Multidisciplinary research and education with open tools: Metagenomic analysis of 16S rRNA using Arduino, Android, Mothur and XSEDE

Kristin Muterspaw
Earlham College
Richmond, IN 47374
kmmuter11@earlham.edu

Tara Urner
Earlham College
Richmond, IN 47374
tmurner12@earlham.edu

Ruth Lewis
Earlham College
Richmond, IN 47374
rylewis22@gmail.com

Ivan Babic
Earlham College
Richmond, IN 47374
ibabic09@earlham.edu

Deeksha Srinath
Earlham College
Richmond, IN 47374
dsrina13@earlham.edu

Charles Peck
Earlham College
Richmond, IN 47374
charliep@cs.earlham.edu

David CerdaGranados
National Autonomous
University of Nicaragua
Leon, Nicaragua
david.cerda@ct.unanleon.edu.ni

Peter Lemiszki
Pellissippi State College
Pellissippi, TN
pjlemiszki@pstcc.edu

María Sánchez-Miranda
National Autonomous
University of Nicaragua
Leon, Nicaragua
marimirsa@yahoo.es

Mercedes Mayorga-Méndez
National Autonomous
University of Nicaragua
Leon, Nicaragua
mrcds_mayorga27@yahoo.es

Olafur Petursson
Skalanes Research Station
Seydisfjordur, Iceland
skalanes@skalanes.com

Ben Smith
US Peace Corps
Nicaragua
bents2012@gmail.com

ABSTRACT

Modern scientific research is often multidisciplinary, involving scientists from two or more backgrounds. A multidisciplinary approach is frequently necessary to advance our knowledge in a diverse range of fields, from genomics to climate change. Many of the projects undertaken in these areas involve a combination of field, lab, and computational analysis components. Our research initiatives demonstrate how the principles of active learning – performing tasks while engaging in analysis, synthesizing and evaluating the tasks being performed [10] – can be applied to undergraduate science education using 16S rRNA metagenomics as the basis. Beginning with development of the scientific questions, students work through the entire process of designing, testing and implementing physical and digital sampling protocols, hardware and software platforms for collecting geographically coded (geocoded) environmental metadata, and lab protocols; they work in the field taking samples and in the lab preparing them; they perform the computational analysis of the sequencer output and synthesis of metadata and metagenomic data; and finally they disseminate the results. The students come primarily from backgrounds in computer science, biology, geology, and physics. This broad range makes it possible to select teams that cover many of the traditionally underrepresented groups in science. Working together for a year or more, the students learn the science, vocabularies, skill sets, etc. of all the disciplines, as well as how their own discipline, in conjunction with others, contributes to addressing large, complex questions.

XSEDE '15 July 26-30, 2015, St. Louis, MO, USA

ACM ISBN 978-1-4503-3720-5/15/07.

DOI: http://dx.doi.org/10.1145/2792745.2792767

CCS Concepts

•**Applied computing** → **Education;** *Bioinformatics;*

Keywords

Multidisciplinary science, education, computer science, biology, geology, archaeology, ecology, bioinformatics, metagenomics, GIS, mobile app, active learning, International collaboration

1. INTRODUCTION

Our research group is comprised of undergraduate students, recent graduates, and a faculty member. The students come from a variety of academic backgrounds, and our projects are developed at the intersection of our fields of study. We describe our work through two lenses, the research initiatives that organized our work, and the underlying educational approaches that shape how the work was done. The science is further broken down into a discussion of how the biology (field, lab, computational analy-

sis), computer science (software/hardware development, sequencer output analysis are visualization) and used in our work.

Research Initiatives

The underlying science for this project comes primarily from biology and computer science, with influences from ecology, geology, and archaeology. The central component is biological, the metagenomic analysis of DNA coding for 16S ribosomal RNA (rRNA) that is extracted from microbes in soil in both Iceland and Nicaragua, and from microbes on and within the leaves of coffee plants in Nicaragua. The 16S rRNA is a component of the 30S small sub-unit of prokaryotic ribosomes. The DNA which codes for the 30S sub-unit is referred to as the 16S rRNA gene coding and is used to reconstruct phylogenies, both because it is highly conserved in prokaryotes and the slow rate of mutation in this region of the gene. In our research, we used 16S rRNA gene sequencing to identify and compare the microbes present in our samples [16].

In the field, we collected geocoded biological samples and environmental parameters - e.g. pH, fertility, moisture, temperature, and humidity - using Arduino® and Yoctopuce sensor platforms and Android™ application software. Our Android application, Seshat (named after the Egyptian goddess of measurement), connects through USB or BlueTooth to the sensor platforms and collects readings at a user-specified interval. Over the past three years, we designed and built Seshat, along with all the Arduino based sensors platforms. We performed all field and lab work and developed new protocols for extracting DNA from sand and leaves. The results from the Illumina MiSeq sequencer were analyzed using Mothur (an open-source software package) [14] running on local, XSEDE [15], and Blue Waters [1] computational resources. The results of both studies are being used by local scientists and students as part of their teaching and research. An associated Blue Waters Petascale Internship project is using the sequencer output as the basis of a profiling, parallelization, and tuning review of Mothur, with the goal of increasing the scalability for large metagenomic studies and porting it to HPC platforms which do not support the *fork()* parallelism model, e.g. Blue Waters.

