MMI GRANTEE NARRATIVE REPORT

Grantee Organization	J. David Gladstone Institutes
Project Lead	Dr. Katherine Pollard
Grant Title	Global mapping of microbial functions: The Environmental Niche Atlas
Grant Number	#3300
Grant Amount	$1,513,352
Due Date of Report	6/1/14
Purpose Statement	In support of the development of predictive models of the global distributions of marine microbes based on functional potential and protein expression patterns in addition to their taxonomy - an Environmental Niche Atlas - to determine the biogeography of marine microbes and their roles in the world's oceans.

MMI Definitions:

1. Outcome = Change occurring in understanding, behavior, institutions, and/or conditions, preferably with an enduring positive impact.

2. Output = Product, service, and/or knowledge resulting from a grant's Activities ("deliverable").

3. Activities = Actions and processes employed to accomplish Outputs and/or Outcomes ("task").

MMI GRANTEE NARRATIVE REPORT

1. GRANT OUTCOMES, OUTPUTS & ACTIVITIES

*Please describe the **cumulative progress towards achieving each Outcome and Output** listed below. Mark the **approximate percent complete** for each Output. In general, report only against Outputs, not Activities; the Activities are listed for your reference and should be used only to guide your responses. For all report answers, if this is not your first report for the grant, please use a colored font to indicate text that is new or revised since the last report. Also, in the appropriate space below, please describe how, **in aggregate**, the progress towards achieving these Outputs culminates in progress towards achieving the Grant's Outcome and a greater understanding of marine microbial ecology.*

OUTCOME 1

Outcome 1, Output 1 (100% completed): Define protein families from sequenced genomes, catalog existing metagenomic and metranscriptomic data, and develop geographic distribution models. *(6/15/13)*

1.1.1 Catalog protein families and compare these with other protein family databases (e.g., Pfam).

We publicly released the Sfams database, as well as open source software for generating and updating Sfams. We published a manuscript describing Sfams and comparing them to other protein family databases. Sfams specifically include sequences and gene families from recently sequenced genomes, and is therefore a particularly large and diverse, albeit not manually annotated, resource. Sfams fills a critical gap between raw sequence data and smaller, more curated databases. Sfams are being used by a number of labs, including the NIH-funded Enzyme Function Initiative, a consortium aimed at identifying and characterizing novel enzymes, and has been downloaded over 1,000 times since the data was published.

To address limitations of using a single gene for taxonomic analysis of microbial communities, and to better leverage information from shotgun metagenomes, we identified phylogenetically informative protein-coding gene families. These phylogenetic and phylogenetic ecology (PhyEco) gene sets alleviate many of the complications introduced by horizontal gene transfer, convergent evolution, or evolution rate variations. To use the PhyEco makers for Operational Taxonomic Unit (OTU) building, we developed TreeOTU, an algorithm that depends on phylogenetic tree structure (rather than nucleotide sequence similarity) to identify taxonomic groups. The method takes into account differences in rates of evolution between taxa and between genes, and it is more robust than non-phylogentic ssu-rRNA OTU classification methods. We made our manuscript, datasets, and software packages available to the public. We are currently developing the application of TreeOTU to metagenomic data sets to enable comparison between OTUs from different PhyEco markers and different environmental samples.

1.1.2 Develop models for protein family geographic distributions with statistical techniques such as niche modeling.

We developed and publicly released niche-modeling methods and software that can be applied to community taxonomic or functional (i.e., protein family) profiles. These tools will continue to be updated. They have been used in multiple publications (see below).

1.1.3 Identify, process, and assess usefulness of current data sets (public and/or available through MMI collaborations) for making niche maps.

We assigned a team member to routinely check data resources for new data sets and to maintain documentation of these datasets on our project wiki. We also actively networked with marine microbiologists through workshops and conferences. These efforts established some new collaboration. But we are still awaiting access to marine shotgun metagenomes at a regional or global scale. We have been using other techniques to analyze the L4 English Channel metagenomes and metatranscriptomes, which have temporal but not spatial variation, and therefore are not ideal for niche modeling. To test, validate, and demonstrate the utility of the new tools developed through iSEEM2, we have also been working with shotgun metagenoms from other environments (e.g., soil, air, human host-associated).

