DOE Joint Genome Institute FY15 Q2 Progress Report

Q2. Report on progress developing new experimental capabilities to analyze complex genomic or metagenomic datasets.

Background

The functional annotation of genomes remains an extremely difficult problem that is exacerbated by the continuing accumulation of new genomes encoding a large number of new genes of unknown function. Computational prediction of gene function based on DNA sequence similarity to known genes is successful when closely related, well-annotated model organisms are available. However, a scarcity of functionally well-characterized microbes means that the genomes of many organisms, including those with important roles in bioenergy, the environment and human health, remain poorly annotated. There is therefore a great demand for novel high-throughput approaches to determine gene function from phylogenetically diverse microbes.

Transposon mutagenesis and sequencing (TnSeq) is a promising approach for high-throughput annotation of gene function (**Figure 1**). In this approach, mutagenesis of an bacterium of interest gives rise to a population (or pool) in which each cell has a single random transposon insertion, and which cumulatively contains at least one insertion for every gene in the genome. Transposons used in this strategy contain ‘tags’, which can be identified by DNA sequencing, and used to track changes in the relative abundance of thousands of mutant strains in a single tube reaction. By repeating the experiment under a variety of growth conditions, the combined patterns of mutant fitness across conditions can be used to infer a first-order functional annotation for many genes in the genome.

We previously demonstrated that transposon-based mutant fitness provides specific functional annotations for a substantial number of genes with previously unknown function (Deutschbauer et al., 2011). Over the last year, we have developed and applied new technologies for the functional genomic annotation of microbes, including a subset with direct relevance to DOE missions of bioenergy and carbon sequestration.

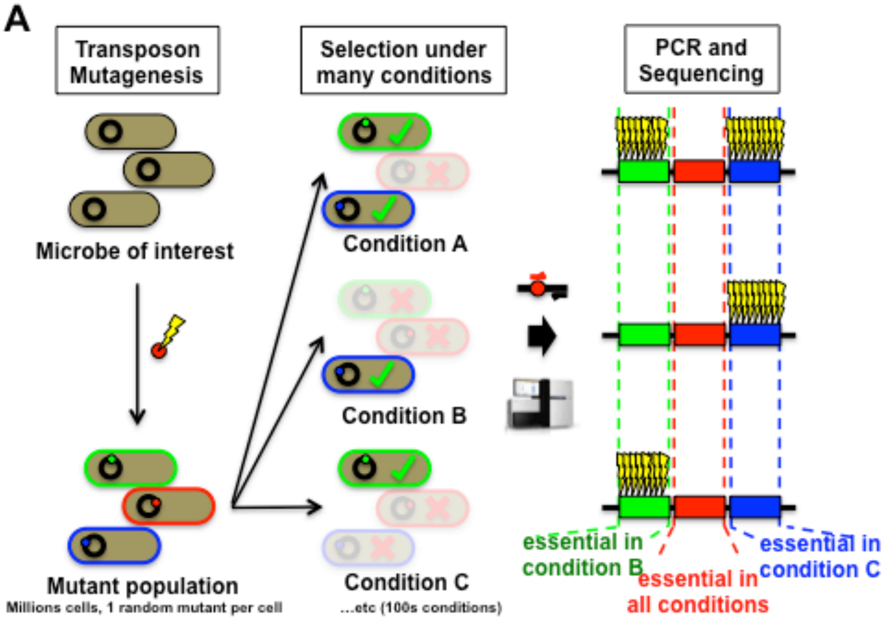

Figure 1. General strategy for gene function annotation using transposon mutagenesis and sequencing (Tn-Seq). i) Transposon mutagenesis generates complex mutant populations with at least one inactivating mutation for every gene in the genome. ii) Mutant populations are grown under a diverse set of growth conditions. iii) Changes in the abundance of individual strains are determined by high throughput sequencing and used to determine gene function. An absence of recovered reads across all conditions (e.g. red gene) indicates that a gene is essential for survival. An absence of recovered reads only after growth under specific conditions provides phenotype information for the gene (e.g. blue gene is essential for growth under condition C).

Progress

In the past year, substantial progress has been made in the development of the TN-seq method at JGI. This progress is described below.

1. High throughput gene function annotation using RB-TnSeq

There are numerous implementations of TnSeq–like approaches to study microbial genome function (van Opijnen and Camilli, 2013). However, existing approaches involve a large number of experimental steps for each sample processed (DNA shearing, DNA end repair, adapter ligation, PCR, with multiple intermediate purification steps), and become prohibitively time-consuming and expensive when applied across more than a handful of conditions or samples.

