First Ultra Conserved Element (UCE) probe-sets for Diptera will decipher the phylogeny and life history evolution in three distantly related fly radiations

Principal Investigator: Eliana Buenaventura (Mauren Turcatel & Torsten Dikow, co-PIs)

Projected Timeline: June/July 2017 start date, 12 month duration

Budget detail: Budget costs for this project include field work, flesh-fly genome sequencing, flesh-fly transcriptome sequencing, MYbaits UCE probe set kit (48 capture reactions to be used for 100 captures), Kapa Hyper library prep kit (for 100 library/samples), NGS adapters, and HiSeq2000 NGS enrichment (2 lanes).

Project Synopsis:
Diptera (true flies) is one of the most familiar insect groups because they are ubiquitous, cosmopolitan, and have had tremendous impacts on human civilization. Morphological and molecular data-sets have advanced our understanding of the phylogenetic relationships within Diptera. To this day, however, the placement, sister-group, and/or monophyly of several taxa continues to be contentious even after the collaborative AToL FlyTree project. Here, we embark on the in-house development of the 1st UCE probe-sets for Diptera. Utilizing these ultra-conserved elements we will propose novel phylogenetic hypotheses for three disparate radiations and study the processes that have led to an increase in taxonomic and life history diversity through evolutionary shifts in adult or larval feeding habits, i.e., Tabanoidea (155 Million years old, 4,600 spp., blood-feeding horse flies versus primarily nectar-feeding athericid flies), Asiloidea (145 myo, 15,000 spp., predatory assassin flies versus nectar-feeding apiocerid and mydas flies), and Oestroidea (50 myo, 17,000 spp., necrophagous-parasitoid flesh fly larvae versus either necrophagous blow fly or parasitoid tachinid fly larvae).

Despite a recent increase in phylogenetic studies on Diptera, the molecular data for horse-, assassin-, and flesh flies are still limited, thus our knowledge about their evolution and diversity is still obscure. The most limiting factor in Diptera systematics has been the identification of orthologous loci across evolutionarily distant taxa, which causes the recovery of poorly supported phylogenetic relationships of the main fly lineages. Our aim is to drastically increase the molecular data for these lineages by developing and utilizing the first two UCE probe-sets for Diptera—one for Orthorrhapha and one for Calyptratae: Oestroidea.

2017 Update:
Our project proposed to drastically increase the molecular data for the fly lineages Tabanoidea, Asiloidea, and Oestroidea by developing and utilizing the first two UCE probe-sets. Designing these UCE probe-sets requires both sequencing genomes and
building a genomic collection of these flies. Accordingly, we sequenced the entire genome for *Sarcophaga (Liopygia) crassipalpis*, which is a model organism commonly utilized to study gene expression and diapause in insects. The genome was assembled in a collaborative effort with Rebecca Dikow, a bioinformatician working for OCIO (Smithsonian Institution). Currently, we are working on annotating this genome, and to facilitate this, we are also generating transcriptomic data. We expect to have a publishable paper on this genome by November 2017. This genome will be used for the design of UCE probes, which will capture genomic data of several species that we recently collected during a field trip to Florida.

This field trip allows us to build a genomic collection. During two weeks of April, we collected Tabanoidea, Asiloidea, and Oestroidea flies in several State Parks and a Biological Station in Florida, as follows: Highlands Hammock State Park, Collier-Seminole State Park, Myakka River State Park, and Dagny Johnson Key Largo Hammock Botanical State Park, and the Archbold Biological Station. In this expedition we collected 191 flies, of which 92 were deposited in the NMNH Biorepository. At least, half of the collected taxa were not represented in the Biorepository. The database of all collected flies is now accessible on the Emu with full taxonomic and collection data associated to them. The remaining 99 flies were pinned as vouchers and deposited in the Entomological collection of the NMNH, which allowed for taxonomic identification and will support future morphological studies. Our project continues, and we expect to have the UCE probe-sets by November 2017, and start producing genomic data of the most relevant species collected by January 2018.

Blog: Forensic indicator flies, their identity and evolution
*By Eliana Buenaventura*

I am Eliana Buenaventura, and I am fascinated with carrion flies. I have a Ph.D. in BioSystematics, and currently, I am a Postdoctoral researcher at NMNH. I study evolution, taxonomy, and ecology of carrion flies. In the last year, I conducted...
biodiversity surveys of flesh flies in several states, and in Denmark, to build a genomic collection and study the taxonomy and evolution of these insects.

There are several insects that colonize and feed on corpses of humans and other animals. Flies are especially attracted to decaying organic matter, where they lay eggs and maggots that feed on the soft tissues until the body is reduced to the bones. These insects are very often the first in finding corpses, and they rapidly start using it as a source of nutrients. For hundreds of years, we have known that flesh flies (Sarcophagidae) and blow flies (Calliphoridae) are useful as forensic indicators of the time of death, and in very specific cases, they can also give relevant information to accurately determine the place and cause of death. However, using flies for forensic research requires determining the fly species with the highest accuracy. This is not an insignificant issue, since the incorrect estimation of the time of death can decide the fate of a potential murderer. Or put in jail an innocent witness.

Identifying flies is not an easy task, as there are thousands of flies in urban and natural ecosystems, but only a few are able to feed on corpses. On top of this, flesh flies and blow flies belong to one of the most diverse insect groups, the Oestroidea flies, which add to the number of potential flies colonizing decaying organic matter. It is important to highlight that by feeding on decaying organic matter, carrion flies also contribute to recycling nutrients in ecosystems, which is essential for the ecosystem’s health and sustainability. Modern cutting-edge techniques ensure accurate identification of flies by using their genetic material, which at the same time provides data for disentangling the evolution of these interesting insects.

Figure 2: Phylogenetic tree decorated with flies of Sarcophaginae. Photo: Eliana Buenaventura
The Global Genome Initiative sponsored this project, which allowed me to spend several weeks in the states of California, Florida, and Washington and in the islands of Zealand and Lolland in Denmark collecting flesh flies. Some of these collecting trips were possible thanks to collaborations with my colleagues Torsten Dikow and Mauren Turcatel (NMNH), Brian Brown (Natural History Museum of Los Angeles County) and Katherine Noble (University of Utah).
Some of the most interesting places to collect were the hilltops. By congregating on the small mountaintops, male flies reduce the area to facilitate finding a female to mate. Once there, males sit on rocks and perch on tree branches. For an enthusiastic fly entomologist, hiking to these places always offers a great landscape view and enormously eases locating and sweeping the net to collect these flies.

Other interesting places to collect flies are in the low lands, where beaches and coastal environments bring also other larger fauna, such as crocodiles and turtles, which are not part of our target species to collect but are interesting visitors during field work.
Collecting carrion flies of the family Sarcophagidae can be done by sweeping the net on hilltops, but also by setting bait such as rotten chicken liver or dog poop, among other decaying materials on the ground. I also use baited Van Someren-Rydon traps.

Flies will be attracted to the bait. But! Researchers should consider the food preferences of flies, because some flies feed on very rotten tissues, while others will go to fresh meat. Thus, some flies arrive earlier when the corpse is still fresh, while others come later when the tissues are softened to almost liquid. The quality of the tissue used for baiting will determine the picture that you capture during the decomposition phases. In my experiments, I used fresh and rotten tissues, in order to capture as much diversity of carrion flies as possible.
I successfully collected more than 300 flies of all target taxa, 50% of which were preserved in liquid nitrogen and deposited in the NMNH Biorepository. The remaining flies (duplicates) were pinned as vouchers and deposited in the Entomological collection of the NMNH, which allowed for taxonomic identification and will support future morphological studies. Thus, I am helping GGI to capture the genomic biodiversity in the guild of carrion flies. At the same time, I use the collected data and specimens for ensuring correct identifications, document their distribution patterns and feeding modes on different habitats, and reconstruct their evolution.

This project also resulted in the sequencing of the first complete genome of a flesh fly. We are currently analyzing and annotating the genome of *Sarcophaga (Liopygia) crassipalpis* Macquart, 1839, which is a model organism commonly used to study gene expression and diapause in insects. This genome will be used for the design of UCE probes.

The specimens preserved in liquid nitrogen and the resulting UCE sequences will be available for ongoing genomic projects, such as the phylogenomic studies of flesh flies, which is the subject of my current postdoctoral project, as well as for future projects of other researchers affiliated with the Smithsonian.
Phylogenomics of the lichen-forming Dictyonema clade: insight into the evolution of an important but forgotten branch in the fungal tree of life

Principal Investigator: Manuela Dal Forno (Eric Schuettpelz, co-PI)

Projected Timeline: October 2016 start date, 12 month duration

Budget detail: Budget costs for this project include purchasing equipment/supplies: Qiagen DNeasy Plant Mini Kit (for 250 samples), primers (220 primer pairs), Fluidigm Reagent Kit, Fluidigm Access Array Integrated Flow Circuit (5 circuits), FastStart High Fidelity PCR System, MiSeq Reagent Kit, and Agencourt AMPure XP.