To automate data searching, we are currently implementing a software tool for updated downloads of the short read archive with computer generated queries, as well as notifications and plots to keep track of publicly available marine metagenomic datasets that have been generated. We are also search IMG for assembled metagenomes. We will scan these results regularly for data we can analyze with iSEEM tools.

With completion of this Output, we have established the basic computational approaches and reference gene families needed to model microbial community functional diversity on a global scale. Associated publications, software, and databases are all freely available (see tables for details).

Outcome 1, Output 2 (90% completed): Develop methods to classify metagenomic and metranscriptomic data into protein families, to update protein families with new genome data, and to model the environmental niches of protein families. *(7/1/14)*

1.2.1 Design and validate a novel, fully automated method to classify metagenomic sequence reads into protein families.

We implemented an automated bioinformatics pipeline for read classification, called Shotmap. This tool has been thoroughly optimized (e.g., classification thresholds specific to different read lengths) and validated via computer-simulated metagenomes. We are now applying Shotmap to the L4 metagenomes and metatranscriptomes, as well as metagenomes from other environments. The open source code is publicly available. A manuscript describing Shotmap is in preparation, and we expect to submit it this summer.

Estimation of gene family abundance from classified reads requires adjusting for the number of reads expected to hit each family, which depends on the length of the gene and the average coverage of the genomes in a metagenome. In microbial communities with larger average genome size (AGS), fewer reads will be classified into each gene family. To correct for this effect, we developed a novel software tool, MicrobeCensus, which rapidly and accurately estimates the AGS of a microbial community from shotgun metagenomics data. Applying this tool to real data from the Human Microbiome Project, we confirmed that AGS is a major source of bias when computing the relative abundance of gene families, and that normalization improves the detection of differentially abundant functions across samples. We also showed that AGS correlates with evolutionary and functional properties of the community and is an interesting biomarker/metric for ecological analyses (1.3.1).

1.2.2 Develop methods for updating protein families as new genomes are sequenced.

We implemented a software pipeline for automatically updating Sfams as new genomes come

on line. Our approach is much more computationally efficient than regenerating the Sfams from scratch at each new iteration, because it first compares new sequences to existing families and then only performs de novo clustering on the novel sequences. The code is publicly available.

1.2.3 Design methods to estimate taxon and protein family ranges in environmental niche space, defined by biogeochemical data.

We developed mathematical theory, based on ideas from convex geometry, that ties the shapes of taxon ranges to three well-known attributes of ecological diversity: the taxon-area relationship, the endemics-area relationship, and beta-diversity. Our main result is that these three diversity metrics are determined almost entirely by the area and perimeter of geographic ranges. We empirically validated this mathematical result using IUCN data on the ranges of all known amphibians, birds, and mammals.

The theory also applies to microbial taxa and gene families. However, it is unlikely that we could empirically validate these findings for marine microbial communities, because this would require an exhaustive census for all taxa or genes. Nonetheless, we might be able to turn the equations around and make inferences about the shapes of microbial ranges from measurements of diversity. We are currently exploring this possibility. We are also using these results to derive a total diversity estimator (see below), which only requires a deep census at a few sampling locations – this estimator will be directly applicable to microbial metagenomics.

Niche modeling provides an alternative method to estimate ranges of individual microbes and microbial gene families as a function of biogeochemical data, climate, or other environmental variables. While this approach does not provide a discrete range boundary, it is potentially more informative and definitely more applicable to microbes. The output is a map of the probability of seeing a given taxon or gene family at any location. For groups of taxa or genes (e.g., pathways), niche modeling can also be used to map various summary measurements, such as alpha- and beta-diversity metrics. For both types of maps, the underlying model can be dissected to infer which environmental variables are driving the geographic patterns.

1.2.4 Evaluate what models (e.g., from Activities 1.1.2 and 1.2.3) are appropriate for existing data. Determine if sufficient metagenomic and metranscriptomic data exists for accurate modeling.

Direct estimation of microbial ranges from metagenomic data is probably too inaccurate to be applied with currently available data. In many ecosystems, however, environmental niche modeling can be applied to accurately predict the distributions of taxonomic and functional groups. We have demonstrated this modeling technique in two publications, and we are now broadly applying it to data from marine and other environments.

As we approach completion of this aim, we have produced all the tools needed to produce accurate functional niche models and maps from shotgun metagenomes and metatranscriptomes.