Our approach to using soil microbe metagenomics to address diverse research questions is interesting in-part because we focused on maximizing the amount of geocoded environmental metadata which was collected with the biological samples, and providing a structure for incorporating that metadata in the analysis and visualizations. Making it easy to generate geocoded visualizations which tie biological, ecological and environmental data and metadata to the (relatively) exact spot being studied provided new opportunities for students and faculty. Having location be a first-class notion provided broad context, an important component of linking results from different disciplines into a larger, more comprehensive picture.

Thus far, our work has involved projects in two different settings: a research station in Iceland and coffee farms in North-central Nicaragua. We used many of the same sensor platforms and sampling protocols in both locales but our research questions were different in each case.

Iceland was our first complete field, in-country lab extractions, DNA transportation, DNA preparation and sequencing, and computational analysis test. Our research focused on a small research station, Skálanes, located on the eastern fjord of Seydisfjordur [2], see Figure 1. There are two study areas at Skálanes featured in this paper: a bird sanctuary that encompasses heath, wetlands, and the coast of the fjord, and the archaeological site of an early European settlement (circa 1000CE). We collected soil samples and metadata at each of the sites at Skálanes with the goal of testing all of our field and lab protocols, hardware, and software.

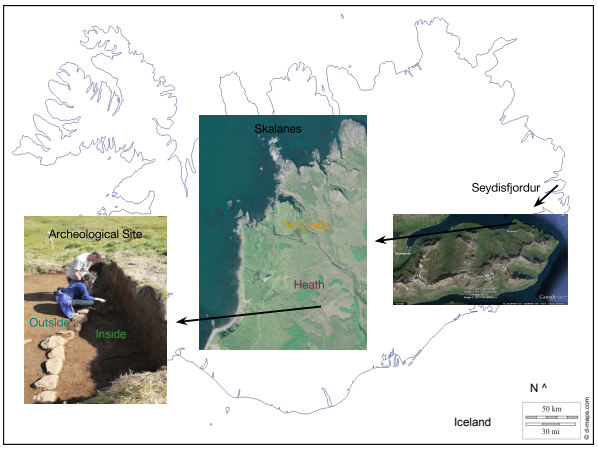

Figure 1: The Eastern peninsula of Iceland where Skálanes is located.

In Nicaragua, we took soil and leaf samples from approximately five randomly chosen spots at each of the six farms in the cloud forest in North-central Nicaragua, see Figure 2. At each of the sampling spots, one infected and one uninfected coffee plant, and the surrounding soil were sampled and metadata was collected. The coffee rust fungus, *Hemileia vastatrix*, is spreading throughout the coffee growing regions of the world, including the cloud forest of North-central Nicaragua, where it is referred to as *La Roya* [4]. Working with local scientists we designed our study with the goal of analyzing how microbes living in the soil and leaves of coffee plants, as well as environmental conditions around the plants, might affect the growth and spread of the coffee rust.

Figure 2: North-central Nicaragua where the six coffee farms are located.

Education

The ability for people with different intellectual backgrounds to work together - in this case, geologists, biologists, physicists, and computer scientists - is essential in the world of multidisciplinary scientific research, which itself is becoming more common. Our experience has been that starting this mixing process as part of undergraduate student/faculty

research, and courses, appears to be an effective way to mitigate some of the compartmentalization which tends to occur in academia. Giving students the opportunity to explore a problem from beginning to end over multiple semesters, doing work which included designing/building/testing hardware and software, developing field and lab protocols, collecting the samples and processing them, and managing the research up through computational analysis and dissemination, helps foster collaborative work between disciplines. It also gives the students a sense of investment and ownership in the work, a key principle of active learning.

2. RESEARCH INITIATIVES

The main components of this research are collecting soil and leaf samples, the associated geographical and environmental metadata, processing the samples in the lab to prepare them for sequencing, analyzing the sequencer output, and visualizing the results. Throughout the workflow, from sample collection to distance matrix visualizations, metadata is collected, organized, and curated.

Motivation

In Iceland, we sampled at three different sites at the Skálanes research station: a wetlands, a heath, and an archaeological site. Skálanes is a hummocky, vegetated, boggy landscape sitting on an uplifted shoreline terrace. Nearby cliffs expose bedrock consisting of discrete basalt flows with intervening paleosol horizons. Basalt rubble (talus deposits) most likely underlies the site upon which the recent soil has formed. The soil at the archaeological site has been disturbed due to human settlement as indicated by the absence of tephra layers and the preservation of distinct charcoal deposits that record the location of fire pits within the walls of the house. The soil is in part a silty loam, but the specific characteristics of the soil at the site have not been investigated. Existing soil maps indicate that the undisturbed residual soils in the area are a variety of andosols [9]. Andosols form in volcanic rich material and are characterized by special kinds of clay minerals (e.g., allophane), the ability to store large quantities of water, and the tendency to bind organic matter [8]. In Iceland, we worked towards generating a geocoded profile of microbial species and associated environmental parameters, such as pH and moisture in the soil at different sites within Skálanes.