Project Synopsis:
Lichenization is among the dominant lifestyles in fungi, with approximately 17,500 species recognized and up to 28,000 predicted (Lücking et al. 2009). Almost all described lichens (99.7%) are found in the Ascomycota; however, lichen-forming fungi in the Basidiomycota (basidiolichens) have evolved independently in at least five separate clades (Redhead et al. 2002; Ertz et al. 2008; Lawrey et al. 2009; Hodkinson et al. 2014). Until recently, *Dictyonema* s.l. was thought to include only six species in a single genus, *Dictyonema* C. Agardh ex Kunth (Parmasto 1978; Marcano et al. 1996), but taxonomic and molecular phylogenetic studies have suggested that this number is a gross underestimate of the real diversity in this clade (Lawrey et al. 2009; Dal-Forno et al. 2013, 2015; Lücking et al. 2013, 2014, in prep). *Dictyonema* s.l. is nested within the tribe Arrhenieae Lücking (Lodge et al. 2014) and now includes five lichenized genera, namely Acantholichen P. M. Jørg., *Cora* Fr., *Corella* Vain., *Cyphellostereum* D. A. Reid, and *Dictyonema* sensu stricto (Dal-Forno et al. 2013). To date, studies of lichen phylogenomics have been undertaken only in the Ascomycota, and even these are very limited (Leavitt et al. 2016). A recently published putative single copy gene dataset for Agaricales (Dentinger et al. 2016) will be tested for initial loci mining in the study group, the *Dictyonema* s.l. clade, which has never before been included in genome level studies.

2017 Update: In summary, I designed 144 primer pairs I had proposed to develop based of data from the genome of *Hygrocybe conica*, the closest relative of the *Dictyonema* clade with subgenomic data available. These taxa belong to the same Agaricales family, Hygrophoraceae, but are very distantly related. In the validation process, none of the aimed regions amplified successfully with the new primers when I tried with samples of *Dictyonema*. Nonetheless, many of the new primer pairs have worked for the species they were based from (*Hygrocybe conica*). Even though this might be important and useful in species delimitation in the genus *Hygrocybe*, it didn't help me towards my main goal, to detect more loci to help with species relationships in *Dictyonema*. Unfortunately, there are no more genomic resources available currently that would allow me to take a different approach to design new ones. At this moment, however, I have turned into an alternative also included in the original proposal, that is to use commonly used loci in Fungi and Cyanobacteria, but adapting those to be used...
with the microfluidics PCR technique, since it has never been utilized in lichens. That way, I will still have novelty data for several loci to create multi-locus phylogenies including many specimens to accomplish the objectives of this research project.

Blog: It’s All About Lichens…
By Manuela Dal Forno

After an incredible three-hour workshop in my very first year as an undergraduate student in Brazil, I decided what I wanted to be in my life: a LICHENOLOGIST. A lichenologist is someone who studies lichens. But, do you know what lichens are?

Lichens are symbioses involving fungi (mycobionts), photosynthetic green algae and/or cyanobacteria (photobionts), and a microbiome, which together form the lichen thallus we see in nature (Figure 1). Lichens represent just one of many fungal lifestyles; others include saprotrophs (decomposers), parasites, pathogens, yeasts, and mycorrhizae.

The symbiotic lichen habit has evolved numerous times in a variety of fungal lineages, meaning that lichens do not form a natural biological group. They occur most frequently (over 99% of the time) in the Ascomycota (i.e., ascotichens) and much less frequently in the Basidiomycota (less than 1%, known as basidiolichens).
The lichen lineage that I have focused on since 2010 is a basidiolichen clade usually referred to as *Dictyonema* sensu lato and it comprises five genera: *Acantholichen*, *Cora*, *Corella*, *Cyphellostereum*, and *Dictyonema* sensu stricto (Figure 2). *Cora* is usually the most exciting to see in the field (Figure 3), because of its huge morphological variation, while *Dictyonema* is more fun in the lab (Figure 4), since the diversity lies mainly in microscopic components.

In 2016, I joined the NMNH as a NSF postdoctoral research fellow in the Botany Department, to further my work on species delimitation within *Dictyonema* sensu lato and to specially investigate the hypothesis that historical collections kept in museums can provide insight into lichen microbiomes, even after decades of collection. While collecting data for my main project, I was immediately interested in the possibility of generating new genetic markers for *Dictyonema*, since we have mostly been using the ITS fungal barcoding marker to study the diversity and relationships within these genera.

Although the collection at the U.S. National Herbarium is probably one of the richest in the world for this lineage, the almost 400 specimens are identified as just six different species. My preliminary morphological and ecological assessments, however, have already detected at least 20 new species among these samples!

With funding from GGI, I plan to generate additional molecular data for species delimitation in *Dictyonema* sensu lato. By adding genetic data to morphological, anatomical, and ecological data, I can take an integrative taxonomic approach, which I believe is the strongest way to build robust species concepts. Observations acquired in
the field and in lab are important for generating hypotheses, but molecular data is necessary for testing them. Building strong phylogenies allows us to detect taxonomically important characters, sometimes not noticed before.

While I am in the process of generating and assembling two genomes within Dictyonema, which can be used for my future research and designing primers from newly detected variable regions, I first tried to design primers based on Hygrocybe conica, the only fungus with a whole genome sequenced in the entire Hygrophoraceae family, to which Dictyonema sensu lato belongs. Unfortunately, this species was too distantly related and the primers did not work for Dictyonema. This emphasizes the importance of programs like GGI, because aside from conserving genomic quality tissues and facilitating a network of genomic and knowledge exchange, it also awards researchers trying to understand the evolution of different organisms, generating resources applicable to many research areas.

Due to the current lack of genomic resources in basidiolichens in general, I am going to adapt dozens of typically used primers to amplify different DNA regions utilizing Microfluidic PCR on the Fluidigm Access Array. With this technique, utilized for the first time in lichens, I will be able to generate data simultaneously from the: (1) mycobiont basidiomycete Dictyonema sensu lato; (2) photobiont cyanobacteria Rhizonema; and (3) microbiome bacteria, archaea, and fungi.
With these data, I can address questions related to many areas of interest in lichenology. At this moment, it is especially important because it allows me to infer phylogenetic trees for my studies of lichen microbiome in space and time, with which I will examine evolutionary patterns of the symbiotic components and possible co-evolution with microbial taxa. For whichever research questions one may have, species delimitation is the key to know you are comparing apples to apples, and therefore the simple fact of knowing of a species is remains one of the most important concepts in biology.

![Field trip photos](image1.jpg)

Figure 5: NSF funded collecting trip in the Serra do Itatiaia, Brazil in October 2016. Photo on the left shows the Atlantic Forest, with a view of the high-altitude fields on the right, found in the Pico das Agulhas Negras. Photo: Manuela Dal Forno.

![Field trip photos](image2.jpg)

Figure 6: Landscape in Äkäslompolo, Lapland, Finland, during a field trip after the 8th IAL Symposium, held in Helsinki, Finland in August 2016. Photo by: Manuela Dal Forno.

Even though I spend most of my time in the lab or in front of the computer, one of the perks of my job is that I get to do field work and receive training in many different countries in the world. Brazil is the place I have been the most (Figure 5), but other countries include Colombia, Costa Rica, Ecuador (including the Galapagos Islands), Finland (Figure 6), Puerto Rico, the United States, and Thailand. I am ready to pack my bags for Austria next month, and Panama and Jamaica next year! #WeAreAllLichens
Unraveling co-evolutionary dynamics of frog and fungal genomes to understand infectious disease

Principal Investigator: Robert Fleischer (Kevin Mulder and Anna Savage, co-PIs)

Projected Timeline: January 2017 start date, 9 month duration

Budget detail: Budget costs for this project include Illumina library prep, capture baits, and sequencing.

Project Synopsis:
The global fungal disease chytridiomycosis devastates some amphibian populations but causes minimal damage to others. Whether differences in susceptibility are caused by genomic variation of the host, the pathogen, or co-evolutionary dynamics between both organisms remains unknown. Understanding variation in susceptibility to chytridiomycosis is critical for global mitigation strategies, but is limited by two factors: (1) our lack of genomic resources for amphibians, which have large and complex genomes, and (2) our inability to sequence the pathogen genome of almost all amphibian infecting strains, as most cannot be isolated and grown in pure culture. Whole genome analyses of culturable strains of the pathogen, *Batrachochytrium dendrobatidis* (Bd), have uncovered low genomic variation among virulent strains and have not been able to identify a region of origin. However, genomic diversity within and among unculturable Bd strains is unknown. Similarly, immunogenetic and gene expression studies of frog populations have uncovered candidate genes for Bd tolerance, but we lack genome-wide association studies. We therefore propose to simultaneously characterize genome-wide polymorphisms of host and pathogen genomes from five North American frogs in the genus *Lithobates* across different environmental regimes.

2017 Update:
We have completed the field sampling and have built Illumina libraries for most DNA samples which are ready and waiting for the sequence capture hybridization step with both the fungal and frog probes. The fungal baits have been ordered from Microarray and will arrive this month and the frog probes are in the final stages of development. We expect to complete the hybridization and the sequencing by the end of October and work on the bioinformatics analyses in the final three months of 2017. Publications will be prepared and submitted in early 2018.