Outcome 1, Output 3 (20% completed): Annotate protein families based on phylogenies and functions to generate niche maps and to characterize global microbial diversity. *(7/1/15)*

1.3.1 Develop metrics to quantify protein distributions and diversity from metagenomic and metranscriptomic data.

We are developing metrics based on (i) the spatial distribution of genetic variation within protein families ("phylogeography of function") and (ii) comparing taxonomic diversity to functional diversity within and across microbial communities ("comparative diversity"). Both projects use PhyloSift to analyze the raw data and have similar underlying frameworks. We have almost completed the scripts that automate the steps shared by the two projects. These pipelines are a novel way to utilize metagenomic data; we aim to understand better the diversity of microbial communities and their evolutionary history by adding phylogenetic and functional context to their spatial distributions.

The theory developed in 1.2.3 has allowed us to develop estimators of regional or global diversity (e.g., the total number of OTUs occurring across an ocean). The estimators are based on knowing the areas of a sample of species ranges (or OTU or gene ranges), and the total (or close to total) diversity at a few sampling locations. To apply these estimators, the species ranges can be inferred using niche modeling, so our method allows for incorporating remotely-sensed environmental data to estimate regional or global diversity. Through the incorporation of these data, it may be possible to arrive at improved estimates of diversity.

Finally, we are currently developing a method to identify significantly high- and low-variance protein families across samples. This method corrects for the mean-variance relationship, as well as the impact of phylogenetic diversity on protein family variability. Protein families found to have
high variance across sampling locations may be involved in adaptation to specific environmental niches, while low-variance families may be involved in more fundamental or essential processes. An initial application of this variability metric to human gut metegenomes showed that it pinpoints ecologically important protein families that reflect differences in host diet and microbial community composition, such as type III secretion systems, carnitine metabolism, and lipid A metabolism. We are eager to apply this metric to marine metagenomes across environmental and geographic gradients to identify the ecologically responsive protein families.

1.3.2 Design bioinformatics tools for mapping known protein functions onto classified metagenomic and metranscriptomic sequence reads.

We expanded our Shotmap analyses to include a broad collection of databases ranging from diverse but poorly annotated (Sfams) to smaller but highly annotated (MetaCyc). We are comparing the conclusions reached with different databases, and with metagenomes versus metatranscriptomes. To computationally predict the functions of unannotated Sfams (or other gene families), we designed a method based on protein sequence similarity and the Extreme Set algorithm, which detects communities in a network graph. Essentially, functions are transferred from highly similar gene families as long as the group of co-annotated gene families is highly connected and not overlapping with other groups of gene families. Results were compared to predictions generated by Markov Clustering algorithm (MCL) and tested based on GS2 (GO-based Similarity of gene sets) metrics. We are now extending this approach to utilize information beyond sequence similarity. This includes using structural role discovery to identify gene families with similar positions in a network, such as being equally well connected or equally peripheral.

1.3.3 Implement algorithms for drawing environmental and geographic niche maps.

We have implemented methods for drawing taxonomic and functional niche maps in geographic space. We are exploring visualizations of diversity along environmental axes.

1.3.4 Given sufficient metagenomic and metranscriptomic data, apply tools developed to characterize protein families and predict niches to publicly available marine microbial data. If sufficient data do not exist, apply tools for taxon niche modeling only.

We demonstrated the utility of microbial niche modeling through an application to global marine microbial taxonomic data (16S) and North American soil microbial taxonomic and functional data (shotgun metagenomes). Both studies were published.

1.3.5 Given sufficient metagenomic and metranscriptomic data, compare functional and taxonomic diversity maps to elucidate patterns of selection and convergence; identify diversity hot spots; and propose testable hypotheses about the roles of microbial functions in marine ecosystems. If sufficient data do not exist, determine and document the sampling needed in order to perform niche modeling.

In an application to soil metagenomes, we found a correlation between Verrucomicrobia diversity and gene families for recalcitrant carbon degradation in relatively nutrient limited environments. In another soil study, we used the tools from this project to predict the widespread changes in microbial distributions in response to climate change in Tibet and North America. If we acquire data for making marine niche maps, we plan to look for hotspots of diversity, as well as functions at risk of extinction.

Completion of this aim, which depends on data availability in the next year, will demonstrate the utility of the tools developed by iSEEM and will uncover novel ecological and evolutionary trends in marine microbial community structure and function.