These projects are part of on-going work by the archaeologists to interpret and extend their dig site, and the ecologists as they work to understand the effects of lupin on the avian habitats in the surrounding heath and wetlands.

In Nicaragua, we applied what we had learned about the effectiveness of our methods in Iceland to a particular problem. Coffee is the most important entity in the agricultural sector and one of the main export products of Nicaragua, generating 338 million USD per year [3]. Three-hundred thousand direct and indirect jobs are generated by the coffee industry, which represents the 53% of the total of jobs in the agricultural sector and 14% of the labor positions in the country. 93% of the coffee farmers are small producers (from 0.70 to 3.52 hectares) with a family income that depends on this product [3]. *La Roya* damaged 37% of the coffee plantations by 2013 and caused losses of 60 million USD in the harvest of 2012-2013 [4].

The results of this study will give Nicaraguan biologists, agroecologists, and coffee farmers a better understanding of *La Roya*, providing them opportunities to develop better methods to manage the problem.

Physical Sample Collection

We took soil, water, and sand samples in Iceland, and soil and leaf samples in Nicaragua. The specific protocol for each material described the order of operations, storage, and cleaning procedures. For example a 10% bleach solution was specified for cleaning tools and latex gloves for people handling the physical samples. Each sampling protocol also included defined roles for each of the people required for a team to sample that material, typically: two for physical sampling, one for storage, one for tool cleaning, and one for recording digital samples.

In order to obtain soil samples that were uncontaminated by modern DNA at the archaeological site at Skálanes, we worked with geologists and designed a three-tool protocol. First a 7cm x 25cm hole was augured (A), that hole was carefully lined with a 6cm OD PVC tube (B), and finally an approximately $4cm^3$ soil sample was carefully removed from the bottom of the hole with a 30cm long soil corer (C) and placed in a labeled sterile 50ml Falcon tube. The Falcon tube was placed in a labeled bag on ice in a cooler. See Figure 3.

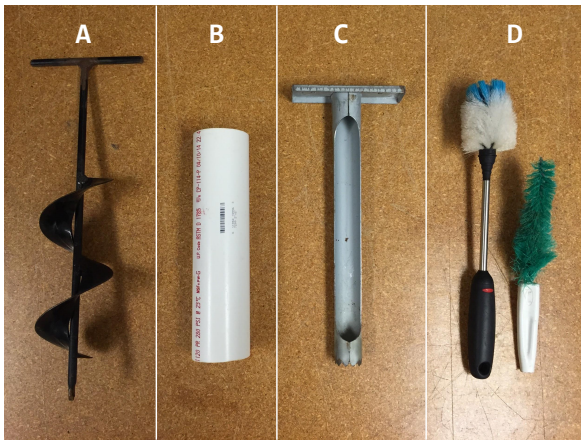

Figure 3: Devices for taking ancient soil samples. A: Soil auger. B: PVC tube. C: Soil corer. D: Cleaning brushes.

For our work in Nicaragua, we developed an application that, given our starting geocode on a farm as the origin, randomly generated a heading and distance to five sampling spots on that farm. At each spot, we took soil and leaf samples from the nearest healthy tree and the nearest infected tree, soil metadata, as well as photos of each tree.

Digital Sample Collection and Curation

The collection, organization and curation of metadata has been a first-class component of this project from the start. We described, structured and defined the metadata as we developed the physical sampling and lab protocols for Iceland and Nicaragua. We organized our metadata following a site-sector-spot hierarchy: a site is composed of one or more sectors, a sector is composed of one or more spots, and each spot corresponds to a single set of physical and digital samples. In Nicaragua, each farm was a site, farms with very different micro-climates were given more than one sector,

Figure 4: An ambiance platform being calibrated before sampling on Eldfell, the volcano on Heimaey, an island located off the Southern coast of Iceland.

ID	Site	Spot	pH	Latitude	Longitude
21	Wetlands	Northeast	6.9	+65.2922	-13.7058
...					
25	Heath	Northeast	6.1	+65.2933	-13.7052
...					
29	Archaeological	North inside wall	5.9	+65.2938	-13.7050
...					

Table 1: Example of soil metadata from Iceland.

and each tree was a spot. In Iceland, each of the heath, wetlands, and archaeological dig were unique sites. An example can be found in Figure 1.

We used the ambiance platform, which contains sensors that measure carbon dioxide, altitude, temperature, humidity, barometric pressure and volatile organic compounds, to provide an overall environmental and weather context to the field work (see Figure 4). The sensor platform connects to a Nexus tablet running Seshat, a locally built Android application. The Nexus hardware provides geographic coordinates which are combined with sensor values and stored locally in comma separated values (CSV) format by Seshat.