Blog: Unravelling co-evolutionary dynamics of frog and fungal genomes to understand infectious disease

The Nile will teem with frogs. They will come up into your palace and your bedroom and onto your bed, into the houses of your officials and on your people, and into your ovens and kneading troughs.
— Exodus 8: 2–3
In historic times frogs were much more abundant and even considered to be plagues, but nowadays quite the opposite is true and frogs have been declining worldwide (Stuart, 2004). Habitat loss and pollution have been two of the main drivers of the declines and many frog species are currently threatened with extinction. Except for some species like the common American bullfrog that have adapted to more anthropogenic habitats, the majority of species are now restricted to what little natural habitat is left.

It was therefore a shock in the early 1990s when scientists found dead frogs even in pristine natural habitat, and forests once loud with frog-calls suddenly went silent. It was not till later that decade that a team of scientists, including Smithsonian researchers, found the cause to be an infectious disease caused by a newly discovered fungal pathogen they called *Batrachochytrium dendrobatidis* (Bd for short) and managed to isolate the fungus from a frog that died at the National Zoo. The fungal strain causing the majority of the die-offs (Bd-GPL) has since been found across the world, infecting and killing frogs and driving many species to the endangerment and extinction.

The environment plays a major role in the disease process, with especially temperature affecting fungal growth and consequently disease outcome. However, even in the same environment some frogs succumb quickly to the pathogen while others keep hopping along and carrying them potentially infecting other frogs along the way. This difference is due to two main factors; (1) the virulence of the infecting fungal strain, and (2) the frog’s genetic background and immune system.

In order to better understand these processes, our team at the Smithsonian Center for Conservation Genomics (https://nationalzoo.si.edu/center-for-conservation-genomics) and at the University of Central Florida (http://sciences.ucf.edu/biology/annasavage/) has been looking at the DNA of the host to try identify the important parts of the genome that confer genetic resistance to the fungus (Mulder et al., 2017). We have mainly focused on the genus *Rana* due to the strong differences found in disease outcome both across and within species. For example, the American bullfrog (*R. catesbeiana*) is very resistant to the disease, whereas its neighbour in Arizona, the closely related Chiricahua leopard frog (*R. chiricahuensis*), is highly susceptible and currently
considered endangered. And even within the same species there are differences in individual survival found to be related to certain genetic markers (Savage et al., 2017) and immune processes (Savage et al., 2016 & unpublished).

One important missing component in our study has been the fungal genome whose information until now has been solely dependent on the subset of strains that willingly grow in a petri dish in the lab. Until now, nobody has screened the genomes of the fungal strains that do not grow in the lab and we are thus missing a large and important part of the puzzle. Due to generous funding by the Global Genome Initiative we were able to set out and investigate both the frog and fungal genome in parallel, using state-of-the-art sequence capture techniques to investigate both host and pathogen genomes and to complete the triangle and better understand disease outcome. This investigation will allow us to get a more complete picture of how the fungus kills the frogs and pinpoint what parts of the frog genome help certain frogs survive. It will also help us sequence and describe all the fungal diversity living on the frog skin that is yet unknown to science.

We have sampled frogs across three distinct environmental regimes in the US (Arizona, Florida and New York) and collected tissues from both resistant frogs (R. catesbeiana, which occurs in all three states) as well as susceptible frogs (R. yavapaiensis in AZ, R. sphenophala in FL and R. clamitans in NY) for several locations. Frogs were measured, inspected for any signs of disease, sampled for their DNA, and subsequently released back into the wild.

We have since extracted DNA and prepared Illumina sequencing libraries for the majority of samples in our genetics lab at the National Zoo. We have also designed two array’s of 20,000 probes each, for both the frog and fungal genome. These arrays were specifically designed to target DNA variation across the genomes with a focus on candidate genes and important mutations using a variety of different bioinformatics software package and publicly available data. Once the RNA probes arrive in the next few weeks we will sequence capture both the host and fungal genomes in parallel at our lab and send out the resulting genetic libraries for high-coverage sequencing.
Our data will help describe the fungal pathogen and its virulence better and help us identify the parts of the frog genome that allow some frogs to survive while others perish. In the future this can help both zoos as well as wildlife managers better screen and manage frog populations to ensure more species survive the epidemic and continue to thrive on our planet. Given the importance of frogs in eating and controlling the other plagues like mosquitos and flies and in maintaining the ecosystem balance we hope our work will help these beautiful creatures survive the this epidemic.

The funding by the Global Genome Initiative has allowed us to take the step from looking at single loci exclusively in the host, to investigating 1000s of loci across both the frog and fungal genomes concurrently for different environmental systems. We hope that this triangular approach will not only help the frogs, but will also serve as an outline for other researchers investigating the genomic background of host-pathogen dynamics.

Bibliography:


‘Symbioses on the Rocks’: Exploring Microbial and Nutritional Couplings of Coral Reef Architects

Principal Investigator: Laura Núñez Pons (Andrew Altieri, PI Supervisor)

Projected Timeline: September 2016 start date, 12 month duration

Budget detail: Budget costs for this project includes equipment/materials/supplies and bench fees.

Project Synopsis:
Symbiosis is key for survival and evolution, and this is particularly true in coral reefs. Endosymbionts provide nutrients and bioactive metabolites that confer adaptability and ecological competence to metazoan hosts, permitting niche expansion and environmental resistance. Due to global climate change and other impacts, symbiotic partnerships get disrupted, thereby affecting the stability of whole ecosystems. After such mutualistic breakdowns, there may be resistant taxa that through genetic or non-genetic changes (epigenetics) can recover and acclimate to stress, with the potential for tolerance to be transmitted across generations. Among epigenetic units providing acclimatization are microbial symbionts, but the taxonomic composition and functional mechanisms that lead to ecological competence are poorly understood. In the marine realm, sponges and corals are major players in the foundation of benthic frameworks that provide housing to one of the richest biodiversity, the marine reefs. A hypothetical coral-sponge coupling with symbionts performing internal “microbial loops” for nutrient recycling is suspected to be the process closing up this vital circle that nourishes these relevant oligotrophic ecosystems, but the evidence remains ambiguous.

2017 Update:
In November 2016 we did our trip and stage at Bocas del Toro Station in Panamá and performed the aquaria experiments. At the beginning of December, we extracted the DNA and left the DNA samples at Kristin’s Saltonstall Lab for MiSeq sequencing. We are still pending to run the sequencing for all amplicon targets (16S, ITS1, ITS2) on the optimization of the PNA clamping protocol that we are developing now. We have done Sanger on our sponge and coral hosts and we have already designed a working PNA for sponges, while we are still optimizing the corals PNA or primer blocker (not easy!). Regarding the isotopes, sample were processed for host/symbionts cell separations in December, and the analyses have a major delay because the Truner Lab at STRI (where the samples should be analyzed) had lost the samples for some months, but luckily there were found in the back of their freezer last mid-July. For the metabolomics samples are being chemically extracted and analyzed on HPLC and NMR to optimize the conditions for the HPLC-MS, and in September we hope to start the metabolomics profiles of the most significant family of compounds. We estimate all analysis to be completed by December 2017, the Methodology paper for fungal characterization to be submitted by December 2017, and putting all data for other publications together by July 2018.
Blog: Symbioses on the Rocks: Exploring Microbial and Nutritional Couplings of Coral Reef Architects

I am Laura Núñez Pons, a marine biologist working at Stazione Zoologica Anton Dohrn (SZN) in Naples (Italy), and at the Altieri’s Lab at Smithsonian Tropical Research Institute (STRI) in Panamá. My recent research focuses on the metabolic roles of microbiomes in nutrition, production of bioactive compounds and adaptability to environmental changes and stress. Our studies use mainly sponges, and other invertebrates like corals to estimate the weight of microbes in health, immunity and ecological competence.

In our trip and stage at Bocas del Toro Station in Panamá we performed an aquaria experiment with six of the most abundant taxa in the foundation of Bocatorenian marine reefs, three hard coral (Agaricia tenuifolia, Porites porites, Siderastrea siderea) and three sponge (Cliona delitrix, C. varians, Chondrilla nucula) species. Such species were selected as model holobiont experimental systems for our research for their ecological relevance. These species are two by two space competitors in the reef and are likely to interact among each other with a sort of “chemical war” of allelochemicals.

We started collecting replicates of these six species of benthic organisms by scuba diving. We first placed them in large aquaria, divided them in clones for the experimental design, and left them to heal for two weeks with running seawater pumped from the ocean.

In the meantime, we built a circulating aquaria system that would feed several aquaria by gravity from mother tanks. The head tanks were treated with UV light pump treatments to kill most microorganism particles in the water column, without eliminating the organic material. This water was then by gravity distributed to feed smaller aquaria. Here we demonstrated ourselves we could also be part of a Mario & Luigi Bross company for marine pumping systems...!