Please describe cumulative progress towards achieving Outcome 1, including mentioning the Outcome 1 milestones you expect to accomplish in the coming year. Lastly, how is progress towards achieving this Outcome culminating in a greater understanding of marine microbial ecology?

Outcome: Predictive models of the global distributions of marine microbes at both the taxonomic level (e.g., genera and operational taxonomic units) and the potential functional level (e.g., protein families and their functions) are developed. *(7/1/15)*

We have already successfully produced, validated, and publicly shared all the bioinformatics tools proposed in the iSEEM2 grant. We have also applied these tools to taxonomic niche modeling of the world's oceans. Our results showed the marine microbial diversity peaks in high latitudes in the winter, with strong temporal dynamics. This pattern was previously unappreciated due to a lack of spatially and temporally explicit models. It is also very different from diversity maps for macroorganisms, highlighting the unique ecological and evolutionary forces that shape marine microbial communities. We additionally applied functional niche modeling to soil metagenomes and compared taxonomic and functional diversity maps, which provided new insight into the role of Verrucomicrobia in nutrient-poor soils. Finally, we showed that soil microbial distributions follow climate from 30+ years ago, rather than contemporary climate, which allowed us to predict how distributions will shift in the future in response to climate change that has already happened up to the present time. We are eager to conduct similar studies for marine communities if appropriate data becomes available in the next year.

MMI GRANTEE NARRATIVE REPORT

2. PUBLICATIONS & PATENTS

*Please list all publications and patents that **have resulted from this project since grant inception**, including papers that are **in press** and **submitted**. For papers published previously as **in press** or **submitted** and that have been published since, please replace with final citation. **Please send PDFs** of all published papers (final versions) at the time you submit this report. Please add additional rows at the bottom of the table if you need more space.*

1. Nayfach S and Pollard KS. Average genome size estimation enables accurate quantification of gene family abundance and sheds light on the functional ecology of the human microbiome (submitted to Genome Biology).
2. Sharpton TJ (2014). An Introduction to the Analysis of Shotgun Metagenomic Data. Frontiers in Plant Science 5:209. doi: 10.3389/fpls.2014.00209.
3. Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, Knight R, Gilbert JA, McCulley RL (2013). Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States, Science 342: 621-624.
4. Riesenfeld SJ, Pollard KS (2013). *Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis*, BMC Genomics, 14: 419.
5. Wu D, Jospin G, Eisen JA (2013). Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups, PLOS ONE 8(10): e77033.
6. Ladau, J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, Koeppel AF, Green JL, Pollard KS (2013). *Global marine bacterial diversity peaks at high latitudes in winter.* ISME 7(9): 1669-1677.
7. Wu D, Doroud L, Eisen JA (2013) TreeOTU: Operational Taxonomic Unit Classification Based on Phylogenetic Trees (http://arxiv.org/abs/1308.6333).
8. Kembel SW, Wu M, Eisen JA, Green JL (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Computational Biology 8(10): e1002743. PMID: 23133348.

9. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D, Tringe SG, Singer SW, Eisen JA, Banfield JF (2013). *Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment.* Nature Communications 4:2120.

10. Rinke C, ... Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459): 431-7.

11. Shih PM, Wu D, ... Eisen JA, Woyke T, Gugger M, Kerfeld CA (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. PNAS 110(3): 1053-8.

12. Sharpton TJ, Jospin G, Wu D, Langille MG, Pollard KS, Eisen JA (2012). *Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource*, BMC Bioinformatics, 13(1): 264.

13. Eisen JA. Phylogenetic and phylogenomic approaches to analysis of microbial communities. In "The Social Biology of Microbial Communities – A Report from the National Academy of Sciences Forum on Microbial Threats."

14. Jiang X, Langille MGI, Neches RY, Elliot MA, Levin SA, Eisen JA, Weitz JS, Dushoff J. 2012. Functional biogeography of ocean microbes: dimension reduction of metagenomic data identifies biological patterns across scales. PLoS ONE 7(9): e43866.