We developed soil and water sensor platforms based on parameters of interest to the particular problem at hand. The Iceland platform measured water temperature, soil temperature and soil humidity. The Nicaragua platform contained soil temperature, humidity, fertility, and conductivity sensors. Both platforms used Arduino boards with attached sensors and recorded geocoded data, exactly as the ambiance platform, via Seshat. Figure 5 illustrates how readings collected from the sensor platforms and the data collected during the DNA extraction steps come together to create the metadata for the physical samples. The metadata is used in conjunction with the output from Mothur in an R workflow for analysis and visualization.

Methods

Once sites, sectors, and spots were identified in both Iceland and Nicaragua, the physical samples were collected, placed in sterile containers in the field, and kept on ice in coolers until DNA extraction in the lab (often days or weeks later). DNA was extracted using MoBio PowerSoil® and PowerWater® kits [7]. While water and soil samples were extracted using their respective kits, extracting DNA from

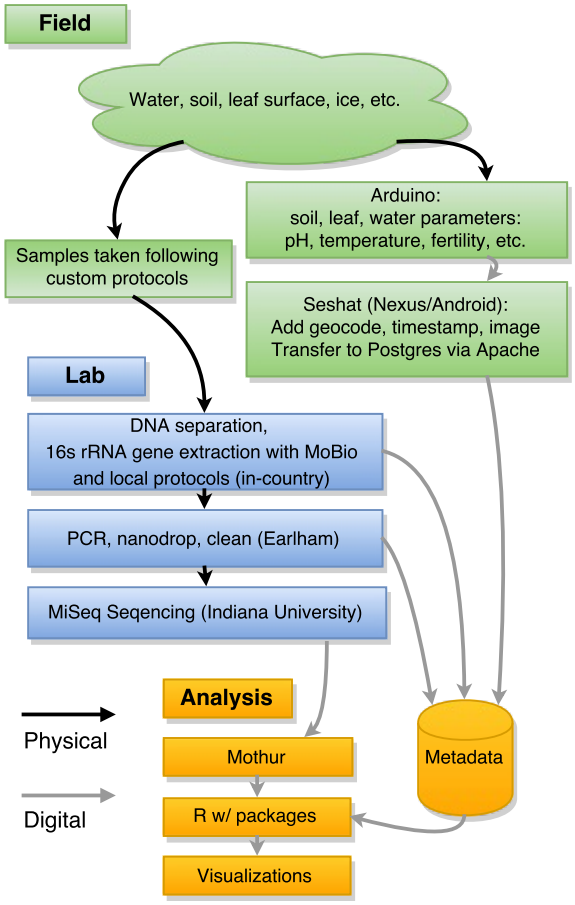

Figure 5: Overall workflow showing the relationship between the field, lab, and computational analysis components of the project, the established and custom protocols, and the physical and digital sample streams.

sand does not have a definitive kit. We developed a protocol for this type of extraction. The sand was mixed with DNase-free water and processed with a water kit. Similarly, there was not a definitive method for extracting DNA from the coffee leaf samples taken in Nicaragua. The leaves were processed using both PowerSoil® and PowerWater® MoBio extraction kits. We homogenized the leaves using a blender and treated the resulting substrate as soil that could be processed with PowerSoil® kits. The blended leaves were suspended in a buffer and put on a chilled shaker overnight to collect any bacteria living in and on the leaves. The buffer from the leaves was then processed using the MoBio PowerWater® kit. For the ancient samples, the soil was shipped to the Ancient DNA lab at Earlham College. Ancient DNA is more prone to contamination, so extraction must be performed in a USDA-certified and completely sterile lab. While many of the individual tasks were based on established protocols, the overall approach and some of the low-level methods were developed as part of this work. For example, we designed and built a sampling device which enabled us to extract ancient soil samples without contamination from the archaeological site, and a small application for selecting random sampling spots within a given range of

the current location. Figure 5 illustrates how the established and custom protocols fit together, and the flow of physical and digital data streams.

Processing

All prokaryotic organisms possess the 16S rRNA gene, but it is different enough in each such that the gene can be used for identification down to the species [16]. The 16S rRNA gene encodes for a specific section of prokaryotic ribosomes composed of RNA. This portion of RNA is vital to the function of the ribosomes, which is why the gene is present among all prokaryotic organisms.

International regulations on shipping organic matter make doing so expensive and complicated. To avoid this, we extracted DNA from the samples in-country, at the University of Akureyri in Iceland and the National Autonomous University of Nicaragua-León. DNA was extracted from all modern soil, water, sand, and leaf samples that were collected in the field. The 1.5ml tubes of DNA were transported back to Earlham College in a custom-built carrier that incorporates a sealed "cool cell", which held packaged frozen vegetables to maintain the quality of the DNA. Frozen vegetables rather than, water or dry ice were used because they can be taken through carry-on baggage screening for domestic and international flights.