To enable the testing of transposon mutant populations across hundreds of experimental conditions we developed a new method, random barcode transposon-site sequencing (RB-TnSeq). This approach combines the advantages of TnSeq (large numbers of mutant strains with no archiving) with DNA "barcoding" (easy and scalable quantification). In this approach, bacteria are mutagenized using a randomly barcoded transposon (**Figure 2**). The resulting transposon mutant library only needs to be characterized by a one-time TnSeq experiment to link the transposon insertion location to the DNA barcode incorporated in the transposon at each insertion site. All subsequent experiments to assay strain and gene fitness in a competitive growth assay are performed by sequencing only the barcodes to determine their relative abundance (BarSeq). This modification makes the technique simpler, less expensive, and scalable.

We applied RB-TnSeq to study gene function in five bacteria, grown across a total of 800 experimental conditions. This study demonstrated that RB-TnSeq is highly reproducible, and results in phenotype information in agreement with other technologies and consistent with known gene functions. A manuscript describing this technique and showing its reproducibility and scalability was recently submitted for publication.

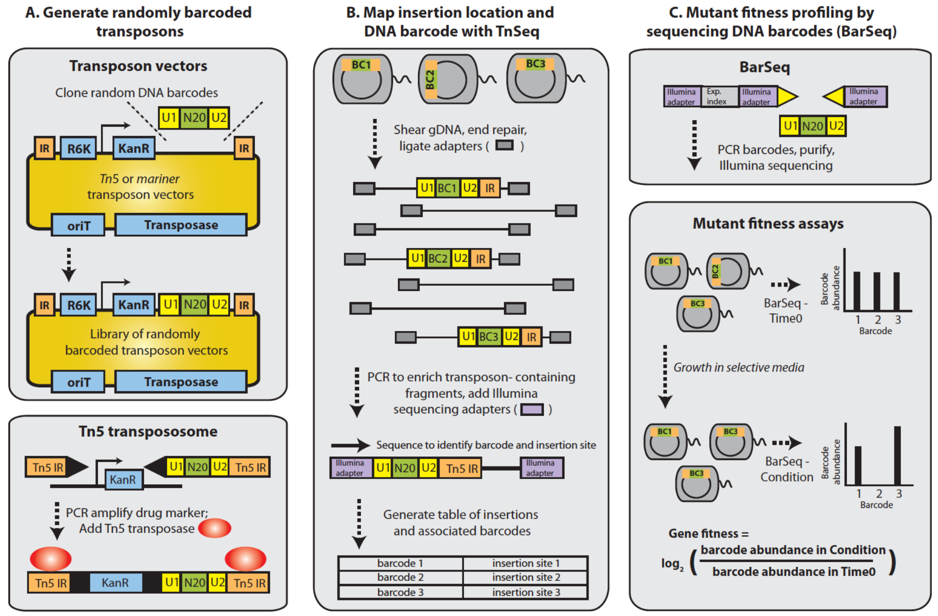
A. Generate randomly barcoded transposons
Transposon vectors
Clone random DNA barcodes
U1 N20 U2
IR R6K KanR IR
Tn5 or mariner transposon vectors
oriT Transposase
IR R6K KanR U1 N20 U2 IR
Library of randomly barcoded transposon vectors
oriT Transposase
Tn5 transposome
Tn5 IR KanR U1 N20 U2 Tn5 IR
PCR amplify drug marker; Add Tn5 transposase
Tn5 IR KanR U1 N20 U2 Tn5 IR
B. Map insertion location and DNA barcode with TnSeq
BC1 BC2 BC3
Shear gDNA, end repair, ligate adapters ( )
U1 BC1 U2 IR
U1 BC2 U2 IR
U1 BC3 U2 IR
PCR to enrich transposon- containing fragments, add Illumina sequencing adapters ( )
Sequence to identify barcode and insertion site
Illumina adapter U1 N20 U2 Tn5 IR Illumina adapter
Generate table of insertions and associated barcodes
barcode 1 insertion site 1
barcode 2 insertion site 2
barcode 3 insertion site 3
C. Mutant fitness profiling by sequencing DNA barcodes (BarSeq)
BarSeq
Illumina adapter Exp. index Illumina adapter Illumina adapter
U1 N20 U2
PCR barcodes, purify, Illumina sequencing
Mutant fitness assays
BC1 BC2 BC3
BarSeq - Time0
Barcode abundance
1 2 3
Barcode
Growth in selective media
BC1 BC3 BC3
BarSeq - Condition
Barcode abundance
1 2 3
Barcode
Gene fitness =
log₂ (barcode abundance in Condition / barcode abundance in Time0)