3. PRESENTATIONS

Please list key presentation titles and abstract citations (if appropriate) ***that have resulted from this project since grant inception. Please indicate if you were a plenary or invited speaker or if it is a student presentation****. Please add additional rows at the bottom of the table if you need more space.*

1. May 2014: Thomas Sharpton -- Metagenomic Investigations of the Human Microbiome. OSU Department of Nutrition Seminar. Corvallis, OR. Invited.
2. April 2014: Patrick Bradley -- Assessing the stability of protein family abundance in the gut microbiome. Exploiting and understanding chemical biotransformations in the human microbiome, Keystone Symposium, Big Sky, MT. Invited.
3. April 2014: Josh Ladau -- Mapping historic, current, and future soil biodiversity. iDIV Workshop: A framework to improve our understanding of the distribution of global soil biodiversity: establishing the first quantitative synthesis. Leipzig, Germany. Invited.
4. April 2014: Thomas Sharpton -- From Noise to Signal: Mining Human Exome Sequences Reveals the Hunter-Gatherer Oral Microbiome. Center for Genome Research and Biocomputing Spring Symposium. Corvallis, OR. Invited.
5. March 2014: Katie Pollard -- Statistics and bioinformatics challenges in shotgun metagenomics. Genome Sciences seminar, U Washington, Seattle, WA. Invited.
6. January 2014: Katie Pollard -- Quantifying the taxonomic and functional diversity of metagenomes. BCATS conference, Stanford University, Palo Alto, CA. Invited keynote.
7. January 2014: Katie Pollard -- Promises and pitfalls of studying microbial communities with shotgun metagenomics data. Bioinformatics & Systems Biology Colloquium, UC San Diego, San Diego, CA. Invited.
8. December 2013: Thomas Sharpton -- Metagenomic Investigations of Microbiomes. University of Oregon META Center Seminar. Eugene, OR. Invited.
9. September 2013: Katie Pollard -- Bioinformatics and statistical challenges in metagenome analysis. Next-generation sequencing seminar, UC Berkeley, Berkeley, CA. Invited.
10. August 2013: Josh Ladau -- Universal scaling of beta-diversity across taxa and terrestrial and marine ecosystems. Ecological Society of America Annual Meeting. Minneapolis, MN.

11. May 2013: Jonathan Eisen - Phylogeny-Driven Approaches to Genomics and Metagenomics. Fresno State. Invited.
12. April 2013. Jonathan Eisen – The Need for a Phylogeny Driven Genomic Encyclopedia of Eukaryotes. Society for Molecular Biology & Evolution Eukaryotic 'Omics Meeting. Davis, CA. Invited.
13. April 2013. Katie Pollard – Quantifying the taxonomic and functional diversity of metagenomes from shotgun sequencing data. Society for Molecular Biology & Evolution Eukaryotic 'Omics Meeting. Davis, CA. Invited.
14. January 2013. Katie Pollard – Decoding our genomes: what can our DNA tell us about our close relatives and who is traveling with us? Castro Valley Educational Foundation. Castro Valley, CA. Invited public lecture.
15. November 2012. Katie Pollard – The ome inside us. Systems Biology Symposium, Johns Hopkins Medical School, Baltimore, MD. Invited keynote.
16. October 2012. Jonathan Eisen - Sequencing and Microbes, the Now Generation. At UC Davis Food for Health Symposium. Slideshow on Youtube. Invited.
17. October 2012. Tom Sharpton – Seminar at the Center for Genome Research and Biocomputing, Oregon State University. Invited.
18. September 2012. Jonathan Eisen. Phylogenomic approaches to the study of microbial diversity. At Bay Area Illumina Users Meeting. Slideshare. Slideshow on Youtube. Invited.
19. September 2012. Jonathan Eisen. Phylogenomic approaches to the study of microbial diversity. Lake Arrowhead Microbial Genomes Meeting. Invited keynote.
20. September 2012. Katie Pollard – Seminar at the VanBUGS cross-institutional bioinformatics group, Vancouver, BC. Invited.
21. July 2012. Jonathan Eisen. Phylogenomic approaches to functional prediction. At AFP "Automated functional prediction" submeeting during ISMB 2012. Slide show on Youtube.
22. July 2012. Katie Pollard – Quantifying the taxonomic and functional diversity of microbial commuities with nextgen sequencing data. Joint Statistical Meetings, San Diego, CA. Invited.

MMI GRANTEE NARRATIVE REPORT

4. ADDITIONAL INFORMATION

A. Exclusive of grant requirements, please describe any new research collaborations and any updates on previously reported collaborations that resulted from this research. Please indicate if this was with other MMI-funded grant project leads, with other marine microbial ecologists, or with researchers from outside the field of marine microbial ecology. Please list any proposals that have been submitted as a result of this effort since inception of the grant.

