Allegheny-Singer Research Institute

Annual Progress Report: 2009 Formula Grant

Reporting Period

January 1, 2010 – June 30, 2010

Formula Grant Overview

The Allegheny-Singer Research Institute received $186,525 in formula funds for the grant award period January 1, 2010 through June 30, 2011. Accomplishments for the reporting period are described below.

Research Project 1: Project Title and Purpose

16S FISH-based FACS Purification of Unculturable Bacteria for Whole Genome Amplification and Sequencing - It has recently become evident that most bioactive secondary metabolites isolated from marine invertebrates such as sponges, bryozoans, and tunicates are actually produced by their bacterial symbionts as part of mutual defense systems. The purpose of the planned research is to obtain pure populations of target unculturable microbial symbionts that will then be used for the generation of species-specific whole genome sequences. From these we can clone the biosynthetic pathways into genetically tractable organisms for the inexpensive engineered production of these medically important compounds. To obtain the single species bacterial cell populations, we will use a combination of metagenomics, deep 16S sequencing, 16S fluorescent *in situ* hybridization, fluorescence-activated cell sorting, whole genome amplification, and whole genome sequencing.

Anticipated Duration of Project

1/1/2010 – 6/30/2011

Project Overview

Once a novel unculturable bacterial species is chosen for sequencing based on matches between: 1) its 16S ribosomal sequence and 2) species-specific sequence features found in assembled metagenomic contigs (which encode parts of the target biosynthetic pathways), we will use a suite of bacterial cell isolation techniques including bacterial fluorescence-activated cell sorting (FACS) to obtain purified populations of the target bacteria for whole genome DNA sequence analysis. These FACS-purified populations of as few as 50-100 bacterial cells will then be subjected to whole genome amplification (WGA) and whole genome sequencing (WGS) to obtain the entire set of genes encoding the biosynthetic pathways that encode the production of the bioactive small molecule of interest. FACS will be accomplished by the production and use of species-specific fluorescently-tagged 16S rDNA probes that will be designed based on the sequences obtained from the deep 16S sequencing. These probes will then be vetted using

standard confocal laser scanning microscopy and 16S fluorescence *in situ* hybridization (FISH)-based imaging to establish their ability to specifically label a subpopulation of the bacterial cells within the invertebrate host. Once probes are shown to identify a population of bacteria within the host, they will be used to label ethanol-fixed single-cell suspensions made from lysates of the symbiome (host and associated microbiome) for FACS in order to obtain highly enriched bacterial cell populations for WGA and WGS. Briefly, we will prepare single cell bacterial suspensions by bead beating the host organism with large [~1 mM] diameter beads (these will not lyse the bacteria). These lysates will be treated with DNAse and proteinase to reduce viscosity and then gently sonicated for 5 minutes (using a water bath sonicator). Remaining host cells will be pelleted by low speed centrifugation and the bacterial-containing supernatant will be FISH-labeled using the specific individual bacterial species probes developed and vetted above. Sorting will be carried out using the new 4-laser FACSAria II custom sorter (BD Bioscience). We find that a simple lysis step consisting of multiple freeze-thaw cycles is preferable to a DNA extraction. Only half of each prep will be used in the Ibis WGA reaction to enable subsequent re-amplifications or to allow the unamplified extract to be analyzed. Once we have WGA DNA we will subject this to standard 454-based sequencing to a coverage depth of at least 40X to ensure we obtain >99.99% genomic coverage according to Lander-Waterman statistics.

Principal Investigator

Garth D. Ehrlich, PhD
Professor of Microbiology and Immunology
Executive Director, Center for Genomic Sciences
Allegheny-Singer Research Institute
320 East North Avenue
Pittsburgh PA 15212

Other Participating Researchers

Josh Earl, MS, Jennifer Bennett, MS, Benjamin Janto, MS – employed by Allegheny-Singer Research Institute

