detection of Salmonella enterica serovar Typhimurium transposon mutants that cannot survive in macrophages and mice. Infect Immun *73*, 5438-5449.

Chenchik, A., Diachenko, L., Moqadam, F., Tarabykin, V., Lukyanov, S., and Siebert, P.D. (1996). Full-length cDNA cloning and determination of mRNA 5' and 3' ends by amplification of adaptor-ligated cDNA. Biotechniques *21*, 526-534.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., DeLong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-1770.

Dai, M., and Copley, S.D. (2004). Genome shuffling improves degradation of the anthropogenic pesticide pentachlorophenol by Sphingobium chlorophenolicum ATCC 39723. Appl Environ Microbiol *70*, 2391-2397.

Dai, M., Ziesman, S., Ratcliffe, T., Gill, R.T., and Copley, S.D. (2005). Visualization of protoplast fusion and quantitation of recombination in fused protoplasts of auxotrophic strains of Escherichia coli. Metab Eng *7*, 45-52.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R.*, et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science *311*, 496-503.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Asmann, I.M., Barbe, V., Dupratt, S., Galperin, M.Y., Koonin, E.V., Le Gall, F.*, et al.* (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proceedings of the National Academy of Sciences, USA *100*, 10020-10025.

Eagon, R.G. (1962). Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. J Bacteriol *83*, 736-737.

Elena, S.F., and Lenski, R.E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet *4*, 457-469.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLoS Biol *5*, e8.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405-1413.

Gill, R.T., Wildt, S., Yang, Y.T., Ziesman, S., and Stephanopoulos, G. (2002). Genome-wide screening for trait conferring genes using DNA microarrays. Proc Natl Acad Sci U S A *99*, 7033-7038.

Goericke, R., and Welschmeyer, N.A. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep-Sea Research (Part 1, Oceanographic Research Papers) *40*, 2283-2294.

Harbers, M., and Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. Nat Methods *2*, 495-502.

Hess, W.R., Rocap, G., Ting, C.S., and Chisholm, S.W. (2001). The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynthesis Research *70*, 53-71.

Hutchinson, G.E. (1957). Concluding remarks. Cold Spring Harbor Symposium of Quantitative Biology *22*, 415-427.

Karp, P.D., Paley, S., and Romero, P. (2002). The Pathway Tools software. Bioinformatics *18*, S225-232.

Lenski, R.E. (1991). Quantifying fitness and gene stability in microorganisms. Biotechnology *15*, 173-192.

Lindell, D., jaffe, J.D., Coleman, M., Axmann, I., Rector, T., Kettler, G., Sullivan, M., Steen, R., Hess, W., Church, G.*, et al.* (2007). Genome-Wide Expression Dynamics of a Marine Virus and its Host Reveal Features of Co-evolution. Nature *In revision.*

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. PNAS *101*, 11013–11018.

Liu, H., Nolla, H.A., and Campbell, L. (1997). *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquatic Microbial Ecology *12*, 39-47.

Liu, J., and Stormo, G.D. (2005). Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. Nucleic Acids Res *33*, e141.

Mann, N.H., Clokie, M.R., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J., Letarov, A., and Krisch, H.M. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. J Bacteriol *187*, 3188-3200.

Martinez, A., Kolvek, S.J., Yip, C.L., Hopke, J., Brown, K.A., MacNeil, I.A., and Osburne, M.S. (2004). Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Appl Environ Microbiol *70*, 2452-2463.

Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. PNAS *103*, 12552-12557.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652-654.

Meibom, K.L., Blokesch, M., Dolganov, N.A., Wu, C.-Y., and Schoolnik, G.K. (2005). Chitin induces natural competence in *Vibrio cholerae*. Science *310*, 1824-1827.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotech *23*, 988-994.

Montecino, V., and Quiroz, D. (2000). Specific primary production and phytoplankton cell size structure in an upwelling area off the coast of Chile (30°S). Aquatic Sciences *62*, 364-380.

Morel, A., Ahn, Y.-H., Partensky, F., Vaulot, D., and Hervè, C. (1993). *Prochlorococcus* and *Synechococcus*: A comparative study of their optical properties in relation to their size and pigmentation. Journal of Marine Research *51*, 617-649.

Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K.*, et al.* (2006). Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res *34*, e84.

Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., McCarren, J.*, et al.* (2003). The genome of a motile marine Synechococcus. Nature *424*, 1037-1042.

Park, M.-O. (2005). New pathway for long-chain *n*-alkane synthesis via 1-alcohol in *Vibrio furnissii* M1. J Bacteriol *187*, 1426-1429.

Park, M.O., Tanabe, M., Hirata, K., and Miyamoto, K. (2001). Isolation and characterization of a bacterium that produces hydrocarbons extracellularly which are equivalent to light oil. Appl Microbiol Biotechnol *56*, 448-452.

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiology and Molecular Biology Reviews *63*, 106-127.

Patnaik, R., Louie, S., Gavrilovic, V., Perry, K., Stemmer, W.P., Ryan, C.M., and del Cardayre, S. (2002). Genome shuffling of Lactobacillus for improved acid tolerance. Nat Biotechnol *20*, 707-712.

Rocap, G., Larimer, F., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., Arellano, A., Coleman, M., Hauser, L., Hess, W.*, et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-1047.

Rodrigue, S., Brodeur, J., Jacques, P.E., Gervais, A.L., Brzezinski, R., and Gaudreau, L. (2007). Identification of Mycobacterial {sigma} Factor Binding Sites by Chromatin Immunoprecipitation Assays. J Bacteriol *189*, 1505-1513.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J.*, et al.* (2006). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res *34*, D394-397.

Sarich, V.M., and Wilson, A.C. (1967). Rates of albumin evolution in primates. Proc Natl Acad Sci U S A *58*, 142-148.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol *48*, 77-84.

Segre, D., Zucker, J., Katz, J., Lin, X., D'Haeseleer, P., Rindone, W.P., Kharchenko, P., Nguyen, D.H., Wright, M.A., and Church, G.M. (2003). From annotated genomes to metabolic flux models and kinetic parameter fitting. OMICS: A Journal of Integrative Biology *7*, 301-316.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728-1732.

Shlyk-Kerner, O., Samish, I., Kaftan, D., Holland, N., Maruthi Sai, P.S., Kless, H., and Scherz, A. (2006). Protein flexibility acclimatizes photosynthetic energy conversion to the ambient temperature. Nature *442*, 827-830.

Sprenger, G.A., Schorken, U., Sprenger, G., and Sahm, H. (1995). Transaldolase B of *Escherichia coli* K-12: cloning of its gene, talB, and characterization of the enzyme from recombinant strains. J Bacteriol *177*, 5930-5936.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005). Three *Prochlorococcus c*yanophage genomes: Signature features and ecological interpretations. PLOS Biology *3*, e144.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLOS Biology *4*, e234.

Sullivan, M.B., Waterbury, J., and Chisholm, S.W. (2003). Cyanophages infecting the oceanic cyanobacterium, *Prochlorococcus*. Nature *424*, 1047-1051.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., Sarma-Rupavtarm, R., Distel, D.L., and Polz, M.F. (2005). Genotypic diversity within a natural coastal bacterioplankton population. Science *307*, 1311-1313.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. Nat Genet *38*, 830-834.

Townsend, J.P., Cavalieri, D., and Hartl, D.L. (2003). Population genetic variation in genome-wide gene expression. Mol Biol Evol *20*, 955-963.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W.*, et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66-74.

Winterberg, K.M., Luecke, J., Bruegl, A.S., and Reznikoff, W.S. (2005). Phenotypic screening of Escherichia coli K-12 Tn5 insertion libraries, using whole-genome oligonucleotide microarrays. Appl Environ Microbiol *71*, 451-459.

Wood, A.M. (1985). Adaptation of photosynthetic apparatus of marine ultraphytoplankton to natural light fields. Nature *316*, 253-255.

Woodward, J., Orr, M., Cordray, K., and Greenbaum, E. (2000). Enzymatic production of biohydrogen. Nature *405*, 1014-1015.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*, 555-556.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006). Sequencing genomes from single cells by polymerase cloning. Nat Biotech *24*, 680-686.

Zhang, Y.X., Perry, K., Vinci, V.A., Powell, K., Stemmer, W.P., and del Cardayre, S.B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature *415*, 644-646.

S6: PROGRESS FOR 2003-2006 & EMERGING QUESTIONS

PROCHLOROCOCCUS – Chisholm laboratory

Under the auspices of the Harvard-MIT Genomics:GTL Center, the Chisholm lab has made significant progress in understanding the systems biology of *Prochlorococcus* — from the genome to the ocean basin scale — through a combination of laboratory and field studies. We summarize our progress below, in understanding (i) genomic diversity and adaptation, (ii) cellular design and machinery of *Prochlorococcus*, (iii) phage as drivers of microbial evolution, and (iv) mechanisms of interaction between *Prochlorococcus* and heterotrophic bacteria. Finally, as our fifth focus we began development of genetic tools for *Prochlorococcus*.

(I) Genomic diversity and adaptation

Distinct *Prochlorococcus* lineages have adapted to various combinations of light, temperature, and other factors in the open ocean. At a coarse scale, *Prochlorococcus* can be divided into phylogenetically distinct high-light (HL) adapted and low-light (LL) adapted groups that generally partition according to available light in the environment (Ahlgren et al., 2006; Johnson et al., 2006; Zinser et al., 2006). Within the HL group, two very closely related ecotypes (>99% 16S rRNA identity) have strikingly different geographic distributions along ocean gradients, which we attribute, in part, to different temperature sensitivities (Johnson et al., 2006; Zinser et al., 2006). These temperature sensitivities are of particular relevance to studies of global change.

Underlying this physiological variability, we have discovered significant genome-wide variability, both in gene content and sequence divergence, within and between ecotypes. We recently completed sequencing the genomes of 12 *Prochlorococcus* strains, and learned that the core genome shared among all strains contains about 1200 genes, while the total pool of distinct genes continues to rise with each new genome – adding 20-200 genes per new genome (Kettler et al., in preparation; Fig. A1.1). Many "non-core" or peripheral genes appear related to stress response, nutrient uptake, and cell surface biosynthesis – yielding clues to the drivers of genomic diversity in this group.

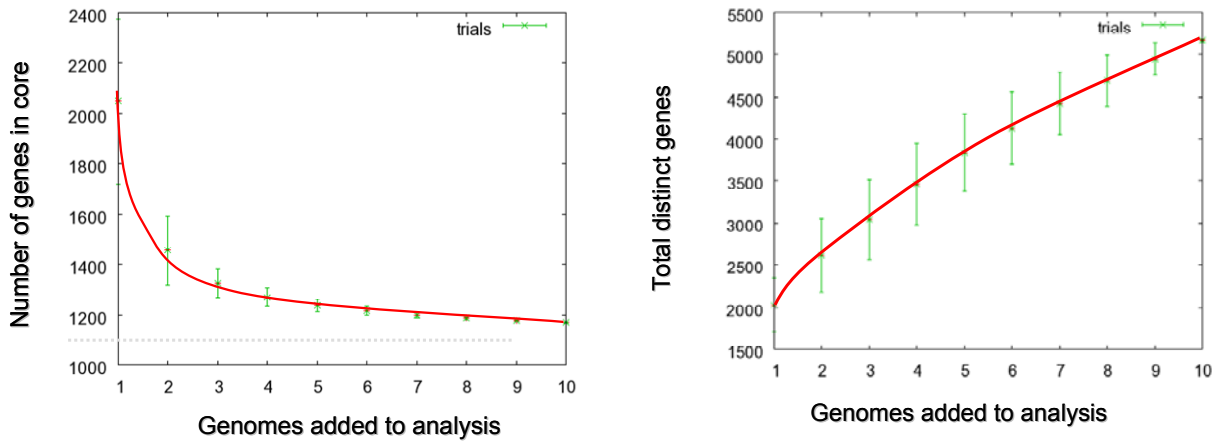


Figure A1.1 The Pan-Genome of *Prochlorococcus*. On the left is shown the number of genes shared among all sequenced *Prochlorococcus* as a function of number of genomes in the analysis. The core genome is converging at about 1200 genes. On the right is the accumulation of totally unique genes in the pan-genome of *Prochlorococcus* as you add more genomes to the analysis. Note that the function is not saturating (Kettler et al., in preparation).

Among HL-adapted *Prochlorococcus*, this variability is concentrated in genomic islands (Coleman et al., 2006); Coleman unpublished data). These islands encode ecologically important functions such as nutrient acquisition, DNA repair, and protection from excess photon flux (Coleman et al., 2006); Kettler et al., in preparation). There is no clear connection between gene content in the peripheral genome and the rRNA or core genome phylogeny. In fact, in the case of genes involved in phosphate assimilation, it appears that gene content is related to phosphate availability in the ocean, regardless of rRNA-based cell lineage (Martiny et al., 2006). This result suggests that adaptations to light and temperature – which appear consistent with the rRNA or core genome phylogeny – occur on longer time scales than adaptations to nutrient availability, and may even occur via different mechanisms. Understanding these time scales and mechanisms is critical for predicting the evolution of natural and engineered populations.