The extracted DNA was primed for the genes encoding 16S rRNA, and then underwent polymerase chain reactions (PCR) – a technique used to amplify specific genes of DNA in a sample. The primers contained specific barcodes making it possible to trace samples back to their particular sample spot during post-processing. The DNA was then sent for sequencing at Indiana University using the Illumina MiSeq platform.

Results and Analysis

Mothur is a popular open-source software package for biological sequence analysis. We use it to process and classify the Illumina MiSeq output from the 16S rRNA gene sequences extracted from the soil, sand, and water samples collected in Iceland and Nicaragua. Our Mothur workflow is patterned after one published by the Schloss Lab at the University of Michigan for processing Illumina MiSeq output [12].

The numerous types of output were used to create a plethora of visualizations, allowing one to find correlations among the different sets of data. We used R[13], a tool for statistical and graphical analysis of datasets, to create visualizations of the Iceland Mothur output. We created two types of visualizations: treemaps and dendrograms.

The treemap in Figure 6 illustrates the abundance of phyla for each site in relation to each other – this makes it easier when analyzing the similarities and differences of each site per phyla. The dendrogram in Figure 6 creates a hierarchical clustering of the samples using a distance matrix. The larger distance between the legs reflects the difference in samples. This shows that the microbial makeup of the samples taken outside the wall in the archaeological site is more similar to that of the heath, than the microbial makeup of samples taken inside the wall. The Y-axis of the dendrogram measures the distance of total quantity, and the X-axis measures similarity between sites.

The treemaps in Figure 7 visualize the normalized relative abundance of phyla for each site. The size of the square indicates the percentage of phyla, thus, the larger the square, the greater the abundance of that phylum. In the treemaps for each sample, Acidobacteria, Proteobacteria, and unclassified are the three most abundant phyla for all four sites.

3. PEDAGOGY

Themes: Active Learning, Open-Ended, Multidisciplinary

Multidisciplinary research is a "juxtaposition of various disciplines" [11]. The research presented here embodies this, with the collaboration of many disciplines including computer science, biology, ecology, archaeology, and geology. While student representation for each discipline was ideal, it is sometimes difficult to accomplish. Where there was a lack of student representation in a particular discipline, one of the students from another field took initiative and learned the issues, consulted with relevant faculty, and worked with the group to develop a solution. This is a key component of what makes multidisciplinary, active learning based research successful.

Each discipline had its own sub-projects to tackle, each typically a significant slice of the whole. Ecology was used as a point of reference for the invasive species, bird sanctuary, and flora and fauna of the current era. Biology was used in the DNA extract component, metagenomics for the analysis. Geology provides context to the soil samples including the much longer time-line associated with the archaeology site, and to discover the time-line in relation to depth at which a soil sample was collected. Computer science was used as a tool for data collection via sensors, data organization, and computational analysis. These sub-projects came together to create a collection of results in the form of an organized and easy-to-understand visualization.

Each person focused primarily, but not exclusively, on work in their discipline. However, it was important that all members of the group understand the work of all the disciplines and how they interact. The students in each area presented their progress to the group at weekly meetings, promoting cross-discipline discussions among the larger group. These discussions leverage the strengths of different disciplines. For example, a computer scientist, whose discipline can often be more creative than scientific, might find inspiration in hearing about how a microbiologist struggles with a rigid set of lab protocols. In contrast, a microbiologist might find it intriguing to watch a computer scientist just dive in and start writing code, and having it fail repeatedly before ultimately succeeding, without a known-to-be successful and documented plan to begin with.

To promote active learning the vast majority of the work was conducted by the students with only "guard-rail" style faculty guidance. They not only had to discover solutions, they also had to discover the problems and develop the questions as well. The questions encompassed metagenomics in different disciplinary contexts. The way they approached problems varied from student to student, with some taking a more analytic approach and others a more experimental one. By their own accounts, the students were able to retain and understand more from active learning compared to traditional teacher-student dialogical learning. Active learning promoted total immersion in the project.