We established collaborations with Jack Gilbert, Haiyan Chu, Noah Fierer, Peter Turnbaugh, and Rob Knight, all of which are excited about applying niche modeling to metagenomic data. These collaborations have improved our access to soil shotgun metagenomes. We have also been working with them to model microbial distributions using 16S data from marine, air, household, and gut environments. We submitted a soil microbial diversity proposal with Jack Gilbert.

We also started collaborating with the Enzyme Function Initiative (EFI) in order to further characterize Sfams of unknown function. In particular, Sfams that are observed to vary significantly across sampling locations and Sfams that are associated with environmental variables will be prioritized for detailed computational and experimental follow-up. The EFI has a track record of success in discovering new enzyme functions through a combination of crystallography, ligand screening, and in silico docking, among other approaches.

The Eisen lab has begun multiple projects on biogeography of microbial communities that have been shaped by work on this project. This includes the recently funded “Seagrass Microbiome project” (a collaboration with Jessica Green and Jay Stachowicz) and projects on biogeography of microbial communities associated with corn and rice.

B. Exclusive of grant requirements, have you provided data, samples, cultures, MMI-funded tools, technology, methods or equipment, etc. to other researchers since inception of the grant? If so, please describe.

Databases and data sets made available broadly:
Sfams: http://edhar.genomecenter.ucdavis.edu/sifting_families/
Vfams: http://derisilab.ucsf.edu/software/vFam/
PhyEco Markers:
http://figshare.com/articles/Systematically_identify_phylogenetic_markers_at_different_taxonomic_levels_for_bacteria_and_archaea/722713

C. Since inception of the grant, have you developed, built or enhanced any instruments, devices, tools, methods or software? Have the codes, designs, protocols, etc. been made available to the public? Please include any relevant information you may have at hand pertinent to the impact of the development (e.g. usage of protocols or code by others, requests for collaboration, etc.). Please include information here even if you listed above the related publication that describes this activity.

Software made available publicly
Shotmap: https://github.com/sharpton/shotmap
MRC: https://github.com/sharpton/MRC
SFam Updater: https://github.com/gjospin/Sfam_updater
Simulations: https://github.com/sriesenfeld/MetaPASSAGE
Niche Modeling: https://github.com/jladau/SpeciesDistributionModeling/
MicrobeCensus: https://github.com/snayfach/MicrobeCensus
TreeOTU: http://figshare.com/articles/TreeOTU/783077

D. Please list any challenges in accomplishing your outcomes for the past year and discuss your plans for overcoming these challenges in the upcoming year.

Our primary challenge is data availability, specifically shotgun metagenomics (or metatranscriptomics) data from marine samples collected on a regional or global scale.

We also devote considerable effort to expanding and maintaining the computer resources needed to support this project, because datasets have significantly grown in size since the start of the grant period. This challenge is largely being addressed through new computer purchases in the Pollard and Sharpton labs.

E. Please describe the most exciting results arising from this award since your last report (if any). If appropriate, please send an attachment (one figure, image, or photo) that conveys the exciting results. MMI will keep this confidential and request your permission for usage.

One of our most exciting findings is that soil microbial community structure and diversity is better predicted from historic than current climate. This observation allows us to predict future distributions, because they should be highly correlated with current climate. Applying niche modeling in this way, we predicted widespread changes in microbial distributions and diversity across most of North America and Tibet (see Figure 1). Our goal is to extend this approach to marine data, if we can identify similar climate lags there.