Beyond genome-wide gene content, we are beginning to grapple with how variability in gene/protein expression can affect a strain's ability to respond to environmental stimuli. Using microarrays, we have shown that HL-adapted strain MED4 and LL-adapted strain MIT9313 differ in their responses to nitrogen and phosphorus starvation, and these differences are mediated by both differences in gene content and differences in regulation of core genes (Martiny et al., 2006; Tolonen et al., 2006). Some of the most up-regulated genes are even located in genomic islands, suggesting that these islands play an important role in adaptation (Coleman et al., 2006; Martiny et al., 2006). We are now exploring physiological differences among even more closely related strains, i.e., within the HL and LL groups. We have observed significant strain-to-strain variability within a single HL clade in response to phosphorus starvation, only some of which can be readily explained by gene content differences (Coleman, Martiny unpublished). Thus the relationship between rRNA and core genome phylogeny, gene content in the peripheral genome, and nutrient physiology is unclear, and elucidation of these links is crucial to understanding adaptation in the environment.

Emerging questions

- *Is the **core genome** of* Prochlorococcus *metabolically functional?*
- *What is the extent of the **peripheral genome**, does it contribute to niche adaptation, and what are the rates of gene gain/loss from individual strains?*

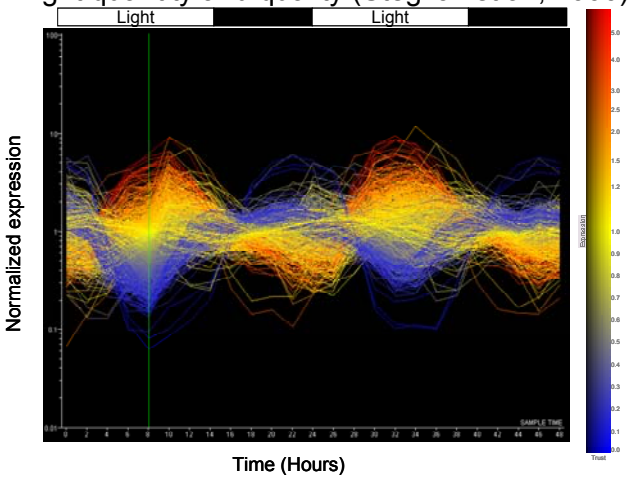
- *To what extent do different types of genomic variation (allelic variation of "core" genes, differences in "peripheral" gene content, alternative gene regulation strategies) lead to observable* ***physiological differences*** *among strains?*
- *How is adaptation to temperature and light different from adaptation to nutrient availability and other factors?*
- *How does this* ***genome-wide diversity*** *change along environmental gradients? How does diversity within a population compare to diversity between populations in different environments?*

(II) Cellular design and machinery

Understanding the cellular design of Prochlorococcus *is an important first step toward engineering this relatively simple but highly-efficient photosynthetic machine (Box 1). We have made significant progress in building the basic scientific and technical infrastructure necessary to characterize* Prochlorococcus *at the genome-wide level, and are using these measurements to elucidate key design principles and genomic adaptations to the environment within this cell.*

Gene expression profiling

Using our custom-designed Affymetrix microarrays we have studied the responses of two ecotypes of *Prochlorococcus* to N-availability (Tolonen et al., 2006), P-availability (Coleman et al., 2006; Martiny et al., 2006), and Fe-availability (Thompson et al, in preperation), and phage infection (see next section; Lindell et al., 2007). These experiments, as described above in (I), have unveiled striking differences between high-light adapted and low-light adapted ecotypes in their response to nutrient starvation, and have begun to reveal the genomic basis for observed differences in physiology. We have also measured the RNA half-life in the cell using the arrays (Steglich et al., in preperation) and studied the response of high-light adapted MED4 to changes in light quantity and quality (Steglich et al., 2006).



oscillated strongly over the diel cycle in patterns consistent with their function, and the expression profiles of regulatory genes offer insight into the processes they control.

Preliminary analyses have revealed several trends that appear to represent a coordinated cellular response to the diel cycle (Fig. A1.3 and A1.4). Based on mRNA levels, photosynthesis genes were generally up-regulated in the morning, allowing the cell to maximize energy capture and carbon fixation when solar irradiance is highest. Carbonic anhydrase (*csoS3*), which supplies Rubisco with CO_2 for carbon fixation, and Rubisco (*rbcL*) itself, were both maximally expressed just after sunrise. Genes for biosynthesis of glycogen (*glgA*, *glgB*, and *glgC*), a sugar storage molecule, were also maximally expressed early in the day. Interestingly, the *psbA* gene, which encodes the core photosystem II protein D1, was maximally expressed later, at around midday, potentially to offset damage from high solar irradiation at this time of day.

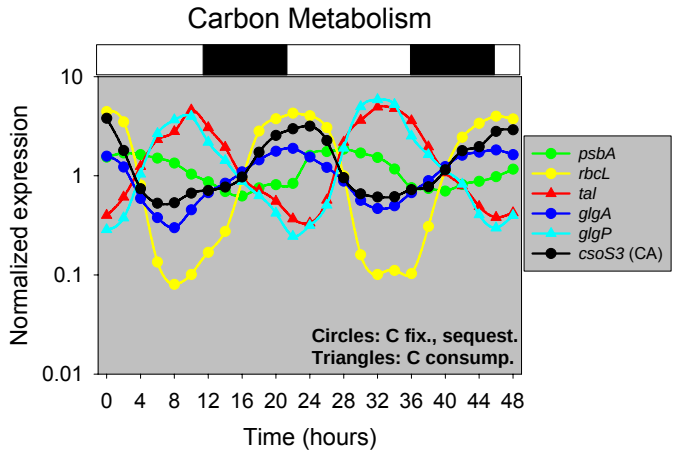


Figure A1.3. Diel variation in genes involved in carbon metabolism. Genes involved in carbon fixation via the Calvin Cycle, *rbcL* (Rubisco) and *csoS3* (carbonic anhydrase), and carbon storage (as glycogen) (*glgA*) have maximal mRNA values at the onset of day, while those involved in catabolism of carbon stores (*glgP* and *tal*) have maximal mRNA values at the onset of night. Interestingly, the photosystem II gene *psbA* has maximal mRNA values at mid-day. Hence, for reasons yet unclear, the light reaction (the photosystem) and dark reaction (Calvin Cycle) of photosynthesis are offset in *Prochlorococcus*.

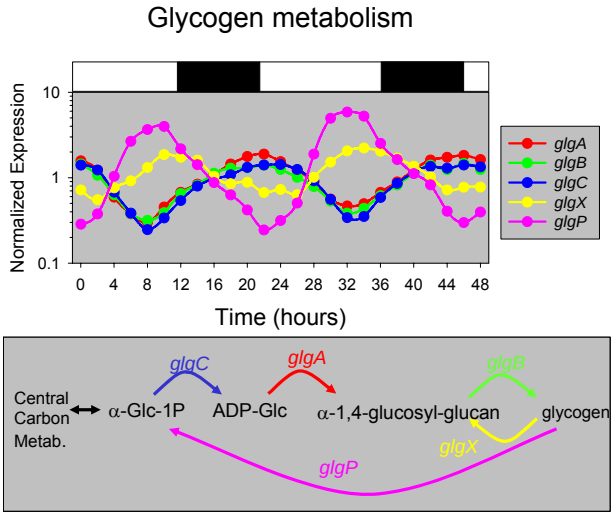


Figure A1.4. Diel variation in genes involved in metabolism of the carbon and energy storage molecule, glycogen. Genes involved in producing glycogen (glgA, glgB,a nd glgC) show nearly identical diel expression patterns, even though they do not form part of an operon. As expected, genes involved in producing glycogen show maximal mRNA expression at the onset of day, when photosynthetically-derived carbon should begin accumulating, whereas genes involved in consuming glycogen show maximal mRNA values at the end of the day, in preparation for respiration at night.

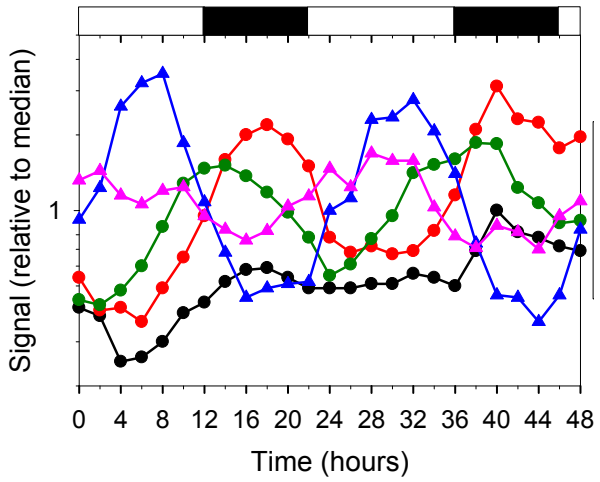


Figure A1.5. Diel expression of different high-light induced proteins *(hli)* showing the different phase relationships in the expression of these genes.

Differential expression of the *hli* gene family in *Prochlorococcus* MED4 over a diel cycle, determined from microarray analysis, suggests that these genes have undergone functional diversification since their acquisition from cyanophage (see phage section, and Lindell et al., 2007).

Conversely, genes for consumption of fixed or stored carbon are up-regulated in the evening (e.g., *glgA*, *glgB*, *glgC*, *tal*) allowing the cell to mobilize organic carbon at night in the absence of photosynthesis (Fig. A1.3 and A1.4). The *tal* gene, encoding a transaldolase, is particularly interesting because its ortholog is found in many of the phages infecting *Prochlorococcus*, suggesting that it may be a key step in cellular carbon metabolism. The transaldolase enzyme of the pentose phosphate pathway helps to generate ribose and NADPH from glucose.

By overlaying expression data onto known pathways, we observe that the full suite of genes for fixing and storing carbon are up-regulated in the morning, and genes for mobilizing that stored carbon and for generating ribose for nucleotides and NADPH for biosynthetic reactions, are all up-regulated in the evening. These data provide an invaluable framework for understanding the regulation of carbon metabolism under a variety of conditions, including phage infection.

The *hli* genes encode small stress response proteins involved in the response of cyanobacteria to high light, nutrient deprivation and temperature shock ((He et al., 2001)). They are thought to protect the cell through the dissipation of excess light energy during stressful conditions (Havaux et al., 2003). High-light adapted *Prochlorococcus* strains such as MED4 encode over 20 *hli* genes, far more than other cyanobacteria (Bhaya et al., 2002). The expansion of the *hli* gene family in HL adapted *Prochlorococcus* appears to have been mediated by cyanophage (Lindell et al., 2004). Members of this gene family are differentially expressed over a diel cycle (Fig. A1.5) and in response to different stressors (Tolonen et al., 2006); Steglich et al. in preparation, Lindell et al., 2007) (Fig. A1.6) suggesting that they have undergone functional diversification since their acquisition from cyanophage.

Proteomic analyses

Protein expression is the essential link between genomic diversity and mRNA expression on the one hand and cellular biochemistry and metabolism on the other. To enable a mechanistic, molecular-level understanding of how *Prochlorococcus* interacts with its oceanic environment, we are developing methods for quantifying the entire *Prochlorococcus* proteome. As a genetically- and metabolically-streamlined organism, *Prochlorococcus* is an ideal target for global proteomic analysis, since its genome encodes fewer than 2000 gene products. We can now identify and quantify as many as 784 proteins in a single mass spectrometric analysis, with

two-thirds of those on the basis of multiple distinct peptides (Leptos, unpublished results). Further refinements to extraction, digestion and chromatography protocols are expected to increase this number. We are developing quantification via both isotopic (^{15}N) labeling (Waldbauer, Krastins, Sarracino, unpublished results) and label-free strategies (e.g., (Jaffe et al., 2006). Initial experiments suggest that the transcript-level microarray analyses of gene expression described above can now be augmented with analyses at the translational level (Lindell et al., 2007); Leptos, unpublished results). The ability to track gene expression through to the protein level will yield new biological insights, particularly given the unique aspects of the molecular physiology of *Prochlorococcus*, such as the environmentally-driven diel cell cycle described above. Further, the small cell size means that the pool of gene products (transcripts and proteins) present at a given time is very limited, and such a small system is likely susceptible to stochastic effects in gene expression, such as translational bursting (Kærn et al., 2005). Ultimately, we aim to have a quantitative, time-resolved picture of global gene expression in *Prochlorococcus*.