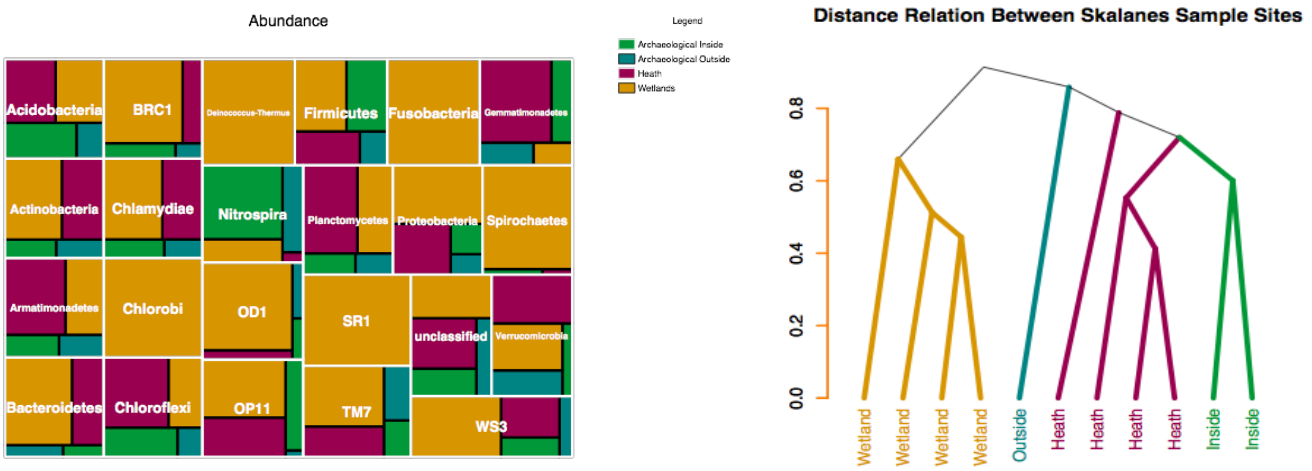

Figure 6: Composite abundance treemap and dendrogram for all sites at Skálanes.

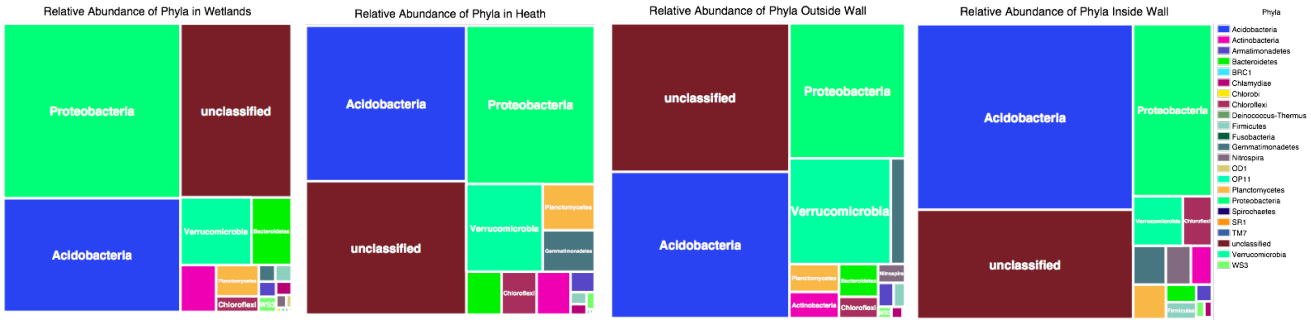

Figure 7: Separate abundance treemaps for each site at Skálanes.

Curriculum Module for BioInformatics Course

The analysis and visualization techniques employed with the 16S rRNA gene for these projects form the basis for a metagenomics curriculum module in Earlham's BioInformatics course. This upper-level course was developed by a biologist and a computer scientist and is team-taught by them to a mixed audience of roughly equal numbers of students from each discipline. Much of the work for the class is done in pairs which are almost always composed of a mix of the two disciplines. This module required the students to develop the scientific questions, process the raw output from the Illumina MiSeq device of the Iceland soil samples with Mothur, and then use R, Python and libraries in conjunction with the metadata to do the analysis and visualizations.

We have observed a few patterns in the feedback students provided for this module. One is that working with data that fellow students collected and curated is much more motivating than using "toy" data sets. The data set for this module was composed of about 481 million bases (R1 + R2 reads) which students analyzed with Mothur on a local cluster. The Mothur workflow took approximately 12 hours of wall-time to complete. This led to the observation that working with realistically sized data sets, while sometimes frustrating, gave them the opportunity to experience first-hand what all the hub-bub around "big data" and the explosion of genomic data really means for working scientists. Lastly, the open-ended, active-learning approach led to comments similar to this one from the most recent offering of the class:

> I had a conversation with a friend of mine who took a similar course at another small liberal arts school, he said that his class involved a lot of worksheets and problem sets that involved doing exercises in Python or R. His class focused so much on teaching the basics through simple worksheets and exercises, that he didn't get to work on any interesting problems or questions. After having this conversation, I was appreciative of the fact that I was working with real data sets, and forming hypotheses in order to ask biologically relevant questions. Although the workload was at times overwhelming, in the end I think it is worth it. I would rather work hard and learn a lot from working on interesting questions than spend the whole class doing exercises about the basics of Python.

Sensor Design and Construction Course

In May of 2014, a three and a half week intensive course on sensor design and construction using Arduino and Intel Galileo boards was taught at Earlham. The diversity of the students was broad – encompassing a variety of academic backgrounds, including physics, computer science, and geology. Participants were asked to brainstorm and research

environmental monitoring platforms and construct them using Arduino or Intel Galileo boards, and sensors from SparkFun and other sources. The devices included soil temperature and humidity reading with an integrated LCD display, earthquake detection using multiple sensor types, and inexpensive portable water quality testing.