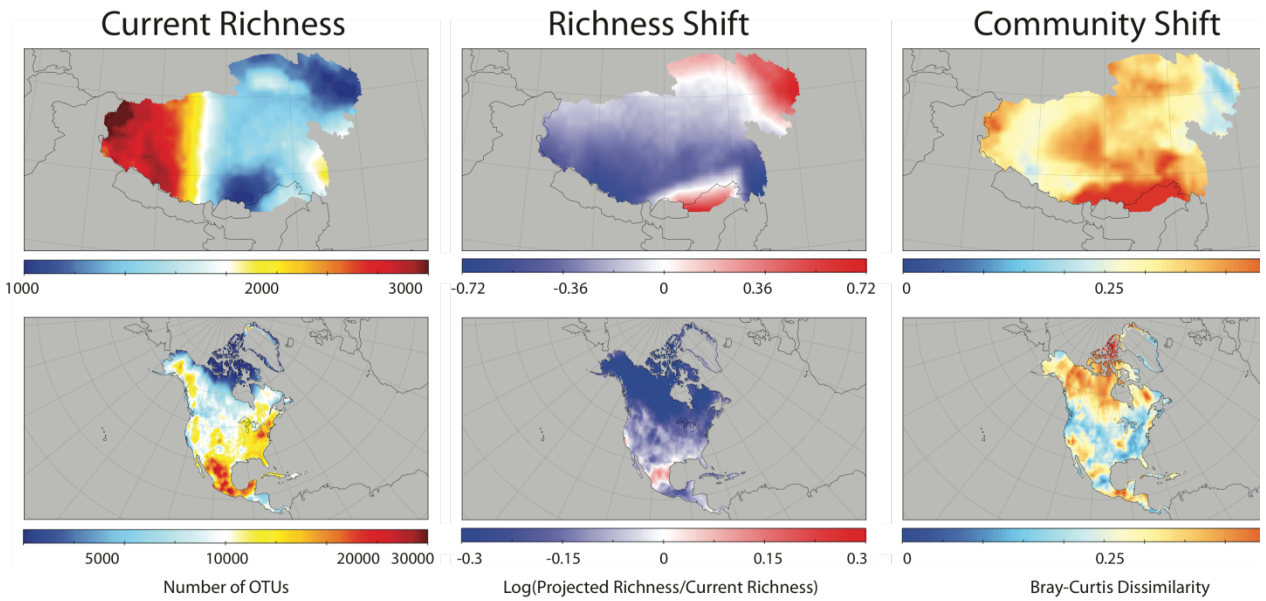

Figure 1. Shifts in bacterial diversity that are committed with current climate change; that is, the shifts that would occur if climate change stopped entirely today. The first column shows current patterns of OTU richness. Declines in richness are expected widely across 78% of the Tibet and 96% of North America (second column). These declines in richness will be accompanied by substantial changes in the composition of communities, as measured by the Bray-Curtis dissimilarity between current and projected future communities (third column). The geographic projections were made using niche modeling methodology.

F. Any lessons learned regarding your research and/or collaborations from a scientific, management, or other perspective? While we welcome reporting on lessons learned across the grant term, please emphasize what you have learned in the past year.

We have a weekly conference call with the whole iSEEM research team. This regular interaction has greatly facilitated our collaboration and improved the quality of every one's work. Our project wiki is also a helpful tool for communication and project management.

For communicating our findings to the broader community, we have started to use a project website, which all team members can easily update: http://iseem2.wordpress.com. We also track the citations and discussions and uses of our outputs using the Impactstory tool: https://impactstory.org/iSEEM2.

G. Please describe special recognition you or your co-PIs have received specifically as a consequence of this project. If a paper published was linked to a press release or covered by the media (on-line, TV, etc.), please provide the URL and other details as available.

Several of our publications have been "highly discussed" online, as documented on Impactstory. For example, see: https://impactstory.org/iSEEM2/product/36chywycsxpzg7dmnw195u06.

Our figure of predicted microbial diversity in the American tallgrass prairie before agriculture was selected by Wired Magazine as one of the best scientific visualizations of 2013. The figure was drawn using the niche modeling methods developed through this grant.

Katie Pollard was confirmed as a Fellow of the California Academy of Sciences.

Postdoc Josh Ladau received a poster award at the Gordon and Betty Moore Foundation's Marine Microbiology Initiative RA and Postdoctoral Scholar Summit.

Sarah Hird has been selected as a "University of California President's Post Doctoral Scholar".

H. Please report any student theses supported by this project, including degree earned.

None.

I. Please list the web addresses for your lab and for any databases, resources, etc. related to this grant.

http://iseem2.wordpress.com

http://openwetware.org/wiki/ISEEM

https://impactstory.org/iSEEM2

http://docpollard.org

http://phylogenomics.wordpress.com

http://phylogenomics.blogspot.com

http://lab.sharpton.org

J. MMI has supported a number of community resources (e.g. CAMERA, MEGAMER, VIROME, and the bacterial, viral, and microeukaryote sequencing projects). Have you used any of these resources? If so, please describe how you have used the resources, emphasizing use over the past year.