Metabolic Reconstruction

To address the fundamental question of how genomic differences translate into differences in metabolism, we have constructed a first draft of a metabolic model of *Prochlorococcus,* based on our genome annotations in collaboration with the Church Lab (Kettler, Martiny, Zucker et al., in preparation) Our goal is to understand the contribution of the core genome and the peripheral genes to each strain's unique physiology. Thus far, strains appear to differ significantly in their nutrient uptake and salvage pathways. We suspect these alternate routes play a role in adaptation, perhaps by allowing a cell to use intermediate products that are in greater supply due to the distribution of intracellular fluxes. The heterogeneity of the ocean at the microscale where it is experienced by microorganisms – e.g., patches of nutrients released by lysis and feeding events, or in the wakes of sinking particles – also contributes to fluctuating metabolite availability. To take advantage of these transient supplies, the metabolic network must be able to respond on an appropriate timescale. The ability to sense fluctuations and induce alternate pathways in response to certain stimuli may confer significant ecological advantage.

Testing these ideas requires a picture of the metabolic capabilities of each strain that is as complete as possible. We have developed software using a mixed integer linear programming (MILP) technique to quickly develop a draft metabolic reconstruction for each newly sequenced genome (Lin, Brandes, Zucker, unpublished results). This software flags particular genes and pathways to facilitate subsequent rounds of manual curation in which genes are reannotated and poorly supported pathways are eliminated.

Emerging questions:

- *Which proteins and metabolites constitute **major nutrient demands** for N in amino acids, P in nucleic acids, and trace metals as cofactors? How does* Prochlorococcus *structure its nutrient uptake systems to meet those demands? Do membrane transport systems show adaptation to different environments?*
- *What are the **temporal patterns of gene expression**, at both the transcriptional and translational levels? Does translation closely follow transcription, or are there significant lags? How does this reflect and impact diel cycling in the oceanic water column?*
- *What is the **metabolic response to environmental fluctuations**, such as nutrient or light availability? The marine water column is actually highly heterogeneous on the microscale. How quickly do different elements of the metabolic network respond to input*

perturbation? How do metabolic shifts contribute to release of extracellular carbon compounds?

(III) Phage as key players in microbial evolution

Work in the Chisholm lab has revealed that phage play an important role in the evolution of their *Prochlorococcus* hosts. Their most obvious role is as vectors for HGT, and indeed genomic islands in *Prochlorococcus* show signs of phage influence (Coleman et al., 2006). By characterizing our extensive cyanophage collection (Sullivan et al., 2003); Sullivan unpublished data), we have shown that some phages are host-specific, while others cross-infect multiple ecotypes and even cross cyanobacterial genera. These host range patterns help establish the boundaries of a single step in phage-mediated HGT.

Analysis of 3 sequenced cyanophage genomes reveals that podovirus P-SSP7, and two myoviruses show similarity to their T7 and T4 counterparts, despite the different habitats of enteric phages and cyanophages. All three cyanophage genomes contain cyanobacterial 'host' genes (Sullivan et al., 2005), including genes encoding proteins involved in the photosynthetic apparatus, such as the core reaction center proteins D1 and D2 (*psbA* and *psbD)* and high light induced proteins *hliP*, (Lindell et al., 2005; Lindell et al., 2004; Mann, 2003; Millard et al., 2004; Sullivan et al., 2005; Sullivan et al., 2006). Other 'host' genes include transaldolase (*talC*), phosphate stress (*pstS* and *phoH*), and ribonucleotide reductase (*rnr*) genes.

Of 33 cyanophage isolates, nearly all contained *psbA* while a subset contained *psbD* (Sullivan et al., 2006). Phylogenetic analyses suggests that whole *psbA* and *psbD* genes have each been transferred from host to phage through only a few discrete events, followed by horizontal and vertical transfer between cyanophages over the course of evolution (Sullivan et al., 2006). Sequence analysis revealed signatures of *intra*genic recombination both from host-to-phage and from phage-to-host, indicating that cyanophage-encoded photosynthesis genes play an important role in photosystem evolution.

Among our most exciting unpublished results to date is the detailed analysis of whole genome response of host and phage during the infective process (Fig. A1.6). (Lindell et al., 2007; Lindell et al., 2005). Podovirus P-SSP7 genome expression progresses in a linear fashion from left to right after infection (Fig. A1.6, top and bottom), except for a few genes that are of notable interest. The cyanophage-encoded "host" genes *psbA*, *hli*, *nrd* and *talC,* form a gene expression group, despite the location of *talC* at the end of the phage genome (Fig. A1.6). It is likely that this cluster forms a functional module, unique to cyanophages, that allows for the enhanced acquisition of energy and/or nucleic acid substrates during infection of resource-limited marine cyanobacteria. High-throughput proteomics showed that 74% of the predicted P-SSP7 proteins were expressed during infection (Fig. A1.6, bottom). Finally, while transcription of most host genes was down-regulated during the 8-h latent period (Fig. A1.6-B), some genes involved in carbon fixation, ribonuclease, RNA modification, *hli* stress response, a *lexA*-like SOS transcriptional regulator, and genes of unknown function were upregulated (Lindell et al. 2007). These genes are likely part of the host's stress response to infection.

Emerging questions

- *What is the breadth of* ***cyanophage diversity****, and how do morphological and genome "types" influence host-specificity?*
- *What* ***additional host genes*** *do cyanophage carry, and how might they influence host metabolism during infection, and evolution of the gene pool in host and phage?*

- *How do the **properties of phage-encoded host genes** differ from their host-encoded analogs in vitro? Do the phage enzymes have advantages over the host enzymes?*

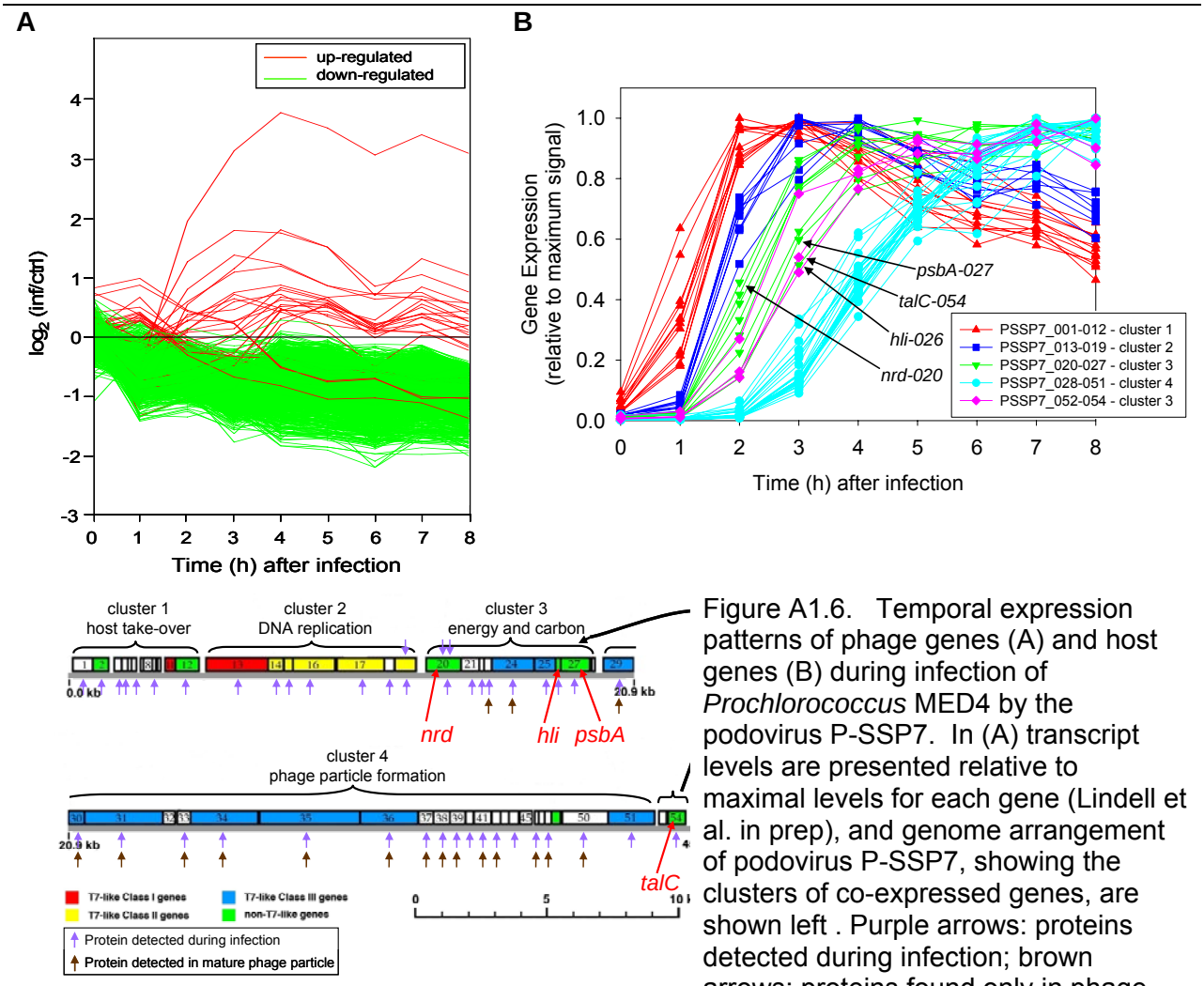


Figure A1.6. Temporal expression patterns of phage genes (A) and host genes (B) during infection of *Prochlorococcus* MED4 by the podovirus P-SSP7. In (A) transcript levels are presented relative to maximal levels for each gene (Lindell et al. in prep), and genome arrangement of podovirus P-SSP7, showing the clusters of co-expressed genes, are shown left . Purple arrows: proteins detected during infection; brown arrows: proteins found only in phage particles. (Lindell et al. 2007)

(IV) Mechanisms of interaction between *Prochlorococcus* and heterotrophic bacteria

Understanding syntrophic interactions is important for designing mass cultures, because the nature of the autotroph-heterotroph interaction will greatly influence yield of such culture-based systems. *Prochlorococcus* cultures are notoriously difficult to free of heterotrophic bacterial contaminants that were co-cultured with them during isolation from the wild. Since no organic carbon is added to the medium, heterotrophs must grow on carbon either secreted by *Prochlorococcus* or released from dying cells. We have embraced this opportunity to better understand a naturally-occurring syntrophy, and discovered that the presence of heterotrophic bacteria (*Alteromonas alvinella*) actually increases *Prochlorococcus* yield in stationary phase

under CO_2-limited conditions (Fig. A1.7). Because this effect is not seen in stationary phase cultures under nitrogen- or phosphorous-limitation, it is thought that heterotrophs partially relieve CO_2 limitation by respiratory release of CO_2 (Drakare, unpublished results).

Our preliminary work also suggests that the presence of heterotrophic bacteria facilitates the growth of *Prochlorococcus* cultures from low density — an effect that could involve supplying something other than CO_2 as it would not be limiting at these densities.

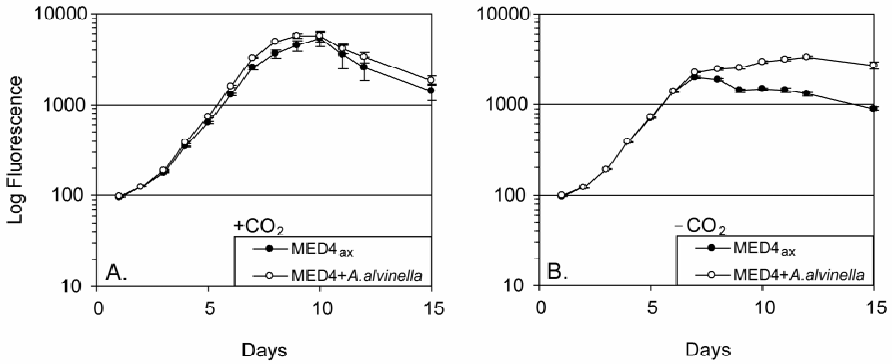


Figure A1.7. Growth curves of axenic MED4 ($MED4_{ax}$) grown with and without heterotrophic bacteria (*Alteromonas alvinella*) in the presence (left) or absence (right) of elevated CO_2 (by bubbling).The only available organic carbon source is that released by *Prochlorococcus*. Note that *Alteromonas* increases the relative yield of *Prochlorococcus* only in the absence of elevated CO_2 (Drakare, unpublished results).