Example of the space competition of the coral Siderastrea siderea (brain coral) and the bioeroding sponge Cliona delitrix (red sponge) in the reefs of Bocas del Toro, Panamá. Photo by scuba diving with camera in underwater chamber by Laura Núñez Pons.
Sponges and corals healing in large aquaria. On top left *Porites furcata* (finger coral) and *Siderastrea siderea* (brain coral); on top right *Cliona varains* (photosynthetic boring sponge); on bottom left *Cliona delitrix* (red boring sponge); on bottom right *Chondrilla caribensis* (chicken liver sponge) and *Agaricia tenuifolia* (lettuce coral) at the wet lab of Bocas del Toro Station, Panamá. Photo: Laura Núñez Pons

Left: Systems aquaria for the experiments built with head tanks and tubes tubing pipelines that distributed treated to feed smaller aquaria. The system was set at the wet lab of Bocas del Toro Station, Panamá. Right: Example of coral clone fragments untreated (on top, looking tan) versus clones treated for obtaining bleached holobionts (on the bottom of the picture), for the 3 coral target species selected (*Agaricia Tenuifolia, Porites porites, Siderastrea siderea*) at the wet lab aquaria of Bocas del Toro Station, Panamá. Photos: Laura Núñez Pons.
Once clones from sponges and corals were healed, we distributed the fragments into experimental groups and exposed the corresponding groups to combined antibiotic manipulative treatments and menthol/light manipulations to create holobionts with modified (distressed) microbiomes (aposymbiotic), mimicking what holobionts experience under environmental stress. Treatments lasted for about 10 days.

After the treatments were concluded we build mosaics of all combinations of manipulated treated and untreated coral and sponge holobionts in aquaria. In the aquaria the competitor pairs were put in forced contact for ~two weeks to monitor the potential capability of allelochemical responses.

After two weeks of competitive contact, all fragments were sacrificed and subdivided in samples for:
- Microbiome characterization by metagenomics of 16S, ITS1, and ITS2 for Bacteria/Archaea, Fungi and Symbiodinium diversity analysis on MiSeq
- Isotopic analyses of the C and N, and when possible separating host cells from major symbionts cells
- Metabolomic analyses

For the central metagenomic approach most interesting aspect for the GGI program we will sequence libraries of 16S, ITS1, and ITS2 for Bacteria/Archaea, Fungi and Symbiodinium diversity analysis on MiSeq, in an effort to examine whole microbiome network compositions related to healthy and distressed states. For the ITS1 analysis we are developing a new methodology by applying a ‘PCR clamping’. We have sequenced on Sanger our host target genomic fragments, and we are designing a PNA (Peptide Nucleic Acid) probe specific against the sponge and coral
ITS regions. This probe binds very strongly to hosts’ DNA and avoids primer coupling successfully blocking undesired co-amplification of host amplicons.

![Diagram of PNA-PCR clamping](image)

<table>
<thead>
<tr>
<th>Species</th>
<th align="right">PNA Probe Sequence</th>
<th align="right">/**/ Tm</th>
<th align="right">Tm Probe Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Scel_14 5.8s</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right">45.1°C</td>
<td align="right">GGAGAGAGATCCACC</td>
</tr>
<tr>
<td><em>Most Fungi</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right">45.1°C</td>
<td align="right">GGAGAGAGATCCACC</td>
</tr>
<tr>
<td><em>Tm Fungi</em></td>
<td align="right">36.3°C</td>
<td align="right"></td>
<td align="right">45.1°C</td>
</tr>
<tr>
<td><em>Most Scelactinia</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right">55.1°C</td>
<td align="right">GGAGAGAGATCCACC</td>
</tr>
<tr>
<td><em>Tm Scelactinia</em></td>
<td align="right">36.3°C</td>
<td align="right"></td>
<td align="right">55.1°C</td>
</tr>
<tr>
<td><em>Other Scelactinia, that mismatch</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right"></td>
<td align="right"></td>
</tr>
<tr>
<td><em>Paragoniastrea_deformis</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right"></td>
<td align="right"></td>
</tr>
<tr>
<td><em>Stylophora_sp.</em></td>
<td align="right">TTTGGGCCGAGATCCACC</td>
<td align="right"></td>
<td align="right"></td>
</tr>
</tbody>
</table>

Diagram of how PNA-PCR clamping works for characterization of fungal communities associated to corals and sponges, avoiding the enormous co-amplification of host DNA (on top). Table with an example of our host organisms Sanger sequences and primers to build the PNAs against our taxa, by Richard O’Roerck & Laura Núñez Pons.

Our project will contribute to exploring Earth’s genomic biodiversity by revealing the whole microbiome composition of ecologically important marine benthic organisms. This will be particular impact for the highly unknown fungal symbiotic compartment. Additionally, the new genetic diversity data will be accompanied by functional and chemoecological information for a mechanistic understanding of mutualistic relationships in a changing world.

Symbiosis is key for survival and evolution. Endosymbionts provide nutrients and bioactive metabolites that confer adaptability and ecological competence to metazoan hosts, permitting niche expansion and environmental resistance. Due to global climate
change and other impacts, symbiotic partnerships get disrupted, thereby affecting the stability of whole ecosystems. Projects like ours that intend to explore on healthy and distressed microbiomes are key to understand this fundamental biological interaction under the prospective of a changing planet. As in humans, microbiome treatments can take part of the tools projected to repopulate and recover disrupted marine ecosystems. We expect that scientifically, our intention to design a new tool to characterize the associated fungal communities in corals and sponges will provide a very useful tool to explore into this significant group of symbiotic organisms. Co-amplification is a major issue in the characterization of marine fungal symbiont communities that has greatly hampered their study up to date, with authors reporting >50% loss of targeted sequences. This protocol that we propose will be a very useful contribution as will allow a sequencing depth necessary to characterize fungal communities, which up to now have been greatly overlooked in marine organisms.

Our main objective is to characterize the microbial community compositions that afford ecological competence to major reef-forming organisms, and determine the contribution of symbionts in nutrition, and metabolite profile. The manipulated and control corals and sponges will be monitored for microbial shifts, isotopic and chemical parameters, aiming to elucidate phenotypic patterns related to healthy microbiome, recovery, spatial competition and nutrition.
Illuminating the drivers, tempo and history of body plan evolution in Chaetopteridae (Annelida using targeted exon capture for phylogenomics)

Principal Investigator: Karen Osborn (Jenna Moore, Co-PI)

Projected Timeline: August 2016 start date, 12 month duration

Budget detail: Budget costs for this project include specimen purchase, DNA extraction and quality assessment, Illumina library prep and purification, Illumina dual-index adapters, Library quality assessment, MYbaits-1 Custom 20,000 Probe kit (48 samples), Dynabeads MyOne Streptavidin C1 beads, 10mM Tris-Cl, 96R Ring Magnet plate, Post-capture library amplification kit, and Illumina MiSeq 300-cycle paired-end sequencing.

Project Synopsis:
The origin and evolutionary consequences of major changes in body plan is a key question in macroevolution. The proposed research will generate a robust phylogenomic framework for research into the evolution of body plan in a uniquely tagmatized annelid clade, used widely in studies of development, biomechanics, and bioluminescence: Chaetopteridae. Their three-part body plan has allowed functional morphological specialization within homologous tagmata for diverse mucus-net filter feeding mechanisms and associated structures. These singular features make Chaetopteridae an excellent model system for investigating how the evolution of functional morphology has influenced diversification rate and radiations into new habitats. Recent phylogenetic analyses of Annelida have included representative Chaetopteridae, but have not resolved with sufficient support its location in the tree, nor relationships within the group. This poor resolution is due to low taxon and gene sampling, tree incongruence, and low support at key nodes. Testing evolutionary hypotheses on body plan, diversity, and distribution requires a well-resolved phylogeny reflecting the diversity of the family.

2017 Update:
Chaetopteridae are a family of approximately 100 species of marine annelid worms. They are distributed globally and have unusual and diverse body plans with functional specializations for filter feeding using one or more mucus nets. This fascinating group presents a natural experiment for investigating feeding mode and body plan evolution in an understudied major phylum of animals. The goal of this project is to generate a well-resolved phylogeny of the family Chaetopteridae as an evolutionary framework to trace the pattern of body plan evolution in the group. The phylogeny is based on a method to reduce genome-scale data to 900 exon regions using a custom-designed set of bait molecules based on sequence data from transcriptomes. The targeted regions are captured for individual species and then pooled for next-generation sequencing. This approach allows targeting many hundreds of specific genes and reduces the cost and effort to generate a well-resolved phylogeny.
The final input to the bait capture included 46 species of Chaetopteridae, and the final sequence dataset will include 50 species, including the type species of three of the four genera: *Chaetopterus*, *Mesochaetopterus*, and *Spiochaetopterus*. Type taxon material for *Phyllochaetopterus* was pursued, but unfortunately no appropriately preserved material was available. Morphological character data for all 50 species has been collected and will be used to trace the evolution of functional morphological structures relating to feeding mode in the data analysis stage of this project. Extracted DNA from 43 species will be deposited in the GGI Biorepository, and most of these species are new representatives in the collections.

**Timeline:** The laboratory work for this project is nearing completion, and the pooled capture products will be ready for sequencing by August 11, 2017. Sequencing on the in-house LAB Illumina MiSeq is scheduled for August 16th, 2017. Sequence data will be processed and analysed in September and October 2017, and phylogenetic analyses will be completed by December 1, 2017. Sequence data will be submitted to GenBank by December 1, 2017. The manuscript resulting from this study is expected to be submitted for publication by February 1, 2018.