Figure 2. Overview of RB-TnSeq. (**A**) RB-TnSeq vectors are created from transposon vectors (top) or transposomes (bottom) by cloning millions of unique DNA barcodes (N20) flanked by common PCR priming sites (P1 and P2) near the edge of the transposon's inverted repeat (IR). These are used to generate a large transposon mutant population such that each strain contains a unique DNA barcode. (**B**) A randomly barcoded transposon mutant library is characterized using "TnSeq". Both the DNA barcode and the transposon insertion site are identified in a single 150 nucleotide Illumina sequencing read, resulting in a table of barcodes and associated transposon insertion locations. (**C**) Top – In BarSeq, DNA barcodes are PCR amplified using indexed oligos that bind the common P1 and P2 regions, and subject to illumina sequencing. Bottom – Mutant fitness is determined by comparing the abundance of the DNA barcodes with BarSeq before (Time0) and after selective growth (Condition). In this simple example, the gene associated with barcode 2 has reduced fitness; the gene associated with barcode 3 has enhanced fitness.

2. Functional Encyclopedia of Bacteria and Archaea (FEBA) project

To test the utility of the RB-TnSeq approach in providing novel gene function information, we are carrying out a large-scale project named 'Functional Encyclopedia of Bacteria and Archaea' (FEBA). The goal of this project is to generate transposon mutant populations from 20 diverse prokaryotes, and obtain gene function information through profiling these mutants under high-throughput growth assays.

To date, we have generated transposon mutant populations from 20 bacteria comprising 19 proteobacteria (with representatives of each class), and one cyanobacteria. These mutant populations contain on average 100,000 mutant strains, corresponding to ~20-40 mutants for the average gene. For high throughput phenotyping of these mutant populations, we compiled a 96-well plate based panel of selective growth media comprising 48 different carbon sources, 48 different nitrogen sources and 96 different small-molecule stresses (antibiotics, etc). Mutant populations have been tested across an average of 80 experimental conditions each.

To evaluate the success of this approach in annotating gene function, we compiled the results from all mutagenesis and high throughput phenotyping assays, and determined the proportion of genes from each FEBA organism for which we are able to provide some kind of functional information (Figure 3). On average, 15% of genes in each genome were never seen in mutant libraries, and are thus classified as being 'essential' (i.e. TN insertions in these genes are lethal). A further ~15% of genes are associated with a major growth phenotype under at least one experimental condition. Collectively these data provide phenotypic information for over 10,000 genes, of which ¼ had no previous functional information.