Emerging questions

- *Do* Prochlorococcus *cells* ***actively secrete organic carbon*** *compounds that the heterotrophic cells exploit? Or do the heterotrophs live off organic carbon that is released from dying cells?*
- *How many* ***different compounds*** *are secreted or released? Does the spectrum of compounds change with ecotype, with growth phase, or after viral infection?*
- *Is there evidence of* ***signaling*** *between these two cells?*

(V) New genetic tools

Developing robust genetic tools for *Prochlorococcus* is vital in order to use it as a chassis for synthetic biology. We have made steady progress in this system, although our experiments are hampered by slow cell growth. Of note are three significant advances. First, we can now introduce foreign DNA at low frequency into HL *Prochlorococcus* MED4 and LL *Prochlorococcus* MIT9313 via conjugation with *E. coli* (Tolonen, 2005). Second, we implemented transposon mutagenesis in MIT9313 using transposon *Tn5*, (Tolonen, 2005). Third, we have identified a more appropriate selective marker for future genetic experiments (Table A1.1, Osbourne unpublished results). While all antibiotics tested inhibited *Prochlorococcus* growth (except for the DNA gyrase inhibitor nalidixic acid), only chloramphenicol treatment prevented spontaneous resistant mutants from appearing. Future experiments will now use chloramphenicol-resistance as a selective marker.

Table A1.1. Sensitivity of *Prochlorococcus* MED4 to Antibiotics[a]

Antibiotic	*Cellular Target*	*Inhibitory concentration*	*Experimental Repetitions* (n=)	*Appearance of Resistant mutants?*
Nalidixic acid	DNA gyrase, A subunit	> 50 µg /ml	4	All resistant
Ciprofloxacin	DNA gyrase, A subunit, and Topoisomerase IV	2–3 µg /ml	3	Yes
Chloramphenicol	Protein synthesis (50S ribosome)	25 µg /ml	3	No
Kanamycin	Protein synthesis (30S ribosome)	50 µg /ml	4	Yes
Rifampicin	RNA polymerase	5 µg /ml	3	Yes

[a]MED4 cells were grown in complete Pro99 medium using Sargasso Sea water, at a light level of 25 µE m^{-2} sec^{-1}. Growth was followed by measuring fluorescence, and was compared to a no-drug control culture.

Other work in progress includes improved plating efficiencies for *Prochlorococcus* and identification of promoters of various strengths to engineer inducible expression systems.

In addition, we have begun to explore the possibility of effecting gene transfer in *Prochlorococcus* by means of generalized transduction. Experiments are currently in progress to determine whether any of a number of broad host range cyanophage are capable of gene transfer by this mechanism.

VIBRIO – Polz laboratory

The Polz lab has addressed *Goal 3* (Characterize Complex Microbial Communities in their Natural Environments at the Molecular Level) and *Goal 4* (Develop Computational Methods to Understand Complex Biological Systems and Design new Systems) under the auspices of the Harvard-MIT Genomics:GTL Center. This led to significant advances in our understanding and analysis capabilities of natural microbial communities. We used a coastal ocean site to address structure-function relationships in microbial communities, and to develop a model system to study genome diversity and dynamics within wild populations. This led to the use of *Vibrio* as our major model, which we propose to continue here with a major focus on bioenergy relevant aspects of *Vibrio* biology (some aspects and important collaborations are developed in the parallel DOE VIBRANT bioenergy research center proposal).

(I) Emergent patterns of community structure and co-existing genomic diversity in the environment

Because genome properties arise in feedback with the environment, they can ultimately only be understood in the context of environmental populations and communities. We therefore have chosen to analyze one model microbial community for coexisting diversity of bacteria and emergent patterns of relationships. This led us to formulate hypotheses about what may constitute ecologically distinct populations and to take initial steps towards testing the genomic diversity of one such population in the environment. Because the vast majority of microbes

remain unculturable, culture-independent approaches (such as PCR-generated and metagenomic gene libraries) remain our main tool for determining microbial community structure and properties. However, these approaches generally reveal so much sequence diversity that there is little redundancy in occurrence of genotypes (Polz et al., 2006). Thus a central question has been how to group sequences to allow identification of ecologically distinct populations.

Under the auspices of the Harvard-MIT Genomics:GTL Center, we showed that despite 20 years of application of culture-independent approaches most previous estimates of microbial diversity are biased due to methodological artifacts and insufficient sampling of communities (Acinas et al., 2004a; Acinas et al., 2004b; Acinas et al., 2005; Klepac-Ceraj et al., 2004). This has led to widely varying estimates of diversity, often based on concessions to methodological issues rather than the apparent diversity encountered. Further, we provided the first well-constrained estimate of bacterial ribotype diversity for a complex, natural community (Acinas et al., 2004b; Marcelino et al., 2006).

Most important, our approach allowed us to observe fine-scale patterns in bacterial diversity and to formulate hypotheses regarding their origins (Acinas et al., 2004a; Klepac-Ceraj et al., 2004; Polz et al., 2006).We unambiguously showed for the first time that the large majority of co-existing bacteria fall into clusters containing extremely closely related (microdiverse) taxa. Over half the diversity in the microbial population is in sequence clusters with <1% divergence (Fig. A1.8). This observation provides rich substance for interpretation by ecological and evolutionary theory as clusters have previously been hypothesized to originate by selective sweeps triggered by ecological specialization (reviewed in Polz et al., 2006).

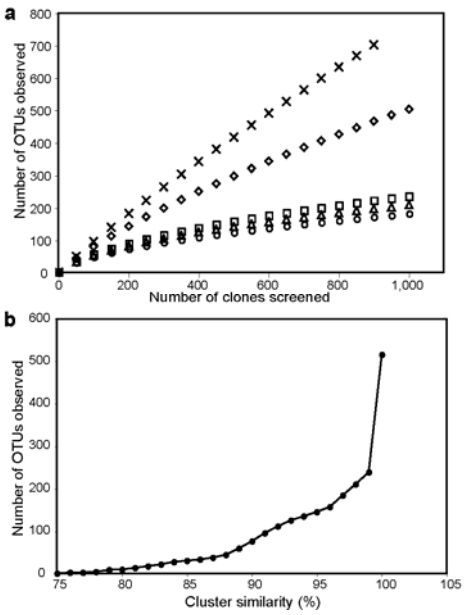


Figure A1.8. Compositional pattern of a coastal bacterioplankton sample. **a**, Rarefaction curves of the number of OTUs (operational taxonomic units) in a 16S rRNA library constructed with standard amplification protocols; modified protocols were designed to remove and constrain PCR-induced artifiacts. Standard deviations were smaller than symbols and are not shown. (x symbol represents 100% sequence similarity cluster, and modified diamond – 100%, square – 99%, triangle – 98%, and circle - 97% sequence similarity clusters) **b**, Number of OTUs vs. changing degrees of cutoffs in 0.5% increments for grouping of sequences into similarity clusters. From Acinas et al. (2004).

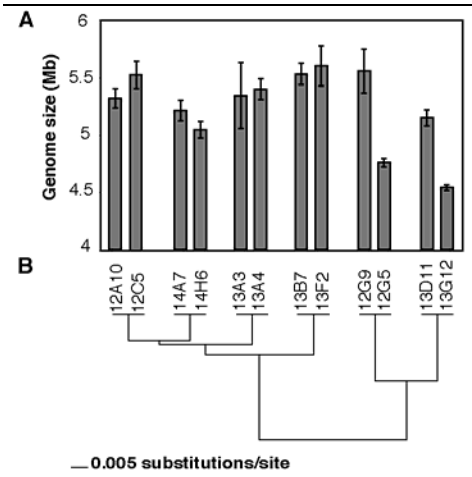


Figure A1.9. Genome size estimates and phylogenetic relationships of Hsp60 sequences for 12 *V. splendidus* isolates chosen as pairs of identical Hsp60 alleles, encompassing all levels of Hsp60 variation observed in the strain collection. **(A)** Genome sizes determined by PFGE as averages of six independent estimates, each obtained from single enzyme digests run to resolve large, medium and small-sized bands, respectively, and repeated three times for each of two enzymes (NotI/SfiI or NotI/AscI) per isolate. **(B)** Phylogenetic relationships of Hsp60 alleles inferred from maximum likelihood analysis with assumption of molecular clock from partial gene sequences. Isolate identifiers correspond to month (12 = 8/12/03; 13 = 9/10/03; 14 = 10/11/03) of isolation and strain name. From (Thompson et al., 2005).

We therefore addressed the diversity (i.e., number and differentiation) of genomes co-existing within a naturally occurring cluster within our model community. We quantitatively analyzed the annual dynamics of a narrowly defined group of coastal bacterioplankton (>99% 16S rRNA identity to *Vibrio splendidus*) by a combination of culture-independent and dependent techniques (Thompson et al., 2005). We showed that this group consists of at least a thousand distinct genotypes, each of which occurs at extremely low environmental concentrations (on average <1 cell/ml). Moreover, the genomes showed extensive neutral allelic diversity and size variation (Fig. A1.9) (Thompson et al., 2005). Individual genotypes rarely recurred in samples and allelic distribution did not show spatial or temporal substructure. The nature of this genetic variation remains unknown; however, the data suggest that many genotypic (and perhaps phenotypic) traits occur at extremely low frequency within the population. Thus a considerable portion of genomic variation may co-exist in a near neutral fashion and/or play a role under changing conditions (e.g., stress) as recently suggested for *Prochlorococcus* (Coleman et al., 2006).

Emerging questions:

- *Is organization into **clusters a universal feature of microbial communities**?*
- *Do observable **genetic clusters** arise via selective sweeps and do they thus represent ecotypes? What may be alternative mechanisms of cluster formation (e.g., homologous recombination, drift)?*

(II) Vibrio population structure

Population structure is a central component of understanding the dynamics of genome evolution. Although it has become feasible to sample bacterial genomes on large scales, surprisingly little is known about distribution of genome types in the environment on appropriate temporal and spatial scales. As part of the Harvard-MIT Genomics:GTL Center, we have laid the foundation for incorporating population genetic models into genome analysis. We have already collected thousands of Vibrio strains, which we have isolated over the last four years from defined environments, which give us the unique opportunity to explore mechanisms of population structure and genome adaptation in an ecological context.

- Spatio-temporal niches and population structure: In a preliminary analysis, we have completed multi-locus sequence analysis (MLSA) of ~300 *Vibrio* strains. We identified multiple hierarchical sequence clusters within named *Vibrio* species (Fig A1.10). Such clusters have been proposed as ecological populations [reviewed in (Polz et al., 2006)]. Because the abundance and dynamics of such clusters can be assayed in samples, MLSA provides the first step towards identification of specific association and response to defined environmental factors on multiple spatial and temporal scales. Our ultimate goal is to analyze genomic diversity within and between such clusters, and to identify genes and pathways unique to (sub)populations and their environmental correlates.

- Phenotypic diversity: We have started to characterize the metabolic diversity of our strain collection by measuring carbon substrate utilization and growth parameters (Fig A1.10). This further enables us to correlate population characteristics with environmental (niche) parameters in order to construct a fitness landscape for the different genomes and clusters of genomes.

- Gene transfer by homologous recombination: The MLSA dataset provides the opportunity to determine rates and bounds on homologous recombination among strains co-existing in the wild. In a preliminary study, we have seen high rates of recombination between 100 and 95% nucleotide differentiation and essential genetic isolation at 90% nucleotide differentiation (Fig A1.10). This indicates that even in the absence of complete reproductive isolation as in (at least some) eukaryotic species, genetic boundaries may exist, which limit homologous recombination among genomes. However, we (with Eric Alm) are currently in the process of developing better tools for identifying recombination events since current methods lack resolution for closely related sequences.

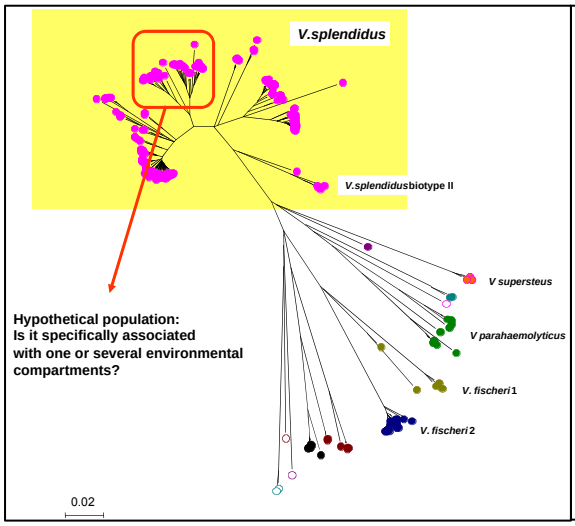


Figure A1.10. Multilocus sequence analysis (MLSA) of vibrios isolated from the water column at our model site (Plum Island Estuary). Note distinctness of named species and hierarchical clusters within the main model *V. splendidus.* Our preliminary results show that a unique MLSA type (*i.e.*, strains with identical alleles in all loci) will be a good marker for a unique genome since two other high-resolution measures correspond: pulsed field gel electrophoresis and O-antigen patterns (Wildschutte and Polz, unpublished). In particular, the O-antigen has been shown to be hypervariable due to diversifying selection [reviewed in (Wildschutte et al., 2004). This high resolution allows fine-scale genotype-environment correlation.