**Significance:** This project is an exciting study on the evolution of morphology in an understudied major phylum of animals with diverse functional morphologies, and will contribute a valuable test of applying targeted exon capture methods to an ancient lineage of animals with a distantly related reference genome. Relating functional morphological characteristics to environmental variables will give insight into factors driving evolutionary diversification and body plan specialization. The phylogeny will be applicable to other areas of research by aiding much-needed systematic revisions of the family and by identifying a species of *Chaetopterus* used extensively as a model organism in developmental biology.

**Blog:**
My name is Jenna Moore, I am a Ph. D. student at the University of Florida. My dissertation focuses on a fascinating, globally distributed family of tube-dwelling marine annelids, Chaetopteridae. I am working at the NMNH with Dr. Karen Osborn on a project to resolve the phylogeny of the family using targeted exon-capture genomic methods. The method we are employing reduces the genome to 900 exon regions using a custom-designed set of bait molecules based on sequence data from transcriptomes. The targeted regions are captured for individual species and then pooled for next-generation sequencing. This approach allows targeting many hundreds of specific genes and reduces the cost and effort required to generate a well-resolved phylogeny.

There are about 100 described species in the family Chaetopteridae. Chaetopterids have an unusual body plan among polychaetes; they are tagmatized, meaning that they have regionally specialized structures along the length of their bodies. Chaetopterid bodies are divided into three distinct regions, with the middle section specialized for filter feeding using a mucus net. Some chaetopterids also use long anterior appendages called palps to feed on particles on the surface of the sea floor, or to capture floating particles directly. The degree of specialization to these feeding mode types varies across the four genera, and makes for an interesting natural experiment to understand
how body plans have evolved in this family, and how local environmental conditions influence feeding mode and specialization. Our approach will generate a well-resolved phylogeny to serve as a framework for integrating evolutionary information with morphology and habitat data in order to understand the evolution of body specialization in this group.

Above left: Chaetopterus pergamentaceus from Puerto Rico. Chaetopterus species share a body plan specialized for mucus-net suspension feeding, and use large modified fan-like appendages to pump water over a single mucus net.

Above right: Mesochaetopterus taylori from Friday Harbor, WA. Mesochaetopterus use thickened posterior segments to pump water over up to three mucus-nets. They can also use their relatively long palps for deposit or suspension feeding.

Bottom left: Phyllochaetopterus arabicus from the Red Sea. Phyllochaetopterus usually have two mucus nets, and use ciliary rings in leaf-like appendages rather than muscles to pump water for feeding. They generally have very long palps for feeding on deposited or suspended particles.

Bottom right: Spiochaetopterus pottsi from Friday Harbor, WA. Spiochaetopterus species sometimes have many serial mucus nets with ciliary water pumping and long palps for feeding.
This work contributes to the understanding of body plan evolution in annelid worms, a major phylum of animals with many diverse body plans and appendage types. Annelids are an ancient group, first appearing in the Cambrian period, roughly 500 million years ago. Recent phylogenetic evidence suggests that chaetopterids are one of the earliest diverging lineages within Annelida. Chaetopterids are used as model organisms in developmental biology and some may have applications in medicine. A well-resolved phylogeny will help resolve the taxonomic issues in the group and aid in the identification of species important to other disciplines. Understanding how body plans have diversified in chaetopterids, and relating it to present distributions and habitat types provides insight into evolutionary patterns and the factors driving differences in body plans and feeding on a global scale. The phylogeny will be valuable for other areas of science, including detailed study of the biomechanics of feeding mode in the family, much needed systematic revisions, and will identify and clarify the evolutionary relationships of Chaetopterus "sp.", a species used as a model organism in developmental biology. This project also provides a test of applying target capture methods to animals with a long divergence history and with a distantly related reference genome.
Targeted sequencing and phylogenomics of the critically imperiled *Pleuroceridae*

**Principal Investigator:** Ellen Strong (Nathan V. Whelan, co-PI)

**Projected Timeline:** October 2016 start date, 12 month duration

**Budget detail:** Budget costs for this project include domestic travel, contractual services (transcriptome sequencing by Macrogen and NGS MYbaits target capture by Mycroarray), materials and supplies (miscellaneous consumable supplies and Qiagen Plant DNeasy extraction kit and associated consumables), and shipping.

**Project Synopsis:**
The southeastern United States is a hotspot of freshwater mollusk diversity, but many groups are understudied, including 166 species in seven genera in the family Pleuroceridae (Cerithioidea). Comprising the second most diverse family of North American freshwater gastropods, pleurocerids can make up over 90% of macroinvertebrate biomass in some streams, but 79% of pleurocerid species (i.e. ~128) are imperiled and at least 33 species and 1 genus have already gone extinct. Traditional Sanger-based barcoding approaches have been unsuccessful in elucidating the diversity and relationships of pleurocerids owing to unusually high rates of mitochondrial sequence divergence. Consequently, the family has been largely ignored from a molecular systematics point of view and has received limited attention since the advent of high-throughput sequencing. To date, no published study has generated whole genome or transcriptome data for pleurocerids, and there are only two published cerithioidean transcriptomes. The few nuclear genes in routine use for gastropod systematics provide poor resolution at the species level, and the data presently available (Strong et al., in prep) indicate that pleurocerid systematics requires extensive revision. Genomic data are urgently needed to robustly resolve pleurocerid systematics and phylogeny, which will be used to elucidate broad macroevolutionary and biogeographic patterns, place their highly variable shells into a solid phylogenetic framework, explore the evolution of their diverse life history strategies, and enhance management plans. We propose to extend target-capture and high-throughput sequencing approaches to a novel taxonomic group, the Pleuroceridae, to infer a well-resolved phylogeny for this understudied and imperiled group.

**2017 Update:**
We have sampled pleurocerid snails from across the family's geographic range, including six federally listed species. To date, over 350 tissue clips for the GGI Biorepository have been made as a part of this project. Four pleurocerid transcriptomes have also been sequenced. Transcriptome sequencing for designing target capture probes was successful, with 98+% of total base pairs having quality scores above 30. The number of assembled contigs and N50 values are consistent with well-assembled transcriptomes that are appropriate for designing target-capture probes (see Table 1). However, a number of unexpected delays in our project have occurred since the last progress report. Most notably, this is my collaborator’s first year in his new position. In
addition to the burden of taking on new responsibilities, he has inevitably run across some unanticipated issues working in his new lab for the first time. We have also bumped up against vacation schedules with the techs at the lab where the last phase of the work is being done. This has led to unforeseen delays. In addition, we had planned to contract with Mycroarray for designing target capture probes and sequencing, but despite what we were told during discussions with Mycroarray when putting together our original GGI funding proposal, Mycroarray does not offer probe design services. After discussion with colleagues, we decided to contract with RAPID genomics for probe design and sequencing. We have the PO in place and we expect to have target-capture Illumina sequencing data back in 8-12 weeks, and we will immediately start bioinformatics and phylogenetic analyses. In the meantime, we are working on data curation (i.e., cataloging shell vouchers and tissues) and characterizing the four transcriptomes that we sequenced. We intend to publish a paper in Molecular Ecology Resources about the transcriptomes before the end of 2017. Given delays with generating target capture data, we anticipate starting to write a Pleuroceridae phylogenomics paper before the end of 2017 with a submission target of quarter 1 of 2018.

Table 1

<table>
<thead>
<tr>
<th>Species</th>
<th>Total base pairs (raw data)</th>
<th>Total reads (raw data)</th>
<th>Bases with quality ≥ 30 (%)</th>
<th>Total assembled contigs</th>
<th>Total &quot;Genes&quot;*</th>
<th>Assembly N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC6</td>
<td>7,745,652,024</td>
<td>76,689,624</td>
<td>98</td>
<td>230971</td>
<td>194447</td>
<td>1053</td>
</tr>
<tr>
<td>PPR1</td>
<td>8,182,882,034</td>
<td>81,018,634</td>
<td>99</td>
<td>235855</td>
<td>192358</td>
<td>939</td>
</tr>
<tr>
<td>LGR1</td>
<td>8,493,835,380</td>
<td>84,097,380</td>
<td>99</td>
<td>250189</td>
<td>207278</td>
<td>1003</td>
</tr>
<tr>
<td>EAC2</td>
<td>8,346,924,416</td>
<td>82,642,816</td>
<td>99</td>
<td>237346</td>
<td>199341</td>
<td>980</td>
</tr>
</tbody>
</table>

*Genes as inferred by Trinity

Blog: Targeted sequencing and phylogenomics of the critically imperiled Pleuroceridae

Ellen Strong, Research Zoologist, Department of Invertebrate Zoology, NMNH
Nathan Whelan, Director of US Fish Wildlife’s Southeast Conservation Genetics lab

We are studying evolutionary relationships within the freshwater snail family Pleuroceridae. The family comprises 166 species as presently understood distributed east of the Rocky Mountains, with a center of diversity in the southeastern United States. Pleurocerids are essential components of freshwater ecosystems but members of the family are highly imperiled owing to human-mediated modifications to their habitats including pollution and damming of rivers. Over 30 species have already gone extinct and the family is considered to constitute among the most highly imperiled freshwater gastropods in North America, with many at significant risk of extinction. However, the biggest barrier to effective conservation is a lack of knowledge about their diversity.
The primary objective of our project is to construct the framework necessary to assess the diversity of the remaining pleurocerids and determine the relationships among them. This phylogenetic framework is necessary to revise the systematics and facilitate effective conservation management. The classification currently in use is based primarily on features of the shell which can be quite variable within and between species of freshwater snails and are now regarded as unreliable indicators of evolutionary affinity. Efforts to explore diversity and evolution of the group using DNA “barcodes” of mitochondrial genes have been unsuccessful owing to exceptionally high levels of mitochondrial divergence within species. Consequently, the family has not been rigorously assessed using molecular data.