We used CAMERA as one place to search for publicly available marine metagenomics data. Some GBMF funded genome sequences were included in the build of Sfams.

K. Please provide a brief narrative description of expenditures to date and planned upcoming expenses. Please include an explanation of any budget variances and surpluses.

Our primary expenses are directly related to project personnel (salaries, benefits, supplies, travel). Capital expenditures refer to funds for a compute server. Other expenses refer to funds for use of the bioinformatics core (personnel, computer time, lab space) at Gladstone Institutes.

Cumulative spending since the start of the grant is below budget in all categories. This is largely due to initial (2012-2013) delays in recruiting personnel and setting up the UC Davis subcontract. In the past 12 months, our spending has more closely matched the budget. We are still under-spending by a small amount, largely due to some personnel changes, including Tom Sharpton transitioning to a faculty position (impacting salaries and wages, employee benefits, travel) and loss of a bioinformatics engineer at Gladstone that resulted in reduced usage of the Bioinformatics Core (impacting other expenses). The engineer has been replaced, and the Pollard lab is currently recruiting a new post doc or graduate student. Capital expenditures are a little below budget, because computer server costs came down since the budget was written. We continue to purchase computer hardware; smaller purchases are budgeted as supplies rather than capital expenditures. Finally, several months of subcontract expenditures have been committed but not yet billed.

L. Please use this space to respond to any additional questions MMI posed when sending you this form. You may also use this space to provide any additional feedback to MMI. How can the Initiative do its job better? How can MMI better facilitate your research?

Nothing to report.

MMI GRANTEE NARRATIVE REPORT

6. PERSONNEL

*Please list the personnel **supported financially since the inception of this grant**. Personnel categories include **lab managers**, **technicians**, **post-doctoral researchers**, **graduate students**, **undergraduate students**, **high school students**, and **others** (please define).*

Year: 6/1/13-5/31/14

Name	Personnel Category	Approximate Duration of Support (Months)
Katherine Pollard	Principal Investigator	2.7
Jonathan Eisen	Principal Investigator	1.8
Thomas Sharpton	Principal Investigator (previously Post doc)	2.5
Alexander Williams	Software Engineer	0.6
Stephen Nayfach-Battilana	Graduate Student	4.3
Ladan Doroud	Graduate student	4
Dongying Wu	Staff scientist	4
Guillaume Jospin	Bioinformatics engineer	6
Sarah Hird	Post doc	6
Josh Ladau	Post doc	9
Stacia Wyman	Bioinformatics engineer	1.5

Year: 6/1/12-5/31/13

Name	Personnel Category	Approximate Duration of Support (Months)
Katherine Pollard	Principal Investigator	2.4
Jonathan Eisen	Principal Investigator	1.8
Thomas Sharpton	Post doc	6

Mariel Finucane	Biostatistician	2.3
Ladan Doroud	Graduate student	4
Dongygng Wu	Staff scientist	4
Guillaume Jospin	Bioinformatics engineer	6
Josh Ladau	Post doc	9

7. OTHER FUNDING

Please list information describing any other funding you have applied for (since grant inception) that was stimulated by your award from MMI—noting whether pending, successful or unsuccessful.

Funding Agency	Proposal Title	Total Request or Amount Awarded	Amount (to be) Awarded to You	Start Date	End Date
NIH	Longitudinal and functional dynamics of autoimmune gut microbiomes	$532,709	$532,709 (pending)	7/1/14	6/30/14
NIH	Collaborative Center for an Enzyme Function Initiate	$5,000,000	$130,247 (subcontract) (successful)	5/1/14	4/30/15
NSF	Collaborative Research: Creating a continent-scale terrestrial microbial bioclimatic envelope model for the US	$618,412	$618,412 (unsuccessful)	NA	NA
NSF	The Elaphomyces microbiome – a model for the ecological evolution of fungal-bacterial interactions	$2,000,000	$500,000	TBD	TBD
GBMF	Data Driven Discovery Investigator: A field guide to the microbes (Eisen)	$1,000,000	$1,000,000 (pending)	TBD	TBD
GBMF	Data Driven Discovery Investigator: Analyzing inter-protein relationships to infer molecular, evolutionary, and ecological functions	$1,000,000	$1,000,000 (unsuccessful)	NA	NA