Expected Research Outcomes and Benefits

The expected outcomes of the larger research program are the elucidation of the entire biosynthetic pathways that encode bioactive small molecules that are useful in the treatment of: cancer, heart disease, infectious disease, neurodegenerative diseases, and immune dysregulation disorders. Once all of the genes are known that encode the proteins that make up these biosynthetic pathways, we will be able to clone each complete pathway into a genetically tractable organism for the inexpensive microbial production of these compounds. Currently, most natural product compounds that are produced for medical treatments are made using very expensive multi-step synthetic organic syntheses (e.g. the anti-tumor compound, ET473) since the underlying genetic machinery has been unavailable to produce them inexpensively using recombinant DNA technology and microbial biosynthesis. Thus, our work will reduce the cost of life-saving drugs while making them available to more patients. Perhaps even more importantly, once we have the actual biosynthetic machinery, this will simplify the production of

large numbers of analogues through standard genetic substitution and mutation technologies, some of which may have improved efficacy profiles with less toxicity. For example, it is known that related bacterial species often use very similar biochemical pathways for the production of small related families of small molecules. Once we have one complete gene set for a particular general class of compound production, this will greatly facilitate our cloning of related pathways through the use of homologous DNA screens. Thus, we will not only be able to find additional natural products within the same families for testing, but we will be able to make "synthetic" pathways by identifying branch points in the pathways between and among multiple species and substituting genes from one organism into another organism's pathway to produce novel compounds.

Summary of Research Completed

Our goal for this project was to conduct 16S FISH-based FACS purification of unculturable bacteria for whole genome amplification and sequencing. To this end novel unculturable bacterial species of interest were: (a) identified based on prevalence of 16S ribosomal sequences obtained from 454-based deep 16S amplicon sequencing of complex medical and environmental microbiomes; (b) purified using 16S FISH-based FACS; (c) whole genome amplified by multiple displacement amplification; and (d) whole genome sequences using 454-based shotgun sequencing. The following is a detailed report of the research activities accomplished and data produced for this reporting period.

Proof of Concept

Our first objective was to test this experimental pipeline for optimization and proof of concept. We took human adenoid and orthopedic tissue samples that had been shown via IBIS technology and confocal microscopy to contain *Staphylococcus aureus.* Next we isolated the bacteria from the tissue and labeled the bacteria with a *S. aureus* specific 16S rRNA FISH 1probe allowing us to FACS sort the specific bacteria population.

Confirmation of Staphylococcus aureus: DNA was isolated from joint tissue and adenoid tissue samples. These DNAs were run on both the IBIS T-5000 universal biosensor and the 454 Lifesciences pyrosequencer to confirm the presence of *S. aureus* DNA. *S. aureus* DNA was detected in the joint tissues and the adenoid tissues. Confocal FISH microscopy was used to confirm the IBIS and 454 results for both types of tissues (Figure 1). Based on the results, samples were selected for sorting.

Bacterial isolation from the tissue: Single cell bacterial suspensions were prepared by bead beating the tissues with large [~1 mM] diameter beads (these will not lyse the bacteria). These lysates were then sequentiallyy filtered through a 50 micron filter and then a 10 micron filter. Any remaining host cells and debris were pelleted by low speed centrifugation.

Fluorescence in situ hybridization (FISH). FISH was performed on the selected tissues. Briefly, 4% paraformaldehyde (PFA) fixed tissues were sectioned into approximately 0.25-0.5 cm thick sections with a sterile surgical scalpel, tissues were then incubated in an ethanol series of 80% and 100% for 3 minutes each and FISH was performed with both the eubacterial (EU338) and

species-specific *S. aureus* 16s rRNA fluorescent probes (Integrated DNA Technologies Inc, Coralville, Iowa), conjugated with Cy3 or Cy5 dyes. Each section was incubated with probe-specific formamide and salt concentrations and then immersed in washing buffer with the probe specific salt concentration. Samples were rinsed in sterile MQ water and observed with confocal scanning laser microscopy (CLSM).

FACS: Sorting was done at the University of Pittsburgh on the MoFlo™ XDP sorter from Beckman Coulter. Our results (Figure 2) showed successful sorting of both the orthopedic and adenoid tissues. An unstained control was used to set the gates for sorting. Bacterial cells were sorted and collected for downstream applications. These results show that we were able to stain for specific bacteria and sort the bacteria in very small numbers (one sort we collected only 300 cells).