To achieve our goal, we have sampled pleurocerids from across the eastern United States in many river systems including those of peninsular Florida, the Atlantic slope system, the Mobile River basin, the Tennessee River basin, and the lower Mississippi River system. A significant outcome of this project has been the collection and archival of over 400 genomic-grade tissues in the NMNH biorepository, including of very rare...
and endangered species. This accomplishes two major goals of GGI: to collect earth's genomic diversity and increase the capacity to sequence genomes.

Thus far we have sequenced four transcriptomes of pleurocerids. When published, these will be the first publicly available transcriptomes for the family and will double the available transcriptomes for the superfamily to which they belong. These data are being used to produce a probe set that will serve as the basis for high-throughput sequencing using target capture. The genetic data generated in this study will allow us to robustly examine relationships within the family and address questions concerning species-level biodiversity for the first time. Genetic analyses being undertaken as a part of this project also contribute to the GGI goal of sequencing earth's genomic diversity. The results of our study are expected to inform a petition, currently under consideration, for listing of 26 pleurocerids under the Endangered Species Act and will improve conservation efforts which we hope will help prevent future pleurocerid extinctions. We also hope that the wealth of genomic data generated during this research will stimulate further work and elevate interest in this fascinating group of snails.
Evolutionary History of Palpimanoid Spiders Using Genomic Data

Principal Investigator: Hannah Wood (Nikolaj Scharff, Vanessa González, co-PIs)

Projected Timeline: October 2016 start date, 12 month duration

Budget detail: Budget costs for this project include DNA extraction and quality assessment, MYbaits predesigned kit, MYbaits custom kit with designed probes, extra MYbaits reagents for dilution of kits, library prep for Illumina sequencing, Illumina library quality assessment, Illumina HiSeq (96 samples per lane), and contractor costs for two weeks.

Project Synopsis:
Palpimanoids, although not very speciose, are nevertheless, an extremely important lineage in the evolutionary history of spiders due to their phylogenetic placement: recent research places palpimanoids as sister to the Entelegynae (Wood et al. 2012, Dimitrov et al. 2016), a clade that contains the bulk of spider diversity, containing ca. 39,500 species or 86% of all spider species. The fossil record for palpimanoids is well documented and suggests that these spiders were at one time more widespread and more diverse in the past, having families and genera that have since gone extinct (Dunlop et al. 2016). Furthermore, palpimanoids have evolved spectacular morphological modifications that are associated with unusual predatory behaviors, such as novel structural mechanisms that allow for high-speed, power-amplified predatory strikes (Wood et al. 2016), and grotesque morphologies that allow for attack-at-a-distance strategies in clades that are specialized to prey on other spiders (Wood et al. 2012, Wood et al. 2015). Extant palpimanoids for the most part are restricted to the Southern Hemisphere, although there is a documented fossil record from the Northern Hemisphere, with fossils going back to the Jurassic (165 Ma), and with distribution patterns in some groups related to Pangaean vicariance (Wood et al. 2013).

2017 Update:
Palpimanoid spiders have evolved spectacular morphological modifications that are associated with unusual predatory behaviors, such as novel structural mechanisms that allow for high-speed, power-amplified predatory strikes, and grotesque morphologies that allow for attack-at-a-distance strategies in clades that are specialized to prey on other spiders. To date, evolutionary relationships among palpimanoid members have been examined using morphology and Sanger sequencing of a limited set of markers, with relationships among families being unsupported and not well resolved. However, Next-Generation Sequencing (NGS) techniques may resolve these issues, specifically the technique of Target Enrichment, which makes use of short fragment sizes typical of degraded DNA. Using a probe set based on published spider transcriptomic data, and also using a probe set based on arachnid Ultra-Conserved-Elements (UCE) we attempted to resolve deep relationships among palpimanoid families and genera. We generated baits to targeted 1500 loci based on transcriptome data and targeted 1000 loci for the previously generated Ultra-conserved elements.
This project is progressing very smoothly. We successfully sequenced hundreds of regions scattered throughout the genome for 48 spider species and have produced a robust phylogeny. We are currently writing up these results to submit and will likely submit this manuscript in one month. Then, in October/November we will do another round of sequencing of an additional 200 arachnid species. This work will result in at least one more publication that we will write up and submit for publication, likely in February or March of 2018. We will have spent all of our awarded funds by Dec. 30th.

Blog: Evolutionary History of Palpimanoid Spiders Using Genomic Data

My research focuses on the evolution of unusual mouthparts in spiders. My name is Hannah Wood and I am the Curator of Arachnids at the National Museum of Natural History and I have been studying palpimanoid spiders for over 13 years. Palpimanoids are an extremely important lineage in the evolutionary history of spiders. The fossil record for palpimanoids is well documented and suggests that these spiders were at one time more widespread and more diverse in the past, having families and genera that have since gone extinct. Furthermore, palpimanoids have evolved spectacular morphological modifications of the chelicerae that are associated with unusual predatory behaviors, such as novel structural mechanisms that allow for high-speed, power-amplified predatory strikes, and grotesque cheliceral morphologies that allow for attack-at-a-distance strategies in clades that are specialized to prey on other spiders. The chelicerae are structures at the front of the spider’s body that have fangs that inject venom and are used for prey capture. The chelicerae can be thought of as being functionally equivalent to jaws or mandibles.

Because of these traits palpimanoids offer the possibility to test hypotheses regarding changes in cheliceral morphology, diversity, and distribution across large time scales. However, an understanding of evolutionary relationships is crucial for addressing these questions. Yet to date, evolutionary relationships among palpimanoid members have only been examined using traditional molecular sequencing techniques to gather data from a handful of genes. These traditional techniques have failed at resolving some relationships among palpimanoid members.

I currently have funding from the Global Genome Initiative to make use of recent advances in genetic sequencing to resolve relationships among palpimanoid members. The goal of this project is to use novel Next-Generation Sequencing techniques to resolve relationships among palpimanoid
families, with a specific focus on species from the biodiversity hotspot of the Eastern Arc mountains of Tanzania. This research will be done using specimens that I've collected over the last 10 years on international fieldtrips mostly to the Southern Hemisphere. We are also using specimens that were collected by a collaborator doing research on arachnids in Tanzania.

An image of myself collecting spiders in the Philippines. To catch ground spiders I have sifted some litter on the ground and then spread it out onto a sheet. When the spiders start to run I use an aspirator to capture them. Yes, the aspirator has a filter so the spiders never make it into my mouth! Photo: Stephanie Stone

Left: A palpimanid spider from South Africa, Palpimanus sp., with her eggcase. Right: This picture was taken at Montagne d'Ambre, a national park in the far north of Madagascar. These rainforest harbor many new species that have yet to be documented by science. Photos: Hannah Wood.
As part of the GGI project we have been sequencing hundreds of regions scattered throughout the genome for different palpimanoid species. Using these data we have constructed a phylogeny, or a branching diagram of evolutionary relationships, that allows us to address specific questions involving cheliceral diversification and biogeography patterns.

This type of research is very useful to arachnologists: traditional sequencing methods of arachnids has not provided enough data to examine deep evolutionary relationships. This work documents the biodiversity of a group of rare spiders that mostly occur in remote parts of the world, furthering GGI’s goals of sampling the world’s biodiversity and generating genomic sequences as a tool to study this biodiversity.
Exploratory phylogenomics and differential expression of genes involved in foliar chemical defense in the hyperdiverse tropical tree genus *Psychotria*.

Principal Investigator: Brian E. Sedio (Rebecca B. Dikow, Paul B. Frandsen, Owen McMillan, S. Joseph Wright, co-PIs)

Projected Timeline: June 2016 start date, 10 month duration

Budget detail: Budget costs for this project include supplies, Life Technologies PureLink RNA kit, Life Technologies Turbo DNA-free kit, Illumina TruSeq Stranded mRNA library prep kit (2), Life Technologies Superscript II Reverse Transcriptase, Agilent Bioanalyser RNA nano, Agilent High Sensitivity DNA kit, plastics, Illumina HiSeq 4000 150 bp paired-end DNA sequencing (2 lanes), Illumina HiSeq 4000 50 bp paired-end DNA sequencing (4 lanes), and Capture design and optimization (50 individuals).