Student/Faculty Research

During the past year, and continuing for another, two students (two of the authors of this paper) are participating in the Blue Waters Petascale Internship program. The students are working on improving the scalability and parallelism for the open-source metagenomics package Mothur, specifically with large data sets such as the one which will be generated by the leaf and soil samples from Nicaragua. Currently, many of Mothur's components use *fork()* to achieve much their parallelism. This is problematic for Blue Waters and other national HPC resources whose run-time environments do not support *fork()*. One goal of the internship is to address this by re-engineering those components of Mothur to replace that approach while preserving the attendant parallelism.

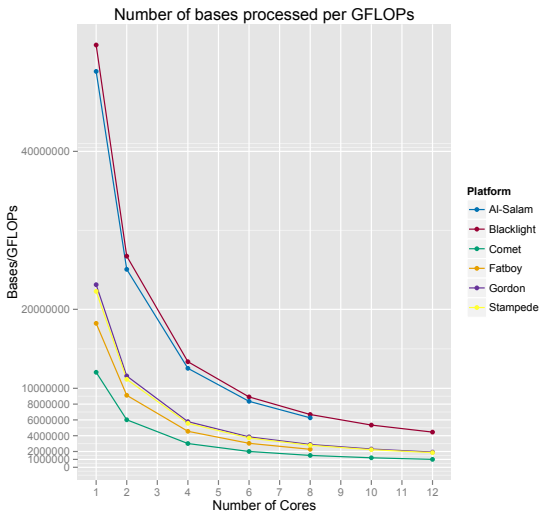

Figure 8: Strong scaling for the Iceland 2014 dataset (≈ 481 Mbases) on a variety of HPC platforms.

Sense of Place

We have worked in different places and collected gobs of data from multiple streams as part of our work. It is often useful to take a step back and look at the bigger picture of where we have worked and the resulting impact.

In Nicaragua, we worked with students and faculty who otherwise do not have access or the opportunity to do this kind of field research. We taught them field and lab methods, supporting hardware/software and computational analysis.

Field training – La Bastilla Technical Centre for Agriculture & Tourism: This ecological educational center provides high school students from local low-income families of the Municipality of Jinotega and surrounding area with the opportunity to earn a technical degree in agriculture and ecotourism, providing Nicaraguan youth with the tools to overcome poverty[5]. During our work at La Bastilla and nearby farms, we trained four students from this center in basic fieldwork, explaining our study of *La Roya* and the work we would do after leaving La Bastilla. Afterwards, they worked with our team collecting the soil and coffee leaf samples. For these students, this was an opportunity to apply science to real problems affecting their communities. It was also an opportunity for us to practice describing our project completely and succinctly, in Spanish, to relative newbies, so they could effectively participate in the work with us.

Lab training – UNAN-León: During DNA extraction at UNAN, we trained two local students in extraction methods for soil and leaf samples. Because of limitations at universities such as UNAN, not all the students in biological sciences can perform such lab activities, making this a unique opportunity for the UNAN students who participated in our work.

In Iceland, the most geologically active country in the world, we are helping a long-running research station that provides opportunities for a wide range of research, education and outreach opportunities for students, faculty, and the public.

Whenever possible, we included local students as part of the in-country lab work. This often required teaching them how to use a micro pipette, a crucial tool for the DNA extraction process. Rather than carry parafilm (the traditional way to learn pipetting), we created a Google Drive sheet with various sized cells for different liquid quantities. Displaying the document on a Nexus tablet provided a glass surface to practice on, see Figure 9. When one student was done, they wiped the surface dry and the next one started.

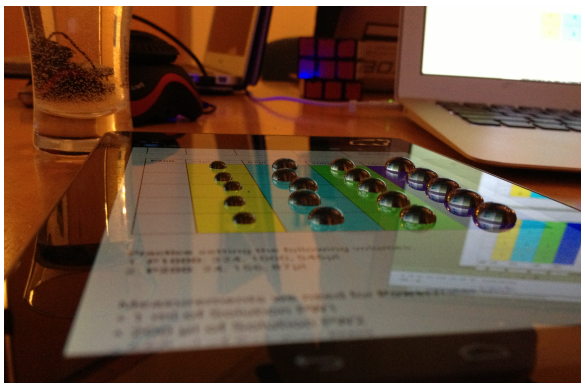

Figure 9: Nexus displaying a Google Drive sheet with cells sized to match the required pipetting quantities.

4. FUTURE WORK

One of our primary goals is to develop on-going relationships with faculty, schools, and students in each of the places we work. Even with months of planning and testing, exactly what will happen in the field is unknown until we are there. Often we encounter unknown unknowns that force us to adjust hardware, software, and/or protocols on-the-fly to meet the needs of the scientists we are working with and their research agendas.

Iceland - Skálanes

Archaeological Site: We plan to return to the site and use 16S rRNA metagenomic analysis to establish a pattern of microbial distribution inside and outside of structures and across the site broadly. With the use of shotgun sequencing for DNA extracted from the samples with ancient DNA protocols, we will look for species that would characterize early agriculture and diet in Iceland (circa 1000CE), e.g. equine, porcine, avian, and piscine.