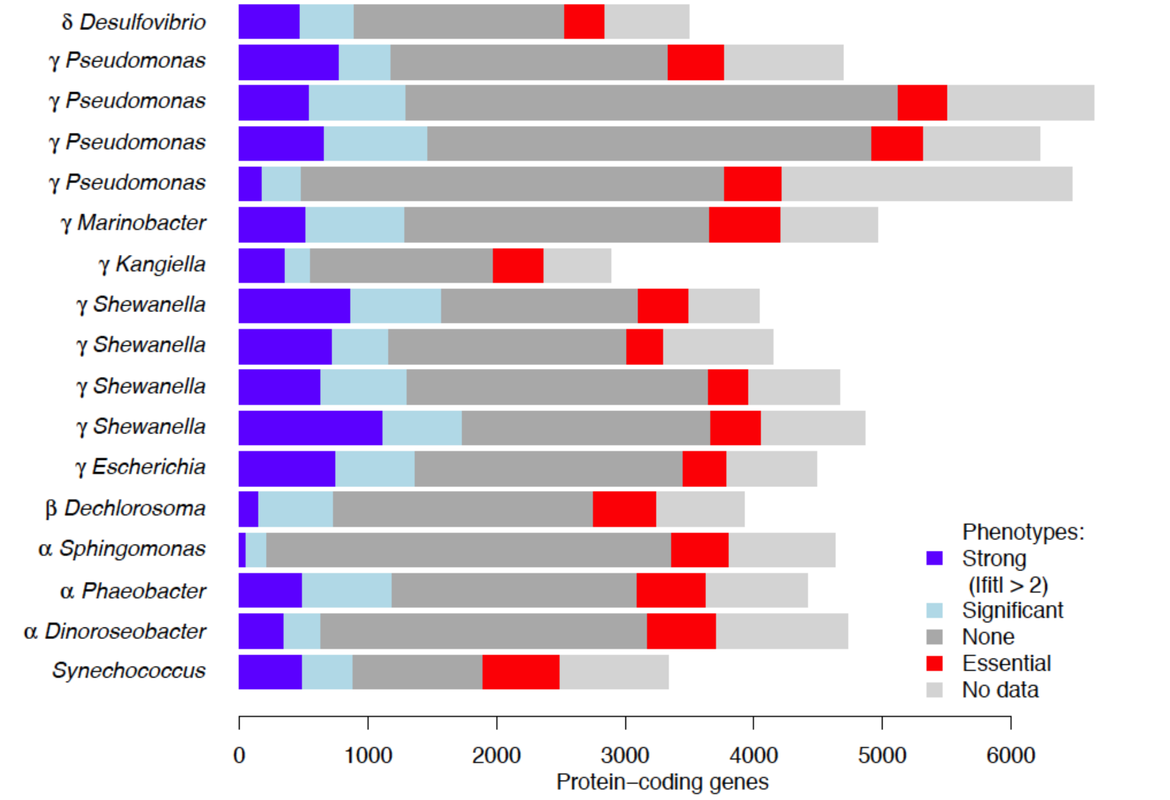

Figure 3: **Overview of FEBA phenotypes**. For each bacterium, we show its division, its genus name, and the number of protein-coding genes with different types of phenotypes information.

The FEBA dataset represents the largest collection of experimentally derived gene function information for bacteria currently available. Furthermore, the availability of phenotype information from multiple organisms tested under the same set of experimental conditions provides us with a unique opportunity for large-scale comparative analyses of gene function. Firstly, by comparing phenotypic data of orthologous genes across organisms, we can use conserved patterns to generate more confident and specific hypotheses of gene function that can be tested in follow up experiments. Secondly, we can evaluate the correlation between gene sequence conservation and gene function conservation. This is important as current automated approaches to genome annotation are based on the assumption that genes with similar DNA sequences have the same functions.

3. New technologies and applications for gene function annotation

i) Development of flexible molecular resources for annotation of diverse gene functions. An important goal of this project is to obtain gene function information from phylogenetically diverse organisms. The RB-TnSeq technology is therefore designed to be easily customizable in order to provide potential access to more diverse organisms. Currently the technology can be adapted to include different transposon vectors (Tn5, mariner), combined with different antibiotic resistance cassettes (Kanamycin, Chloramphenicol, etc), and with different delivery systems (conjugation, electroporation).

ii) In vivo assays of transposon mutant populations. The functional assays described in this report involve growth of transposon mutant populations in liquid culture. However, these studies provide phenotype information for only a subset of genes in the genome, and additional assays that reflect the natural environment of the organism will be required for more comprehensive annotation of gene function. We have recently explored the application of mutant libraries to study extracellular processes (by growing physically isolated mutants in solid culture), and in *in vivo* plant-root systems.

iii) High throughput isolation of individual gene-knockout strains. In the experiments described here, transposon mutants exist in a single pooled population. However, for many applications, it would be beneficial to have access to individual mutant strains (individual gene knockouts). To this end, we are developing approaches to array the pooled transposon mutant populations using automated colony picking of random clones, followed by Bar-Seq to characterize and identify individually isolated mutants. If successful, this may be routinely applied to every mutagenized population.

4. User program

An important goal of the JGI is to make TnSeq based gene function annotation capabilities available to the user community. In recent years, we have worked on a number of collaborative transposon sequencing projects through the CSP program (Christen et al, Moran et al, Xiao et al). However, these were ad-hoc projects using custom sequencing technology of limited scope. We are therefore developing two new ways in which users can access RB-TnSeq capabilities:

i) Provision of existing transposon mutant populations to user community. The 20 existing transposon mutant populations generated as part of the FEBA project, along with any future mutants we generate, are biological resources that can be used for future gene annotation studies. As an example, the *Pseudomonas fluorescens* strain mutagenized as part of FEBA was subsequently used in follow up studies to identify phosphate-solubilizing genes. We will likewise make all transposon mutant populations potentially available to the JGI user community through a dedicated CSP call.

ii) *Generation of novel transposon mutant populations*. Based on proposals from JGI collaborators, we will mutagenize 5 new organisms per year. These will be selected based on scientific merit, phylogenetic novelty, evidence that the organism is transformable, and availability of a suitable scientific assay for screening the transposon mutants. We will perform high throughput phenotyping of these mutant populations (carbon sources and Nitrogen sources) in addition to providing them to the collaborator for testing under specific experimental conditions.

DOE JGI / LBNL team

JGI: Matthew Blow, Cindi Hoover, James Bristow
LBNL: Adam Deutschbauer, Morgan Price, Jordan Waters, Kelly Wetmore, Adam Arkin.

External collaborators

Susan Golden (UCSD)
John Coates (LBNL)
Yongqin Jiao (LLNL)
Mary Anne Moran (University of Georgia)
Beat Christen (Institute of Molecular Systems Biology)

References

Deutschbauer, A., Price, M.N., Wetmore, K.M., Shao, W., Baumohl, J.K., Xu, Z., Nguyen, M., Tamse, R., Davis, R.W., and Arkin, A.P. (2011). Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS genetics *7*, e1002385.

van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nature reviews Microbiology *11*, 435-442.