Project Synopsis:
We will characterize the foliar transcriptomes of *Psychotria* trees and use these data to (1) develop a capture probe set for broad-scale phylogenomic analysis and (2) identify candidate loci involved in chemical evolution. *Psychotria* is the fourth largest plant genus, with some 1,650 species distributed throughout the tropics and subtropics. The genus is comprised almost exclusively of shrubs and small trees of the forest understory. *Psychotria* inspired the botanist Alwyn Gentry to coin the term ‘species-swarm’ to describe it and other genera comprised of seemingly ecologically undifferentiated species with high local species richness.

2017 Update: Our project seeks to characterize foliar transcriptomes of *Psychotria* trees and use these data to (1) develop a capture probe set for broad-scale phylogenomic analysis and (2) identify candidate loci involved in chemical evolution. *Psychotria* is the fourth-largest plant genus in the world. As its name implies, this diversity is thought to be the result of an adaptive radiation driven by the evolution of defensive secondary metabolites, particularly several classes of alkaloids of which some exhibit known emetic, anti-helminthic, hallucinogenic, and psychotropic properties in humans. With support from the Smithsonian Institution Scholarly Studies Program and Grand Challenges Consortia, we have applied recent innovations in mass spectrometry metabolomics to the study of thousands of secondary metabolites in *Psychotria* and other species-rich tropical tree radiations, as well as in whole forest communities in both Panama and Maryland. Our results indicate that temperate and tropical forests are fundamentally different chemically. However, not only are tropical trees in Panama more diverse and less similar chemically than temperate trees in Maryland, the better part of the chemical diversity of Panama trees is comprised of just a few species-rich and chemically diverse genera, including *Psychotria*. Within *Psychotria* and other potential chemical-defense adaptive radiations, even the most closely related species can be remarkably different chemically. Our recent and forthcoming ecological publications strongly suggest that species differences in secondary metabolites comprise a major axis of niche differentiation in tropical forests, and contribute to the ecological
coexistence of high levels of species diversity by imposing narrow host ranges on insect herbivores and microbial pathogens that check the population growth and density of their host plants. However, to link these ecological and diversity patterns to chemical evolution requires: i) well-resolved phylogenies for *Psychotria* and other challenging genera, and ii) an understanding of the genetic underpinnings of the striking chemical differences among closely related species. To this end, we are generating cDNA libraries to identify homologous loci for phylogenomics and differentially expressed loci to begin to identify the genes responsible for the metabolomic differences among closely related species of *Psychotria*. We plan to use these exploratory investigations of *Psychotria* to build infrastructure for large-scale phylogenomics and comparative work in *Psychotria* and to extend these methods to several other species-rich and chemically diverse tropical tree radiations to identify features that consistently characterize lineages that may represent adaptive radiation in secondary metabolites defenses relative to their less diverse sister lineages.

We have successfully generated quality 38 cDNA libraries representing 12 species comprising diverse evolutionary lineages within the 35.6 million year-old genus. These libraries represent 4 stages of leaf ontogeny for each of the 12 species, so as to increase the number of potential homologous, phylogenetically informative loci recovered from the transcriptomes. Furthermore, we have generated additional libraries for six target species that represent two groups of closely related species with dramatic differences in their foliar metabolomes. We have procured sequencing costs for two HiSeq 4000 lanes. Assuming that each lane produces 700 million paired-end reads, we expect to generate approximately 100-120 million, 2x150 Illumina pair-end reads per species or roughly 25 million reads per developmental time point. We expect to receive these data in September, complete bioinformatics work in November, and submit the first manuscript reporting loci for phylogenomics in December 2017. We expect to submit our first manuscript concerning comparative gene expression in *Psychotria* in early 2018.

**Blog: Tropical Forests: Pharmacopeias of Chemical Diversity**
*By Brian E. Sedio*

Tropical forests are remarkably diverse. A quarter of a square mile of forest in Ecuador can contain over 1,000 tree species, or roughly as many as the 1.6 million square miles of temperate forests in North America, Europe, and Asia combined. Understanding how so many species of tree coexist in tropical forests remains one of the greatest challenges in ecology. It is thought that species must differ in some important way in order to coexist, such that each species exploits a distinct “niche” and avoids competing with other species for common resources. Animals, for example, can exploit distinct food resources. Where pumas and jaguars overlap, the lighter pumas tend to focus on fleet-footed deer, whereas the more powerful jaguars focus on larger or tougher game like tapirs and peccaries. Plants, however, all require a small number of common resources, such as light, water, carbon dioxide, and a small number of nutrients. Given these similarities, are there really 1,000 ways to be a tree in a small plot of forest in Ecuador?
I am Brian Sedio, and I am a postdoctoral researcher at the Smithsonian Tropical Research Institute (STRI) in Panama. I have a Bachelor’s Degree in Biochemistry and Genetics and a Ph.D. in Ecology and Evolutionary Biology. My work seeks to identify the characteristics, the niches, that allow tree species to coexist in tropical forests, and I think they have less to do with soil and water than with the astonishing diversity of chemical compounds that distinguish the trees of the rainforest. To test these ideas, the Smithsonian Institution Global Genome Initiative (GGI) has supported my recent study of the relationships among species of plants in the genus *Psychotria* (in the coffee family, Rubiaceae) in Panama, and of the genes responsible for the astonishing chemical diversity found in these plants.

Why are there so many species of tree in the rainforest? A not altogether unrelated problem in ecology is that known as the Green World problem: Why is the world green? Plants are immobile organisms, sitting ducks for hungry herbivores on the prowl. Why don’t the many beetles, caterpillars, monkeys, and deer simply increase in abundance until the world around them is denuded of vegetation? The renowned ecologist Dan Janzen proposed this hypothesis: “To an herbivore, the world isn’t green. It’s colored L-dopa, cocaine, and caffeine.”

That is to say, to the human eye, a tropical forest appears as a sea of green. Yet the forest is likely able to sustain such an impressive standing biomass of vegetation precisely because of its diversity, and that diversity manifests not only in the number of tree species, but in the astonishing diversity of chemical compounds with which plants defend themselves from an endless onslaught of insect herbivores and microbial pathogens.

The chemical diversity of plants has long stymied biologists studying ecological communities or broad-scale evolution. Any individual plant can be characterized by hundreds of small molecule metabolites. A forest can contain hundreds of thousands of unique molecules. Conversely, any given compound can be found in a small number of very distantly related plants, but few species in a given community. Caffeine, for example, is found in coffee (*Coffea arabica* and *Coffea canephora*, native to Ethiopia), tea (*Camellia sinensis*, native to Asia), yerba mate (*Ilex paraguaiensis*, native to Argentina and
southern Brazil), kola nut (*Cola acuminata*, native to West Africa) and chocolate (*Theobroma cacao*, native to Amazonia). None of these species occur together naturally, hence caffeine may be wholly unique to a cacao within an Amazonian tree community. But just how different chemically is *Theobroma cacao* from its neighbors in the eyes of a beetle? Enough to turn away a hungry bug?

Fortunately, our ability to study the chemical differences among plant species was given a significant boost by innovations in mass spectrometry by analytical chemist Pieter Dorrestein and collaborators at the University of California San Diego. Dorrestein, Mingxun Wang, Nuno Bandeira, and others developed a method for comparing the structures of unknown molecules, and using those similarities to build “molecular networks”, the links of which indicate structural similarities among molecules. Their method works because two molecules that are similar in structure will, when broken, break into many of the same pieces. Hence, one can compare the pattern of pieces generated when different molecules are shattered as a means of measuring their similarity, even without knowing the true structure of the molecules in question. This gives ecologists and evolutionary biologists a tool for assessing the chemical similarity of species and for identifying potentially important compounds even in chemically diverse, species-rich, and understudied plant communities such as tropical rainforests.

I have begun collecting mass spectrometry data for hundreds of tree species in both the United States and Panama to attempt to identify the chemical niches that may sustain the coexistence of species and ultimately favor the evolution of new, chemically novel compounds.
species through the process of natural selection. In addition to the chemical data, we collected 2,000 insect herbivores in Maryland and 8,000 in Panama and are sequencing “DNA barcodes” of the insects themselves and the plant tissue in their stomachs. DNA barcodes are short, highly variable DNA regions that can be used to distinguish and identify species, even from otherwise unidentifiable tissue, such as the plant tissue in the stomach of a beetle. If we know which insects eat which plants, and we know which compounds are found in each plant, we can let the insects tell us which compounds determine their choice of host plant. In essence, we can measure the chemical niches of the plants.

In 1964, Paul Ehrlich and Peter Raven imagined that the ecological importance of chemical differences among plants might be so powerful as to drive the evolution of new species of plants and the insects that feed on them, and ultimately generate much of the world’s biodiversity. With regards to plants, Ehrlich and Raven envisioned a mutation resulting in a novel defense, followed by ecological success as the plant population grows unchecked by herbivores. This microevolution of a novel defense in a population could be followed by speciation, as the new, incipient species benefitted from distinguishing itself from less chemically novel populations. Eventually, insects and microbes would themselves adapt to the new defense and colonize the new plant species. Over time, this evolutionary arms race would generate many species of both plants and herbivores. Most interestingly, if some chemical or genetic traits were more evolvable than others, the rate of speciation would accelerate in plant lineages with highly evolvable chemical defenses. Such chemical defenses, or classes of chemical defenses, would therefore be key innovations that distinguish a diverse adaptive radiation of related species from a sister lineage with less evolutionarily labile chemistry.