Avian Sanctuary: *Lupinus nootkatensis*, more commonly known as lupin, is a non-native plant that is rapidly spreading across Iceland, taking over the native flora. The invasion of lupin, alongside years of unsustainable grazing and farming, depleted the natural ecosystem. A large part of coastal heath lands, Skálanes included, are now covered with the plant as a result. This, in turn, has affected the biodiversity. This invasive species encroaches on the habitat of the black-tailed godwit (*Limosa limosa*), a "near threatened" species [6], in the heath and wetlands. We plan to return to Skálanes to map the pattern of microbial distribution across the godwit habitat and survey nest sites with an infrared-equipped low altitude UAV.

Longitudinal Snow Field Study: Year-round, snow fields cap the mountains to the south of the research area. We plan to develop and deploy technology to measure the depth and extent of the snow fields quarterly over many years. This data would be added to the local weather station data and the project would serve as a vehicle for teaching students the techniques of long-term field measurement.

Nicaragua - UNAN

UNAN is working in partnership with Bioversity International to research the Musacea family (plantain and banana). In Nicaragua, plantain is grown in the occident zone of the country (mainly in Rivas, Leon and Chinandega), while banana is more cultivated in the Northern zone where it is grown in association with coffee. UNAN operates three farms at different elevations to study the effects of climate change on coffee, banana and plantain production.

Currently, each of the farms hosting study sites have multiple discrete weather stations. Using hardware and software designs created by Earlham students, we will create weather stations that will automate collection and publication of the data. These weather stations will provide a long-term source for a variety of environmental parameters at these sites for researchers and students. Further support could take the form of facilitating the collection, processing, and analysis of a broad range of metagenomic data across all the UNAN study sites. Changing climate patterns have brought new plant diseases that cannot be explained by plant pathology alone, and a multidisciplinary approach assisted by metagenomics would support a better understanding of the pathogen's behavior.

The work of UNAN with plantain focuses on plant breeding through clone selection. They are selecting to increase production and improve the health and genetic status of the plantations. As part of this study, UNAN is looking at more accurate ways to collect a number of inter-generational data and metadata. Our Android application for sensor data collection, Seshat, could be modified to provide a cost effective solution, while at the same time increasing the types and quality of metadata collected. More data enables more power to make decisions and spot trends, but it also takes more resources to analyze. It is not enough to help UNAN develop ways to collect more and better data. Technical resources and training for data analysis must also be a part of the collaboration, to turn data into information.

5. ACKNOWLEDGMENTS

Samual Kahsay, Earlham College's Spring 2015 BioInformatics class, NSF TUES 1139893, Dr. Kathleen Affholter, Sean Scully, Dr. Chris Smith, Dr. Heather Lerner, and Bernard Lundie. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF ACI 1053575. This research is part of the Blue Waters sustained-petascale computing project, which is supported by the NSF (awards OCI 0725070 and ACI 1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] NCSA Blue Waters system summary, https://bluewaters.ncsa.illinois.edu/hardware-summary.

[2] Skalanes: A place of learning, http://www.skalanes.com/research-learning/.

[3] El café en nicaragua. Tech. rep., Ministerio Agropecuario y Forestal, http://www.monografias.com/trabajos-pdf/cafe-nicaragua/cafe-nicaragua.pdf, May 2013.

[4] Report on the outbreak of coffee leaf rust in Central America and action plan to combat the pest (ed 2157/13). Tech. rep., International Coffee Organization, http://dev.ico.org/documents/cy2012-13/ed-2157e-report-clr.pdf, 2013.

[5] Colegio Técnico Agropecuario. Tech. rep., La Bastilla Ecologe, http://bastillaecolodge.com/colegio.php, 2015.

[6] The IUCN Red List of threatened species, version 2015.1. Report, Bird Life International, http://www.iucnredlist.org, 2015.

[7] Mobio, Inc protocols, January 2015.

[8] Arnalds, O. Soils of Iceland, 2008.

[9] Arnalds, O., and Gretarsson, E. Soil map of Iceland. Map, scale 1:500,000, Agricultural Research Institute, Agricultural University of Iceland, http://www.rala.is/desert/2-1.html, 2001.

[10] Bonwell, C. C., and Eison, J. A. *Active Learning; Creating Excitement in the Classroom.* ASHE-ERIC Higher Education Report No. 1. Washington, D.C.: The George Washington University, School of Education and Human Development, 1991.

[11] Emmelin, L. Environmental education at university level. *Ambio 6*, 4 (1977), pp. 201–209.

[12] Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform.

Applied and environmental microbiology 79, 17 (2013), 5112–5120.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[14] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology 75*, 23 (2009), 7537–7541.

[15] Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., and Wilkens-Diehr, N. XSEDE: Accelerating scientific discovery. *Computing in Science and Engineering 16*, 5 (2014), 62–74.

[16] Woese, C. R., and Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences 74*, 11 (1977), 5088–5090.