Use of FACS to Obtain Purified Populations of Target High Value, Novel, Unculturable Bacteria

Our second objective was to FISH/SORT true unknown bacteria from the tunicate. The following is a detailed description of how the 16S probe was designed for FISH.

454 Sequencing and assembly of the hologenome: Hologenomic DNA was extracted from frozen *E. turbinata* samples using a DNeasy Tissue kit (Qiagen). DNA was used to prepare: 1) a 16S targeted 454 amplicon library and 2) random shotgun 454 libraries. Sequencing was performed on a Roche/454 Lifesciences FLX Sequencer and later with the 454 Titanium upgrade. Raw reads from the shotgun libraries were assembled using the 454 Newbler assembler. Between two shotgun sequencing runs 839,923 reads were generated with an average size of 332 bp yielding 77,754 total contigs and 15,097 contigs larger than 500bp.

Biosynthetic pathway discovery: Raw and assembled data was initially filtered using relatedness of the translated protein sequences to saframycin and safracin non-ribosomal peptide synthases (NRPSs) characterized in *M. xanthus* (U24657), *S. lavendulae* (DQ838002) and *P. fluorescens* (AY061859) using BLASTx and tBLASTn searches. Once candidate sequences were selected, primers were designed off the ends of each filtered read or contig, and PCR reactions were performed based on the location and orientation of the BLAST hit on the reference sequences. Sequencing of positive reactions were performed to link multiple reads or contigs. This approach yielded six contigs of high interest that contained putative nonribosomalpeptide sythetase (NRPS) domains expected to be present for biosynthesis of ET-743.

Determination of the bacterial populations present in the metagenome: Phylogenetic classification of both the raw reads and of the total assembly was performed using the Metagenomic Rapid Annotations with Subsystems pipeline (MG-RAST). Results from raw reads and assembly were consistent, indicating ~40% of the classified sequences were of eukaryotic origin (mostly identified as *Ciona* [sea squirt/tunicate]). The remaining 60% was dominated by proteobacterial sequence (>90%) and of these two major populations stood out: α-proteobacterial (largely *Rhodobacteraceae*) making up 78-85% of the proteobacterial reads and γ-proteobacterial making up 10-17% of the proteobacterial reads.

Analysis of the 16S amplicon sequencing run was performed by assembling the raw 16S reads

with an identity threshold of 95%. The assembled contigs were then submitted to the Ribosomal Database Project (RDP) for classification. Analysis of our 16S rRNA gene amplicon sequencing run identified 30 variants but only three significant ones (making up >1% of the total reads). The largest population of 16S rRNA gene reads are classified as *Rhodobacteraceae* (~78%), consistent with the classification of total reads by MG-RAST. The second most abundant 16S rRNA gene variant is an unclassified γ-proteobacteria (~19%). A third small population of 16S rRNA gene reads (~1%) is observed. These three variants account for >97% of the total number of 16S rRNA gene sequencing reads therefore it is likely the NRPS module identified in Biosynthetic Pathway Discovery (above) is present one of these three organisms.

Use of FACS to obtain purified populations of target bacteria: FISH probes were designed based on the three assembled 16S sequence variants. These probes were used in the proof of concept process. As shown in Figure 2D, we were able to FISH 60,000 specific bacteria from the tunicate tissue.

Additional Details for Methods:

IBIS technology: The IBIS technique is based on the principle that despite the large diversity among microbial organisms, shared sets of common traits are included in their genomes. The extracted nucleic acid from the original tissue sample is aliquoted into the wells of a microtiter plate which contains broad range primers for PCR. The broad-range primers are designed to amplify a product from a group of organisms from a selected domain of microbial life, for example, all bacteria or specific bacterial divisions. The PCRs produce a mixture of products reflecting the diversity of the original mixture of organisms present in the starting sample. The products from the PCRs are desalted in a 96-well plate format and sequentially electrosprayed into a mass spectrometer. The spectral signals are processed to determine the masses of each of the PCR products present with sufficient accuracy that the base composition of each amplicon can be unambiguously deduced. Using combined base compositions from multiple PCRs, the identities of the pathogens and their relative concentrations in the starting sample can be determined.