Emerging questions:

- *What are the **patterns of genomic and phenotypic differences** within or between MLSA clusters?*
- *What are the **dominant modes of genome differentiation** within and between clusters (e.g., gene addition and loss, adaptation by point mutation)?*

- *Do differences in **rates of homologous recombination** sufficiently isolate genome clusters from others to be equivalent to species-like boundaries in gene flow?*
- *What is the **role of horizontal gene transfer** by illegitimate recombination in differentiating genome clusters? Is there a specific, **laterally transferred gene pool** (either directly or via extrachromosomal elements)? What types of genes are subject to horizontal gene transfer and what is their turnover within genomes?*
- *How rapidly does **selection** purify genomes from unused genes (i.e., to what extent can genes persist neutrally in genomes)?*
- *Are clusters specific to **definable environmental conditions**? Conversely, do populations assemble neutrally in at least some environments?*

(III) Interaction with algae

Interactions between heterotrophic bacteria and algae are an important component of natural and engineered bioenergy systems. Such interactions may lead to conversion of sunlight to various bioenergy endproducts, but they also may lead to undesirable respiratory losses in the system. Indeed, the partners can either compete for resources or cooperate in a species or condition dependent manner. We are currently studying the specific interaction of heterotrophic bacteria with algae by experimental manipulation in microfluidics systems and mathematical modeling. The initial work is specifically funded by a grant from NSF (in collaboration with Roman Stocker who is a fluid dynamics specialist); here, we propose to establish Vibrio as a systems biological model to study algae-heterotroph interactions. We have already assembled a collection of different eukaryotic algae, which are axenic so that they can be used to experimentally characterize *Vibrio*-algae interactions. Further, we can use our microfluidic system to quantify the influence of behavior on interactions between algae and bacteria. This setup will be particularly powerful when characterizing mutants, which display show altered characteristics of interactions (e.g., decreased transfer of carbon substrates).

Emerging questions:

- *Are specific strains **adapted for interactions** with algae (e.g., chemotaxis towards algal exudates)?*
- *What are the **genetic traits involved** in carbon transfer between algae and heterotrophs?*
- *What are the specific conditions under which heterotrophs **enhance or depress algal productivity**?*

(IV) Development of a genetic system

We are currently in the process of developing a genetic system for our model organism, *V. splendidus* and *V. parahaemolyticus* to be able to create specific knockouts or allelic replacements.

We have chosen conjugation as the historically most effective means of mutagenesis of vibrios. This is mediated by cell-to-cell contact, which requires specific transfer and replication properties of both plasmid and donor (e.g. suicide plasmid described in (Miller and Mekalanos, 1988). This approach has been used in vibrios to construct chromosomal site-directed insertion and deletion mutants, and we are currently using it to generate mutants of isolates for which genomic information is available (e.g. *Vibrio splendidus*). Additionally, transposons have been engineered into such conjugation vectors (de Lorenzo et al., 1990; Herrero et al., 1990),

enabling further gene disruptions, insertions and promoter probing, which is then not limited to genes or regions for which there is sequence information. Therefore, this type of transposon mutagenesis seems particularly well-suited to environmental isolates and it has been successfully employed in several gram-negative organisms; we intend to investigate the possibility of using this type of genetic manipulation in a wide variety of selections and screens.

An important synergy may arise from our screening for transducing phages. These may be used to move genetic markers and mutations between strains, thereby enabling sophisticated strain construction with relative ease. At least one generalized transducing phage from *V. cholerae*, CP-T1, has been isolated (Ogg et al., 1981), characterized, and optimized (Hava and Camilli, 2001; O'Shea and Boyd, 2002); however, its use is likely to be species-specific. More recently the genome sequence of KVP40, a broad-host-range vibriophage isolated from *V. parahemolyticus* (Matsuzaki et al., 1992), has been determined (Miller et al., 2003) and is likely to be useful in generalized gene transduction in and between a variety of vibrio species. We anticipate the development and use of this transduction system with vibrios in our collection will be particularly insightful regarding questions related to species- or strain-specific genes or markers.

Transformation, the uptake of extracelluar DNA, is yet another means of genetic exchange, but it was not possible with vibrios until recently. Meibom and colleagues have shown that transformation competence in *V. cholerae* can be induced by chitin, as well as nutrient limitation and cell density (Meibom et al., 2005). Their study identified many genes, which are required for chitin-induced competence including *tfoX*, a regulator of competence-specific genes, which is conserved among other vibrios including *V. splendidus* and *V. alginolyticus*. This suggests that competence may not be limited to *V. cholerae*, but rather may be a generalized feature of other vibrios in the environment. Thus we can begin to develop methods of competence induction in order to conduct transformations of vibrios with various plasmids, transposons, and other DNAs for mutagenesis or gene replacement.

Emerging questions:

- *Can natural population properties (e.g., phages, transformation) be used to develop* ***rapid genetic manipulation*** *techniques for bacteria?*

(V) Experimental evolution of reproductive rate

We have developed a chemostat system, which allows efficient selection for maximum reproductive rate in bacteria. Using *V. natriegens*, which already has reported doubling times of ~10 min, we are evolving strains for maximum possible reproductive rate. This gives us the opportunity to ask fundamental systems biological questions of genome adaptation. The ability to reproduce is one of the fundamental characteristics of life, and the entire theory of evolution rests on optimizing reproduction rates. The factors limiting this rate are complex, particularly for sexual organisms. Yet even for bacteria, fundamental questions about the maximum possible reproductive rate remain unanswered. It is widely believed that the minimum doubling time for bacteria cannot be less than 9 minutes; however, our preliminary study suggests that we have experimentally evolved strains towards doubling times of <4 minutes. To our knowledge, this represents the fastest growing organism. We are now interested in the genomic and phenotypic changes that lead to such reproductive rate optimization.

Emerging questions:

- *What **types of mutations** are adaptive for increase in growth rate?*
- *To what extent are these **mutations reproducible** (i.e., does the same selection pressure lead to convergent evolution)?*
- *Can **other organisms** evolve to similarly fast growth rates?*
- *Do organisms evolved to grow faster in complex media display similarly increased growth rates in minimal media?*

(VI) Analysis tools for complex communities

The enormous genetic diversity, which co-exists within natural environments, challenges the analysis of composition and dynamics of microbial communities. We have developed two new tools: (i) a novel analytical tool for application of oligonucleotide arrays to natural communities (Marcelino et al., 2006), and (ii) new sets of universal phylogenetic primers targeting the 23S rRNA (Hunt et al., 2006; Polz et al., 2006)).

Application of microarrays to environmental communities remains particularly challenging since most transcripts exist in low abundance in complex mixtures of sequences with varying degrees of similarity to the target so that spurious contributions may mask the specific signal. To address this challenge, we have developed a new analytical solution to the open problem of microarray quantification of low abundant targets in complex pools of similar sequences (Marcelino et al., 2006). We combined a new analytical predictor of non-specific probe-target interactions (cross-hybridization) with a new optimization algorithm, which iteratively deconvolutes true probe-target signal from raw signal affected by spurious contributions (cross-hybridization, noise, background and unequal specific hybridization response). The method was capable of quantifying with unprecedented specificity and accuracy ribosomal RNA (rRNA) sequences in artificial and natural communities (Marcelino et al., 2006). We accurately identified specific *Vibrio* taxa in coastal marine samples at their natural concentrations (<0.05% of total bacteria) despite the high potential for cross-hybridization by hundreds of different co-existing rRNAs of vibrios and other community members. This suggests that this methodology should be expandable to any microarray platform and system.

Emerging questions:

- *Can microarrays be used to **monitor shifts in thousands of co-occurring populations**?*
- *Can the microarray approach be expanded to **mRNA measurement within natural communities**?*

References for Aim S6 Progress Report

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., and Polz, M.F. (2004a). Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430, 551-554.

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004b). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol 186, 2629-2635.

Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 71, 8966-8969.

Ahlgren, N.A., Rocap, G., and Chisholm, S.W. (2006). Measurement of Prochlorococcus ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environmental Microbiology 8, 441-454.

Bhaya, D., Dufresne, A., Vaulot, D., and Grossman, A. (2002). Analysis of the hli gene family in marine and freshwater cyanobacteria. FEMS Microbiology Letters 215, 209-219.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., DeLong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. Science 311, 1768-1770.

de Lorenzo, V., Herrero, M., Jakubzik, U., and Timmis, K.N. (1990). Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in gram-negative eubacteria. J Bacteriol 172, 6568-6572.

Hava, D.L., and Camilli, A. (2001). Isolation and characterization of a temperature-sensitive generalized transducing bacteriophage for Vibrio cholerae. J Microbiol Methods 46, 217-225.

Havaux, M., Guedeney, G., He, Q., and R., G.A. (2003). Elimination of high-light-inducible polypeptides related to eukaryotic chlorophyll a/b-binding proteins results in aberrant photoacclimation in Synechocystis PCC6803. Biochimica et Biophysica Acta 1557, 21-33.

He, Q., Dolganov, N., Bjorkman, O., and Grossman, A.R. (2001). The high light-inducible polypeptides in Synechocystis PCC6803. Expression and function in high light. Journal of Biological Chemistry 276, 306-314.

Herrero, M., de Lorenzo, V., and Timmis, K.N. (1990). Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. J Bacteriol 172, 6557-6567.

Hunt, D.E., Klepac-Ceraj, V., Acinas, S.G., Gautier, C., Bertilsson, S., and Polz, M.F. (2006). Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. Appl Environ Microbiol 72, 2221-2225.

Jaffe, J.D., Mani, D.R., Leptos, K.C., Church, G.M., Gillette, M.A., and Carr, S.A. (2006). PEPPeR, a Platform for Experimental Proteomic Pattern Recognition. Mol Cell Proteomics 5, 1927-1941.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Malcolm, E., S., S.W., and Chisholm, S.W. (2006). Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science 311, 1737-1740.

Kærn, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: From theories to phenotypes. Nature Reviews Genetics 6, 451-464.

Klepac-Ceraj, V., Bahr, M., Crump, B.C., Teske, A.P., Hobbie, J.E., and Polz, M.F. (2004). High overall diversity and dominance of microdiverse relationships in salt marsh sulphate-reducing bacteria. Environ Microbiol 6, 686-698.

Lindell, D., jaffe, J.D., Coleman, M., Axmann, I., Rector, T., Kettler, G., Sullivan, M., Steen, R., Hess, W., Church, G., et al. (2007). Genome-Wide Expression Dynamics of a Marine Virus and its Host Reveal Features of Co-evolution. Nature In revision.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. In Nature, pp. 86-89.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. PNAS 101, 11013–11018.

Mann, N.H. (2003). Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiology Reviews 27, 17-34.

Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Preheim, S.P., Lien, C., Lim, E., Veneziano, D., and Polz, M.F. (2006). Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. PNAS 103, 13629-13634.

Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006). Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. PNAS 103, 12552-12557.

Matsuzaki, S., Tanaka, S., Koga, T., and Kawata, T. (1992). A broad-host-range vibriophage, KVP40, isolated from sea water. Microbiol Immunol 36, 93-97.

Meibom, K.L., Blokesch, M., Dolganov, N.A., Wu, C.-Y., and Schoolnik, G.K. (2005). Chitin induces natural competence in Vibrio cholerae. Science 310, 1824-1827.

Millard, A., Clokie, M.R.J., Shub, D.A., and Mann, N.H. (2004). Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. PNAS 101, 11007–11012.

Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A., Feldblyum, T.V., White, O., Paulsen, I.T., Nierman, W.C., et al. (2003). Complete genome sequence of the broad-host-range vibriophage KVP40: Comparative genomics of a T4-related bacteriophage. J Bacteriol 185, 5220-5233.

Miller, V.L., and Mekalanos, J.J. (1988). A novel suicide vector and its use in construction of insertion mutations: osmoregulation of outer membrane proteins and virulence determinants in Vibrio cholerae requires toxR. J Bacteriol 170, 2575-2583.

O'Shea, Y.A., and Boyd, E.F. (2002). Mobilization of the Vibrio pathogenicity island between Vibrio cholerae isolates mediated by CP-T1 generalized transduction. FEMS Microbiology Letters 214, 153-157.

Ogg, J.E., Timme, T.L., and Alemohammad, M.M. (1981). General transduction in Vibrio cholerae. Infect Immun 31, 737-741.

Polz, M.F., Hunt, D.E., Preheim, S.P., and Weinreich, D.M. (2006). Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 2009-2021.

Steglich, C., Futschik, M., Rector, T., Steen, R., and Chisholm, S.W. (2006). Genome-wide analysis of light sensing in Prochlorococcus. J Bacteriol, JB.01097-01006.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005). Three Prochlorococcus cyanophage genomes: Signature features and ecological interpretations. PLOS Biology 3, e144.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLOS Biology 4, e234.

Sullivan, M.B., Waterbury, J., and Chisholm, S.W. (2003). Cyanophages infecting the oceanic cyanobacterium, Prochlorococcus. Nature 424, 1047-1051.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., Sarma-Rupavtarm, R., Distel, D.L., and Polz, M.F. (2005). Genotypic diversity within a natural coastal bacterioplankton population. Science 307, 1311-1313.

Tolonen, A.C. (2005). Prochlorococcus genetic transformation and genomics of nitrogen metabolism. In Joint Program in Biological Oceanography, MIT/WHOI (MIT), pp. 148.

Tolonen, A.C., Aach, J., Lindell, D., Johnson, Z.I., Rector, T., Steen, R., Church, G.M., and Chisholm, S.W. (2006). Global gene expression of Prochlorococcus ecotypes in response to changes in nitrogen availability. Mol Syst Biol 2.

Wildschutte, H., Wolfe, D.M., Tamewitz, A., and Lawrence, J.G. (2004). Protozoan predation, diversifying selection, and the evolution of antigenic diversity in Salmonella. Proc Natl Acad Sci U S A 101, 10644-10649.

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006). Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol 72, 723-732.

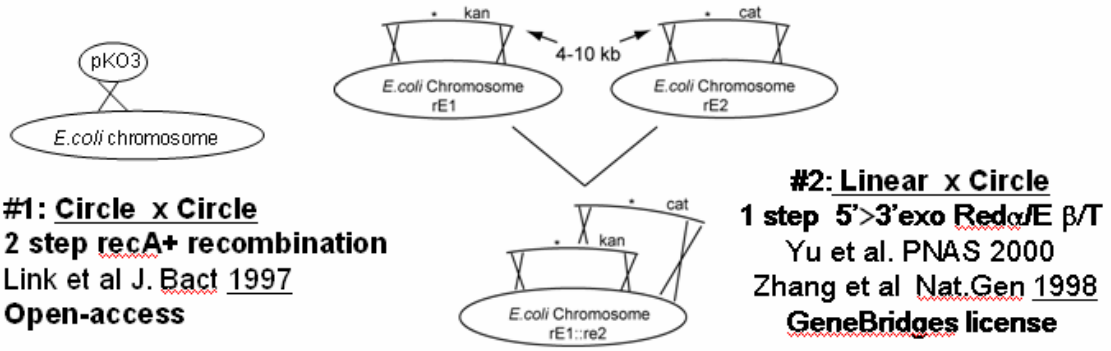
Project S7: Genome Engineering and the Construction of New Genetic Codes

(HMS: Church in collaboration with Jacobson lab at MIT). We are developing methods of engineering large numbers of specific changes into the *E. coli* genome and increasing the efficiency of homologous recombination. One goal is to replace all instances of a single codon (~4,000 substitutions) with a synonymous codon, a problem not tractable with conventional techniques. This codon could be used to generate proteins with non-native amino acids. Also, strains in which codons are reprogrammed to alternative amino acids may be insulated from genetic exchange with other organisms or viruses. Two approaches under development are (i) a large-scale DNA synthesis strategy in which Polymerase Assembly Multiplexing (Carr et al. 2004, Carr et al. 2007, Tian et al. 2004) is used to build a large pool (10^4-10^5) of short (100 bp) DNA fragments that are assembled to generate small pools (~50) of long (~100 kb) DNA fragments that are integrated into the genome by lambda red recombination (Copeland et al. 2001), and (ii) a strategy where large pools (10^2-10^4) of short DNA molecules containing desired

mutations are introduced directly into cells via lambda red beta protein, whereby many independent changes are introduced simultaneously into a population of cells and the whole population evolves towards one containing all the desired changes. We are also developing combinatorial DNA synthesis methods (Tian et al. 2004) that enable generation of high complexity libraries of alternative coding sequences which can be used to optimize metabolic pathways for biofuel production.

Aim S7.1: Conjugation. We are developing methods

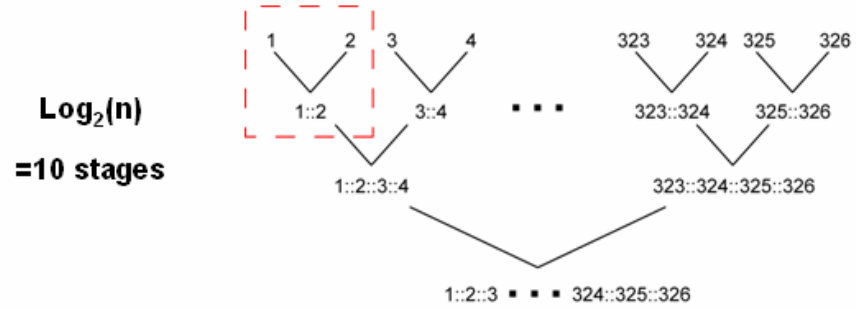
Aim S7.2: Automated homologous recombination. We are developing methods

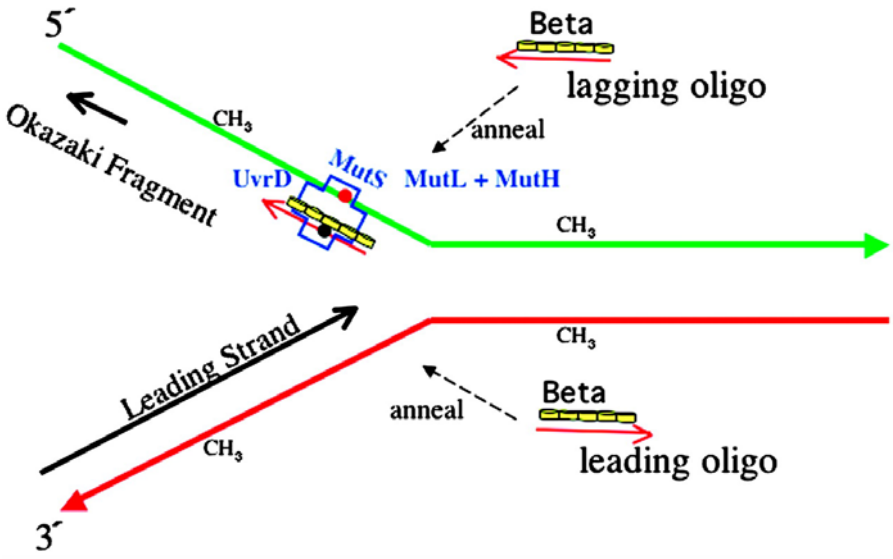


#1: Circle x Circle
2 step recA+ recombination
Link et al J. Bact 1997
Open-access

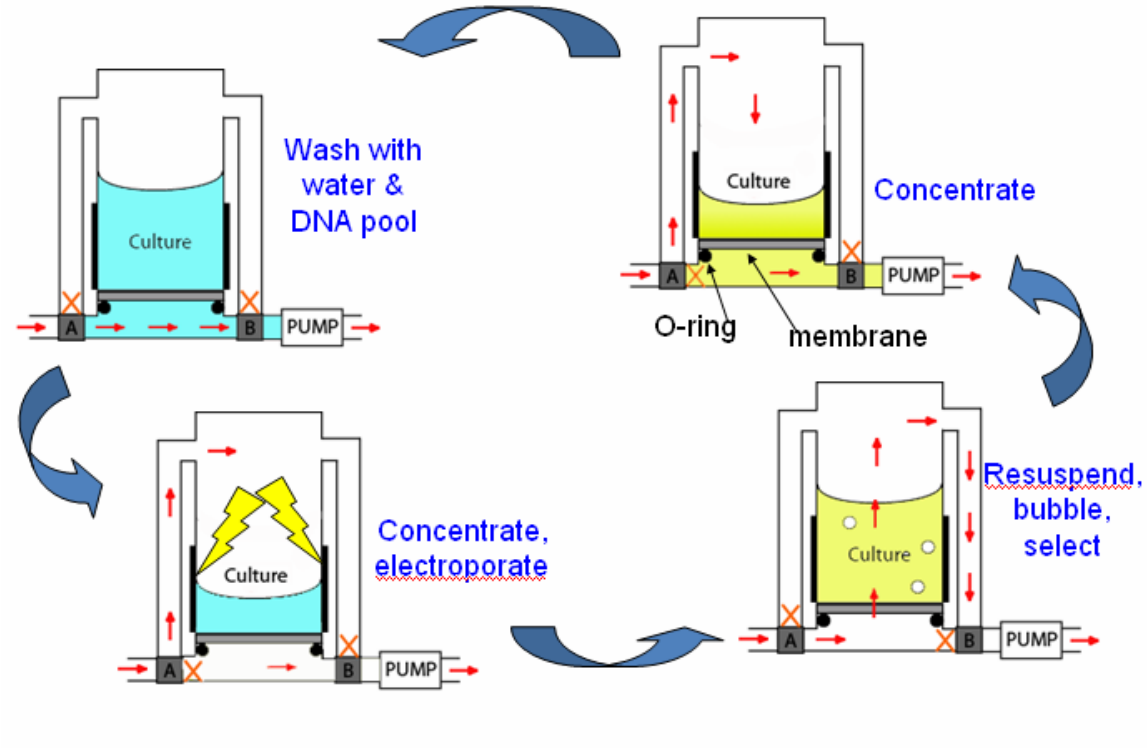
#2: Linear x Circle
1 step 5'>3'exo Redα/E β/T
Yu et al. PNAS 2000
Zhang et al Nat.Gen 1998
GeneBridges license

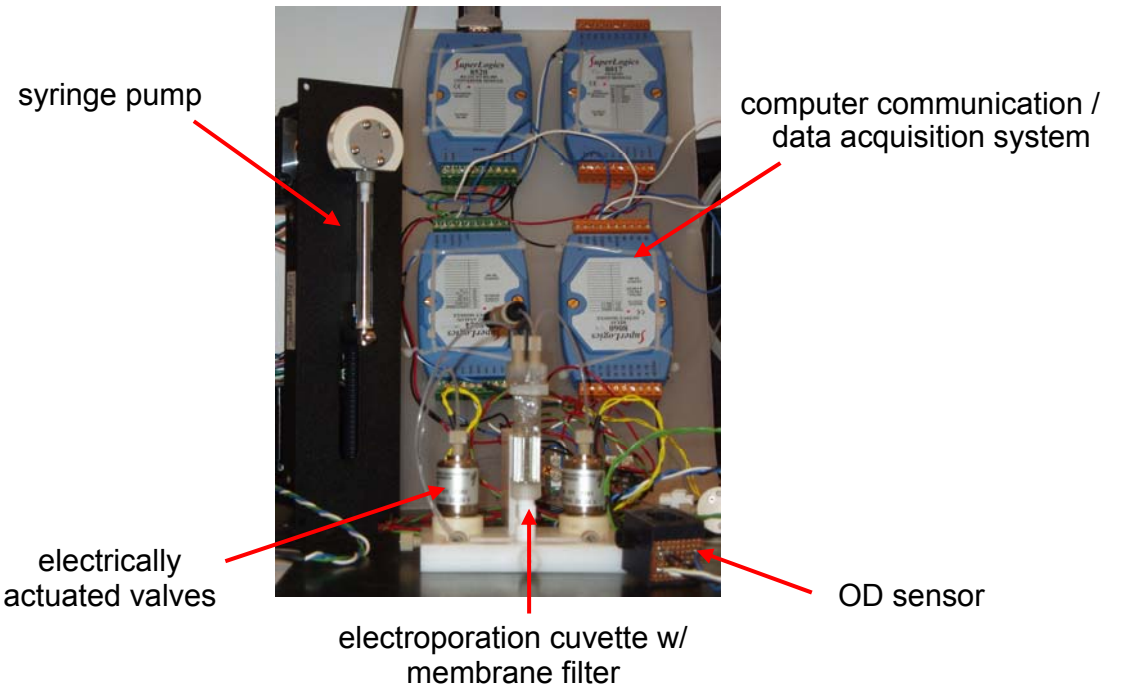
Hierarchical recombination-conjugation strategy