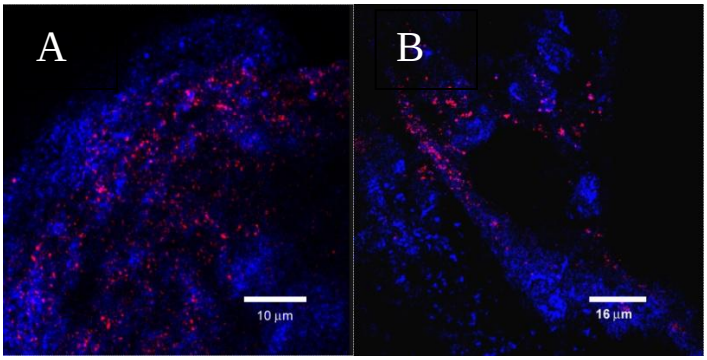

Figure 1: Confocal FISH microscopy confirmed the presence of *Staphylococcus aureus* in both the adenoid and orthopedic tissues (only the adenoid tissue is shown here). (A&B) FISH was performed on adenoid tissue with Eubacterial (EU338) and with species-specific *S. aureus* 16s rRNA fluorescent probes, conjugated with Cy3 or Cy5. CSLM imaging was performed and the images were collected and analyzed using the Leica LCS software. (blue = live/dead stain; pink = *S. aureus* specific probe)

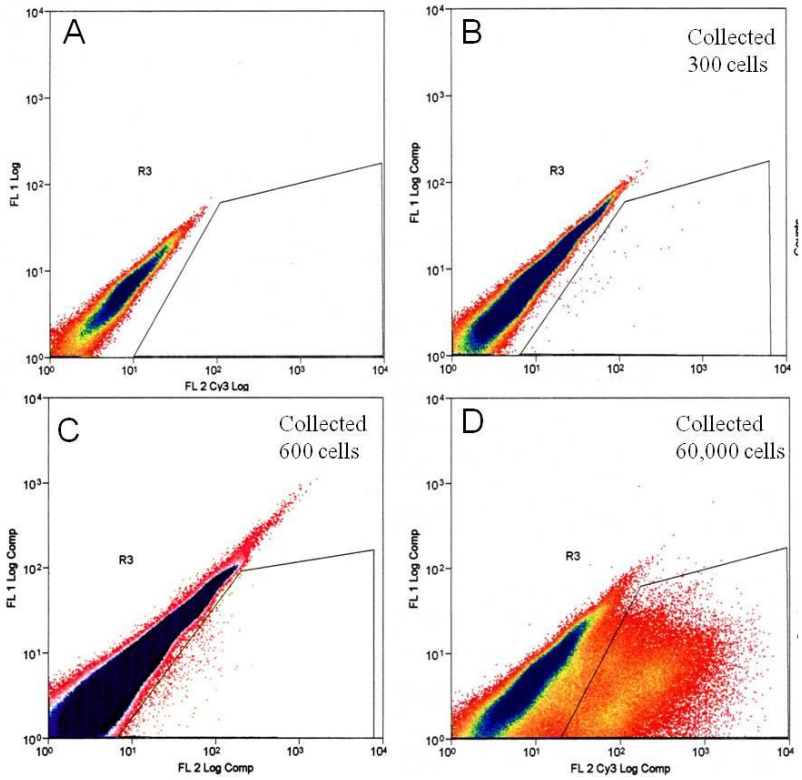

Figure 2: Sorting pure populations of bacteria targeted with species specific 16S FISH probes. (A) Unstained control (used to set the proper gates for the sort) (B) Bacteria isolated from orthopedic tissue and stained with specific 16S probe for *S. aureus* (cy3). 300 cells were collected. (C) Bacteria isolated from adenoid tissue and stained with specific 16S probe for *S. aureus* (cy3). 600 cells were collected. (D) Bacteria isolated from Tunicate samples and stained with specific 16S probe for *Type 1-4* (cy3). 60,000 cells were collected.