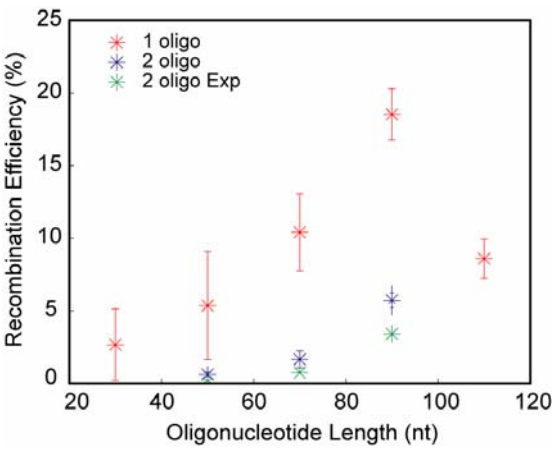


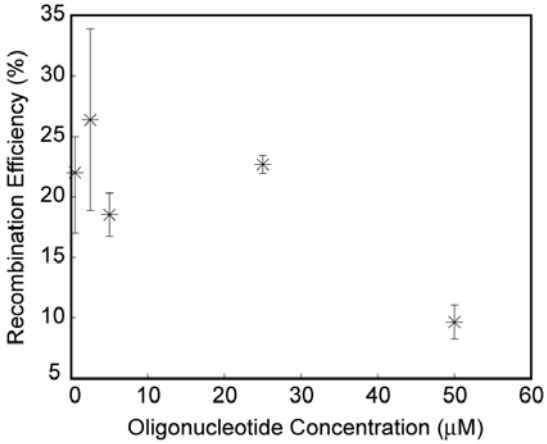


Multiplex Automated Genome Engineering (MAGE)

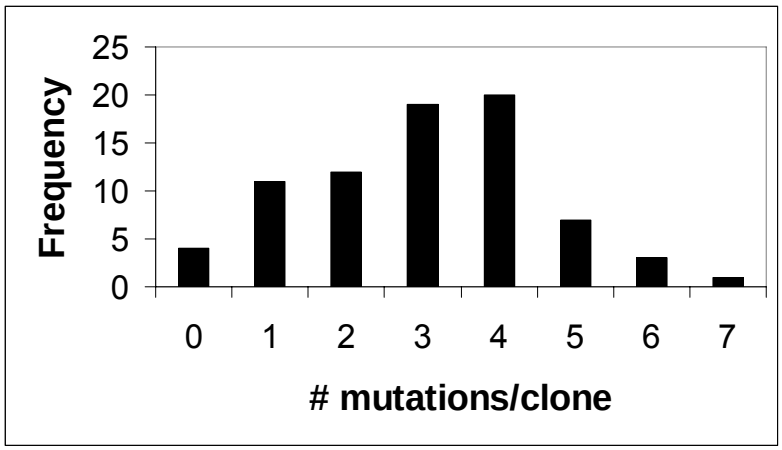








Impact of oligo length 90mer oligos are optimal. Two oligos synergistic High recombination frequencies from 0.25 to 25 mM oligo which permits high multiplexing. Obtain 25% recombination efficiency in *E. coli* strains lacking mismatch repair genes (*mutH, mutL, mutS, uvrD, dam)* (Ellis et al. 2001, Constantino & Court 2003)



Mutation Distribution: 11 oligos, 15 cycles

Oligo Pool	# cycles	Best Clone (98 %tile)	Fraction of mutated sites	Time
11	15	7	7/11	5 days
54	20	16 (estimate)	16/54	7 days

References for Project S7.

Copeland NG, Jenkins NA, Court DL. 2001. Recombineering: a powerful new tool for mouse functional genomics. *Nat Rev Genet* 2: 769-79

Tian J, Gong H, Sheng N, Zhou X, Gulari E, et al. 2004. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432: 1050-4

Carr PA, Kong D, Jacobson J. 2007. Synthetic Biology 2.0 (submitted) *Nucl. Acids Res.*

Carr PA, Park JS, Lee YJ, Yu T, Zhang S, Jacobson JM. 2004. Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res* 32: e162

Ellis et al. PNAS 2001

Constantino & Court. PNAS 2003

Bibliography of work done in our GTL Systems Biology grant 2003-2007.

1 **Aach** JA and **Church** GM. (2004) Mathematical models of diffusion-constrained polymerase chain reactions: basis of high-throughput nucleic acid assays and simple self-organizing systems. J Theor. Biol. May 7;228(1):31-46.

2 Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, **Polz** MF (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 71(12):8966-9.

3 Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., **Polz**, M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature. **430:**551-554**.**

4 Ahlgren NA, Rocap G, **Chisholm** SW. (2006) Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ Microbiol. 8(3):441-54.

5 Ahmad R, Nguyen DH, Wingerd MA, **Church** GM, Steffen MA. (2005) Molecular Weight Assessment of Proteins in Total Proteome Profiles Using 1D-PAGE and LC/MS/MS. Proteome Sci. 3(1):6

6 **Ausubel** FM. (2005) Are innate immune signaling pathways in plants and animals conserved? Nat Immunol. 6(10):973-9.

7 Bais HP, Prithiviraj B, Jha AK, **Ausubel** FM, Vivanco JM. Mediation of pathogen resistance by exudation of antimicrobials from roots. Nature. 2005 Mar 10;434(7030):217-21.

8 Baker, D, **Church G**, Collins, J, Endy, D, Jacobson, J, Jay Keasling, J, Modrich, P, Smolke, C, and Weiss, R (2006) Building a Fab for Biology. Scientific American June 2006

9 Bertilsson, S, Berglund, O, Pullin, MJ, **Chisholm, SW** (2005) Release of Dissolved Organic Matter by *Prochlorococcus*. Vie et Milieu-(Life & Environment*)* 55 (3-4):225-231

10 Bertilsson, S, Berglund, O, Karl, DM, and **Chisholm, SW** (2003) Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. Limnol. Oceanogr. 48: 1721-173

11 Bindschedler LV, Dewdney J, Blee KA, Stone JM, Asai T, Plotnikov J, Denoux C, Hayes T, Gerrish C, Davies DR, **Ausubel** FM, Paul Bolwell G. Peroxidase-dependent apoplastic oxidative burst in Arabidopsis required for pathogen resistance. Plant J. 2006 Sep;47(6):851-63.

12 Branda SS, Chu F, Kearns DB, Losick R, **Kolter** R. A major protein component of the *Bacillus subtilis* biofilm matrix. Mol Microbiol. 2006 Feb;59(4):1229-38.

13 Bush J, Jander G, **Ausubel** FM. Prevention and control of pests and diseases. Methods Mol Biol. 2006;323:13-25.

14 Chang JH, Urbach JM, Law TF, Arnold LW, Hu A, Gombar S, Grant SR, **Ausubel** FM, Dangl JL. A high-throughput, near-saturating screen for type III effector genes from *Pseudomonas syringae*. Proc Natl Acad Sci U S A. 2005 102:2549-54.

15 Choe SE, Boutros M, Michelson AM, **Church** GM, Halfon MS. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol. 2005;6(2):R16.

16 Chu F, Kearns DB, Branda SS, **Kolter** R, Losick R. Targets of the master regulator of biofilm formation in *Bacillus subtilis*. Mol Microbiol. 2006 Feb;59(4):1216-28.

17 **Church** GM. (2005) From systems biology to synthetic biology Mol Syst Biol. 1:32.

18 **Church** G. (2005) Let us go forth and safely multiply. Nature. 438(7067):423.

19 Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, **Chisholm** SW. Genomic islands and the ecology and evolution of *Prochlorococcus*. Science. 2006 Mar 24;311(5768):1768-70.

20 Cui J, Bahrami AK, Pringle EG, Hernandez-Guzman G, Bender CL, Pierce NE, **Ausubel** FM. *Pseudomonas syringae* manipulates systemic plant defenses against pathogens and herbivores. Proc Natl Acad Sci U S A. 2005 Feb 1;102(5):1791-6.

21 DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, **Chisholm** SW, Karl DM. Community genomics among stratified microbial assemblages in the ocean's interior. Science. 2006 Jan 27;311(5760):496-503.

22 Diener AC, **Ausubel** FM. Resistance To Fusarium Oxysporum 1, a dominant Arabidopsis disease-resistance gene, is not race specific. Genetics. 2005 Sep;171(1):305-21.

23 Dorrestein PC, Blackhall J, Straight PD, Fischbach MA, Garneau-Tsodikova S, Edwards DJ, McLaughlin S, Lin M, Gerwick WH, **Kolter** R, Walsh CT, Kelleher NL. Activity screening of carrier domains within nonribosomal peptide synthetases using complex substrate mixtures and large molecule mass spectrometry. Biochemistry. 2006 Feb 14;45(6):1537-46.

24 Dudley, A, Janse, D, **Church**, GM (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. Nature/EMBO Molecular Systems Biology msb4100004-E1-E11 (open access).

25 Follows, M, Dutkiewicz, S, Grant, S, **Chisholm, SW** (2007) Emergent biogeography of microbial communities in a model ocean. Science (in press)

26 Forster, AC & **Church**, GM (2007) Synthetic Biology Projects In Vitro Genome Research 17(1):1-6.

27 Forster, AC & **Church**, GM (2006) Toward Synthesis of a Minimal Cell. Nature-EMBO-Molecular Systems Biology 2:45.

28 Friedman L, **Kolter** R. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. Mol Microbiol. 2004 Feb;51(3):675-90.

29 Friedman L, **Kolter** R. Two genetic loci produce distinct carbohydrate-rich structural components of the Pseudomonas aeruginosa biofilm matrix. J Bacteriol. 2004 Jul;186(14):4457-65.

30 Gao,Y & **Church,** GM (2004) FLEXMOTIF: A Generic Flexible Motif Discovery Algorithm for Unaligned Sequences. Submitted to RECOMB2004.

31 Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, Lamerdin J, **Chisholm** SW. The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynth Res. 2001;70(1):53-71.

32 Hogan DA, Vik A, **Kolter** R. A Pseudomonas aeruginosa quorum-sensing molecule influences Candida albicans morphology. Mol Microbiol. 2004 Dec;54(5):1212-23.

33 Hunt DE, Klepac-Ceraj V, Acinas SG, Gautier C, Bertilsson S, **Polz** MF. Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. Appl Environ Microbiol. 2006 Mar;72(3):2221-5.

34 Jaffe JD, Berg, HC, **Church** GM (2004) Proteogenomic mapping reveals genomic structure and novel proteins undetected by computational algorithms. Proteomics 4(1):59-77

35 Jaffe, J, Mani, DR, Leptos, K, **Church**, GM, Carr SA (2006) PEPPeR: A platform for experimental proteomic pattern recognition. Mol Cell Proteomics. 2006 Jul 19;

36 Jaffe, JD, Stange-Thomann, N, Smith, C, DeCaprio, D, Sheila Fisher, S, Butler, J, Calvo, S, Elkins, T, FitzGerald, MG, Hafez, N, Kodira CD, Major J, Wang S, Wilkinson, J, Nicol, R, Nusbaum, C, Birren, B, Berg, HC, **Church** GM (2004) The complete genome and proteome of *Mycoplasma mobile*. Genome Research 14: 1447-1461. supplement

37 Janse, DM, Crosas,B Finley,D & **Church,** GM (2004) Localization to the Proteasome is Sufficient for Degradation. J. Biol Chem 279 (20):21415-20.

38 Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, **Chisholm** SW. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science. 311(5768):1737-40.

39 Johnson, Z.I. and **Chisholm,** S.W. (2004) Properties of Overlapping Genes are Conserved Across Microbial Genomes. *Genome Research* 14:2268-2272,

40 Kearns DB, Chu F, Branda SS, **Kolter** R, Losick R. A master regulator for biofilm formation by *Bacillus subtilis*. Mol Microbiol. 2005 Feb;55(3):739-49.

41 Kharchenko P, Chen L, Freund Y, Vitkup D, **Church** GM. Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics. 2006 Mar 29;7(1):177

42 Kharchenko, P, **Church,** GM, Vitkup, D. (2004) Filling gaps in a metabolic network using expression information. Bioinformatics Aug 4;20 Suppl 1:I178-I185.

43 Kharchenko, P, Vitkup, D & **Church**, GM (2005) Expression dynamics of a cellular metabolic network. Nature MSB doi:10.1038/msb4100023, 2 August 2005

44 King, OD, Lee, JC, Dudley, AM, Janse, DM, **Church,** GM, Roth, FP (2003) Predicting Phenotype from Patterns of Annotation. ISMB 2003.

Kolter R, Greenberg EP. Microbial sciences: the superficial life of microbes. Nature.
45 2006 May 18;441(7091):300-2.

Kulasakara H, Lee V, Brencic A, Liberati N, Urbach J, Miyata S, Lee DG, Neely
46 AN, Hyodo M, Hayakawa Y, **Ausubel** FM, **Lory** S. Analysis of Pseudomonas aeruginosa diguanylate cyclases and phosphodiesterases reveals a role for bis-(3'-5')-cyclic-GMP in virulence. Proc Natl Acad Sci U S A. 2006 Feb 21;103(8):2839-44.

Kuo W, Liu F, Jenssen T, Trimarchi J, Punzo C, Lombardi M, Sarang J, Maysuria M, Serikawa K, Lee SY, McCrann D, Kang J, Shearstone J, Burke J, Park D, Choi S,
47 Perrin S, **Church** GM, Bumgarner R, Cepko, C (2006) A sequence oriented comparison of gene expression measurements across different hybridization-based technologies. Nature Biotech. Jul;24(7):832-840.

Lee SI, Pe'er D, Dudley AM, **Church** GM, Koller D. (2006) Identifying regulatory
48 mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci USA 103(38):14062-7.

Lee DG, Urbach JM, Wu G, Liberati NT, Feinbaum RL, Miyata S, Diggins LT, He J,
49 Saucier M, Deziel E, Friedman L, Li L, Grills G, Montgomery K, **Kucherlapati** R, Rahme LG, **Ausubel** FM. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol. 2006 Oct 12;7(10):R90

Lee, D.G., N.T. Liberati, J.M. Urbach, G. Wu, and F.M. **Ausubel** (2007) Modeling
50 microbial virulence in a genomic era: The impact of shared genomic tools and datasets. In: Bacterial Pathogenomics, ASM Press, Washington, DC, in press.

Lee, D.G., J.M. Urbach, G. Wu, N.T. Liberati, R.L. Feinbaum and F.M. **Ausubel**
51 (2007) Combining genomic tools to dissect multifactoral virulence in *Pseudomonas aeruginosa*. Stadler Symposium Volume, in press.

52 Leptos, KC, Sarracino, DA, **Church**, GM (2006) MapQuant: Open-Source Software for Large-Scale Protein Quantitation. Proteomics 6(6):1770-82.

Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei
53 T, **Ausubel** FM. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A. 2006 Feb 21;103(8):2833-8.

Liberati, N.T., J.M. Urbach, T.K. Holmes, G. Wu and F.M. **Ausubel** (2006)
54 Comparing insertion libraries in two *Pseudomonas aeruginosa* strains to assess gene essentiality. In: Methods in Molecular Biology, Humana Press, in press.

Lindell, D, Jaffe, JD, Coleman, MI, Axmann, IM, Rector, T, Kettler, G, Sullivan,
55 MB, Steen, R, Hess, WR, **Church, GM, Chisholm, SW.** (2007) Genome-wide expression dynamics of a marine virus and its host reveal features of coevolution. Nature (in revision)

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, **Chisholm** SW.
56 Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci U S A. 2004 Jul 27;101(30):11013-8.

Lindell, D, Jaffe, JD, Johnson, ZI, **Church**, GM & **Chisholm**, SW (2005)
57 Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86-9.

Luyten YA, Thompson JR, Morrill W, **Polz** MF, Distel DL. Extensive variation in
58 intracellular symbiont community composition among members of a single population of the wood-boring bivalve *Lyrodus pedicellatus* (*Bivalvia: Teredinidae*). Appl Environ Microbiol. 2006 Jan;72(1):412-7.

Madan R, **Kolter** R, Mahadevan S. Mutations that activate the silent bgl operon of
59 *Escherichia coli* confer a growth advantage in stationary phase. J Bacteriol. 2005 Dec;187(23):7912-7.

Marcelino LA, Backman V, Donaldson A, Steadman C, Thompson JR, Preheim SP,
60 Lien C, Lim E, Veneziano D, **Polz** MF. Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. Proc Natl Acad Sci U S A. 2006 Sep 12;103(37):13629-34.

Martiny AC, Coleman ML, **Chisholm** SW. Phosphate acquisition genes in
61 *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. Proc Natl Acad Sci U S A. 2006 Aug 15;103(33):12552-7.

Mikkilineni V, Mitra RD, Merritt J, DiTonno JR, **Church** GM, Ogunnaike B,
62 Edwards JS. Digital quantitative measurements of gene expression. Biotechnol Bioeng. 2004 Apr 20;86(2):117-24.

Mitra, RD, Butty, V, Shendure, J, Williams, BR, Housman, DE, and **Church,** GM
63 (2003) Digital Genotyping and Haplotyping with Polymerase Colonies. Proc Natl Acad Sci USA. May 13;100(10):5926-31.

64 Mitra,RD, Shendure,J, Olejnik,J, Olejnik,EK, and **Church,**GM (2003) Fluorescent in situ Sequencing on Polymerase Colonies. Analyt. Biochem. 320:55-65

Moore, LR, Coe, A, Zinser, ER, Saito, MA, Sullivan, MB, Lindell, D, Frois-Moniz,
65 K, Waterbury, J, **Chisholm, SW** (2007) Culturing the marine cyanobacterium *Prochlorococcus*. Limnol. Oceanogr. (submitted)

Morikawa M, Kagihiro S, Haruki M, Takano K, Branda S, **Kolter** R, Kanaya S.
66 Biofilm formation by a *Bacillus subtilis* strain that produces gamma-polyglutamate. Microbiology. 2006 Sep;152(Pt 9):2801-7.

Oates PM, Castenson C, Harvey CF, **Polz** M, Culligan P. Illuminating reactive
67 microbial transport in saturated porous media: demonstration of a visualization method and conceptual transport model. J Contam Hydrol. 2005 May;77(4):233-45.

Perry TD 4th, Klepac-Ceraj V, Zhang XV, McNamara CJ, **Polz** MF, Martin ST,
68 Berke N, Mitchell R. Binding of harvested bacterial exopolymers to the surface of calcite. Environ Sci Technol. 2005 Nov 15;39(22):8770-5.

Petti,A & **Church**, GM (2005) A Network of Transcriptionally Coordinated
69 Functional Modules in Saccharomyces cerevisiae. Genome Research Sep;15(9):1298-306.

Randa MA, **Polz** MF, Lim E. Effects of temperature and salinity on Vibrio vulnificus
70 population dynamics as assessed by quantitative PCR. Appl Environ Microbiol. 2004 Sep;70(9):5469-76.

Reppas NB, Wade JT, Church GM, Struhl K. The transition between transcriptional
71 initiation and elongation in E. coli is highly variable and often rate limiting. Mol Cell. 2006 Dec 8;24(5):747-57.

Rindone, W, Gong, H & **Church,** GM (2006) CADPAM gene/genome synthetic
72 design software (http://arep.med.harvard.edu/cgi-bin/cadpam/worktest/cadpamworks.pl)

Rocap G, Larimer F, Lamerdin J, Malfatti S, Chain P, Ahlgren N, Arellano A,
73 Coleman M, Hauser L, Hess W, Johnson Z, Land M, Lindell D, Post A, Regala W, Shah M, Shaw S, Steglich C, Sullivan M, Ting C, Tolonen A, Webb E, Zinser E, **Chisholm** S (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. (2003) Nature Aug 28;424(6952):1042-7.

Romano JD, **Kolter** R. Pseudomonas-Saccharomyces interactions: influence of
74 fungal metabolism on bacterial physiology and survival. J Bacteriol. 2005 Feb;187(3):940-8.

Segre, D, DeLuna, A, **Church,** GM, Kishony, R (2005) Modular epistasis in yeast
75 metabolism. Nat Genet. 37(1):77-83.

76 Segre, D, Vitkup, D, and **Church,** GM (2002) Analysis of optimality in natural and perturbed metabolic networks . Proc. Nat. Acad. Sci USA 99: 15112-7.

Segre, D, Zucker, J, Katz, J, Lin, X, D'haeseleer, P, Rindone, W, Karchenko, P,
77 Nguyen, D, Wright, M, and **Church**, GM From annotated genomes to metabolic flux models and kinetic parameter fitting. (submitted Jun-2003) Omics in press.

Selinger, DW, Saxena, RM, Cheung, KJ, **Church,** GM, and Rosenow, C (2003)
78 Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation . Genome Research Feb;13(2):216-23.

Shaw, SL, **Chisholm, SW**, Prinn, RG (2003) Isoprene Production by
79 *Prochlorococcus*, a Marine Cyanobacterium, and Other Phytoplankton. Marine Chemistry 80:227–245

80 Shendure J, Mitra R, Varma C, **Church** GM (2004) Advanced Sequencing Technologies: Methods and Goals. Nature Reviews of Genetics May;5(5):335-44.

Shendure, J, Porreca, GJ, Reppas, NB, Lin, X, McCutcheon, JP, Rosenbaum, AM,
81 Wang, MD , Zhang, K, Mitra, RD, **Church**, GM (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome Science 309(5741):1728-32.

82 Steffen, M, Jaffe, JD, & **Church,** GM (2003) Analysis of DNA-Binding Proteins by Mass Spectrometry. Submitted.

83 Steglich C, Futschik M, Rector T, Steen R, **Chisholm** SW. Genome-wide analysis of light sensing in *Prochlorococcus*. J Bacteriol. 2006 Sep 15;

84 Straight PD, Willey JM, **Kolter** R. (2006) Interactions between Streptomyces coelicolor and Bacillus subtilis: Role of surfactants in raising aerial structures. J Bacteriol.188(13):4918-25.

85 Su-In Lee, S-I, Pe'er, D, Dudley, AM, **Church**, GM, Daphne Koller D (2006) Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. Proc Natl Acad Sci USA 103(38):14062-7

86 Sullivan MB, Coleman ML, Weigele P, Rohwer F, **Chisholm** SW. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. PLoS Biol. 2005 May;3(5):e144.

87 Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, **Chisholm** SW. Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. PLoS Biol. 2006 Jul 4;4(8)

88 Sullivan, M.B., J. Waterbury, and S.W. **Chisholm**. (2003) Cyanophage infecting the oceanic cyanobacterium, Prochlorococcus. Nature Aug 28;424(6952):1047-1051.

89 Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, **Polz** MF. Genotypic diversity within a natural coastal bacterioplankton population. Science. 2005 Feb 25;307(5713):1311-3.

90 Thompson JR, Randa MA, Marcelino LA, Tomita-Mitchell A, Lim E, **Polz** MF. Diversity and dynamics of a north atlantic coastal Vibrio community. Appl Environ Microbiol. 2004 Jul;70(7):4103-10.

91 Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, & **Church** GM (2004) Accurate Multiplex Gene Synthesis from Programmable DNA Chips. Nature 432: 1050-4.

92 Tolonen AC, **Aach** J, Lindell D, Johnson ZI, Rector T, Steen R, **Church** GM, **Chisholm** SW. Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol. 2006;2:53.

93 Tolonen, AC, Liszt, GB, and Hess, WR (2006) Genetic manipulation of *Prochlorococcus* MIT9313: GFP expression on an RSF1010 plasmid and Tn5 transposition. Appl. Environ. Microbiol. doi:10.1128/AEM.02034-

94 Tompa M, Li N, Bailey TL, **Church** GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z.(2005) An Assessment of Computational Tools for the Discovery of Transcription Factor Binding Sites. Nature Biotechnology 23:137-44.

95 Urbach, J.M., G. Wu, D.G. Lee, N.T. Liberati, and F.M. **Ausubel** (2006) P. aeruginosa PA14 Genome Sequence Public Database: http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/view_gene.cgi

96 Urbach, J.M., G. Wu, D.G. Lee, N.T. Liberati, and F.M. **Ausubel** (2006) *P. aeruginosa* PA14 Transposon Insertion Mutation Library Public Database: http://ausubellab.mgh.harvard.edu/cgi-bin/pa14/home.cgi

97 Vlamakis HC, **Kolter** R. Thieves, assassins and spies of the microbial world. Nat Cell Biol. 2005 Oct;7(10):933-4.

98 Wade JT, Reppas NB, **Church** GM, Struhl K. (2005) Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites. Genes Dev. 19(21):2619-30.

99 Weaver VB, Kolter R. Burkholderia spp. alter Pseudomonas aeruginosa physiology through iron sequestration. J Bacteriol. 2004 Apr;186(8):2376-84.

100 Wright, MA, Kharchenko, P, **Church**, GM, Segre, D (2007) Chromosome folding from evolutionary constraints. In revision at PNAS

101 Yin J, Straight PD, McLoughlin SM, Zhou Z, Lin AJ, Golan DE, Kelleher NL, **Kolter** R, Walsh CT. Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase. Proc Natl Acad Sci U S A. 2005 102(44):15815-20.

102 Zambrano MM, **Kolter** R. Mycobacterial biofilms: a greasy way to hold it together. Cell. 2005 Dec 2;123(5):762-4.

103 Zhang, K, Martiny, AC, Reppas, NB, Barry, KW, Malek, J, **Chisholm**, SW, **Church**, GM (2006) Sequencing genomes from single cells via polymerase clones. Nature Biotech. Jun;24(6):680-6.

104 Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, Scanlan DJ, **Chisholm** SW. *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol. 2006 72(1):723-32.

105 Zinser, ER, Johnson, ZI, Coe, Karaca, AE, Veneziano, D, **Chisholm, SW** (2007) Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol. Oceanogra. (in revision)

106 Kim JB, Porreca GJ, Song L, Greenway S, Gorham JM, **Church** GM, Seidman CE, Seidman JG (2007) Deep sequencing analysis of gene expression in disease pathogenesis. Science (in revision).