The late botanist Alwyn Gentry noticed that a small number of exceptionally species-rich tree genera contribute a large fraction of the species in many tropical forests. Gentry called genera such as *Eugenia* (the myrtle family, Myrtaceae), *Inga* (the bean family, Fabaceae-Mimosoideae), *Miconia* (the melastome family, Melasomataceae), *Piper* (the black pepper family, Piperaceae) *Pouteria* (the sapote family, Sapotaceae), and *Psychotria* (the coffee family, Rubiaceae) “species swarms” because of the astonishing...
A *Psychotria limonensis* suffers from intensive herbivore damage, whereas its *Psychotria acuminata* neighbor does not. Insect herbivores can have large effects on the plants they feed on. Differences in chemical defenses among plant species can allow one species to avoid the herbivores of another, even if they are closely related, as in these *Psychotria*.

What does a chemical niche look like? Perhaps plant seedlings tend to survive where they differ chemically from their neighbors (left), and to die where they are chemically similar to their neighbors and hence likely to share herbivores and pathogens (right). Following these patterns may allow us to infer which chemical traits define the niches of coexisting species. Figure from Sedio 2017 *New Phytologist* 214:952-958.

number of ecologically similar, closely related species that can be encountered within a fraction of an acre. These genera can comprise between a quarter and a third of the tree species in forests from Mexico to Brazil. My recent work indicates that they comprise an even greater part of the chemical diversity of these forests, and I and others have found that closely related species in these genera tend to be very different chemically. Are Gentry’s species swarms adaptive radiation in chemical defense?
One of the most interesting of Gentry’s species swarms is *Psychotria*, so named because some species contain psychotropic, or hallucinogenic, compounds. *Psychotria* is perhaps the fourth most species-rich plant genus on the planet, with over 2,000 species found in subtropical and tropical forest environments in the Americas, Africa, southern Asia, and even Pacific islands such as Hawaii. *Psychotria* is also diverse at local spatial scales, with perhaps 160 species in Panama, or to give another example, over 60 species in a quarter-square-mile forest plot in Ecuador. Much of this diversity may be a result of the chemical differentiation of species, as Ehrlich and Raven predicted.

*Psychotria* are known for their alkaloids, small organic molecules that contain nitrogen, often in a ring structure. Alkaloids can be very biologically active. The caffeine that makes coffee, tea, and chocolate stimulating is an alkaloid. So are the nicotine in tobacco (*Nicotiana tabacum*) and the narcotic cocaine (derived from several species of *Erythroxylum*). The most well-known alkaloids found in *Psychotria* are emetine and cephaeline, found in a Central American species known as *ipecacuanha*, or more commonly as ipecac. Syrup of ipecac was used historically to induce vomiting if poison had been ingested (physicians no longer recommend this practice prior to diagnosis). These alkaloids also exhibit antiprotozoal activity.

Another known alkaloid of *Psychotria* is dimethyltryptamine, or DMT, derived from the Amazonian species *Psychotria viridis*, or chacruna. DMT is found naturally in many plants and animals, including humans, though *Psychotria viridis* produces the alkaloid in
high concentrations. DMT is structurally similar to serotonin, a neurotransmitter associated with feelings of euphoria.

Interestingly, if one were to consume a leaf of chacruna, nothing would happen, as an enzyme that occurs in human saliva would metabolize the DMT prior to passage to the small intestine. However, indigenous peoples of the western Amazon long ago discovered that the roots or bark of an unrelated vine, *Banisteriopsis caapi* or ayahuasca, can be used to liberate the effects of DMT on the body. This is because ayahuasca contains a monoamine oxidase inhibitor (MOI) that prevents the enzyme in one’s saliva from altering DMT. In traditional societies in western Amazonia, ayahuasca root or bark is mixed with chacruna leaves to yield a psychotropic tea. I must confess that my interest in these practices is more intellectual than experiential, as the first effects of ayahuasca consumption include several hours of projectile vomiting and diarrhea. Afterwards, the participant typically experiences a vision in which they are visited by a powerful animal, such as a jaguar, which imparts wisdom. Or so I am told. It’s hard for me to imagine any wisdom a jaguar could give me that would be worth several hours of vomiting.

Let me point out here that it is easy to view the practice of consuming ayahuasca through the lens of Western, recreational drug use. However, both *Psychotria viridis* and *Banisteriopsis caapi* contain compounds that function as antihelmintic drugs. That is, the consumption of ayahuasca is a treatment for parasitic worms, which I imagine is a widespread malady in traditional Amazonian societies. It turns out the jaguar spirit vision may just be a pleasant side-effect that makes an otherwise unpleasant medical treatment more bearable.

Emetine and DMT are only two of thousands of compounds produce by species of *Psychotria* with no clear function in primary metabolism. Many of these alkaloids, terpenoids, flavonoids, and other secondary metabolites may serve to defend *Psychotria* from insects and pathogens. And the vast diversity of *Psychotria* species may be a result of the ease which with *Psychotria* evolve new chemical compounds, allowing them to temporarily escape pests and pathogens and generate new species of *Psychotria* in the process. In fact, in Panama, the most closely related species of
Psychotria are often remarkably different chemically. This suggests that natural selection imposed by herbivores results in chemical divergence among closely related species, and may result in speciation and hence the radiation of the genus.

To understand how closely related species of Psychotria can differ so dramatically in their chemistry, we need to understand the relationships among species of Psychotria and to identify the genes responsible for the alkaloids and other metabolites. To that end, the Smithsonian Institution Global Genome Initiative (GGI) supported our collecting expedition and genetics research on the Psychotria of Barro Colorado Island, Panama.

Interns Luke Frentsos, from Florida, and Christian López, from Panama, spent months locating individual plants in the forest understory on Barro Colorado Island comprising 21 species of Psychotria. Luke and Christian marked individual leaves and tracked their development, harvesting leaves after 5, 7, 10, and 14 days. These leaves were collected in liquid nitrogen in the forest to preserve the RNA transcripts active in the cells, temporary strips of RNA that encode the proteins and enzymes responsible for turning primary metabolites, such as the sugar glucose and the amino acid tryptophan, into secondary metabolite defense molecules such as emetine and DMT.

Back in the lab at the Smithsonian Tropical Research Institute (STRI), Christian isolated RNA from frozen leaf tissue. Christian and STRI research technician Marta Vargas then reverse-transcribed the RNA into the original DNA code in which the RNA sequences are stored in the cells chromosomes. We then sequenced the DNA to provide us with DNA sequence for hundreds of genes that are actively expressed in developing leaf tissue. We will now use this DNA sequence to identify genes that are shared among all Psychotria, the variation in which we can use to determine the relationships among hundreds or thousands of plant species in the genus. We will also identify genes that
differ, either in their DNA sequence or their expression levels, between very closely related species of *Psychotria*. This will allow us to identify candidate genes underpinning the defensive chemistry that may define the most important ecological niches of *Psychotria* species and to understand why *Psychotria* is so diverse, both in terms of species and chemistry.

For now, with support from GGI, we seek to understand the chemistry, ecology, relationships, and genetic architecture of *Psychotria*. Yet the tools we are developing through this research will make it easier in the future to apply these same methods to other potential adaptive radiations in defensive chemistry—Gentry's species swarms—to understand what makes tropical forests so diverse, and so green. I have a strong hunch that Janzen, Ehrlich, and Raven were right in their predictions that chemistry drives much of plant ecology and evolution.

A vision of the tropical rainforest as a pharmacopeia of chemical diversity: "Los Cachiboleros" by Pablo Cesar Amaringo Shuña, a Peruvian shaman who depicts ayahuasca-induced visions in art. https://s-media-cache-ak0.pinimg.com/736x/a8/78/e3/a878e3f0313fed2595d1d25aec6d03ae.jpg
Testing Historical Presence of Mountain Gorilla and Elephants in an African Biodiversity Hotspot using Targeted Enrichment of Sedimentary Ancient DNA.

Principal Investigator: René Dommain (Rick Potts, Jesus Maldonaldo, Michael Campana, Molly McDonough, co-PIs)

Projected Timeline: October 2016 start date, 12 month duration

Budget detail: Budget costs for this project include permits, travel and accommodations, fieldwork core equipment, core scanning and storage, radiocarbon dating, aDNA Analysis 25 samples, DNA extraction, library prep, quality checking, and Illumina runs and MYbaits.

Project Synopsis:
Climatic fluctuations during the Quaternary in Africa resulted in repeated fragmentation of forests and therefore expansion and contraction of species distributions. Today's most diverse African forests are thought to represent Pleistocene refugia where forest-adapted species survived during dry climate phases. Whereas most proposed African rainforest refugia were located in the central African Congo Basin, few refugia existed in East Africa. This supposedly includes the most biodiverse East African forest—the Bwindi Impenetrable Forest in Uganda. This forest harbors one of the two remaining populations of mountain gorillas and is a contact zone for savannah and forest elephants—both highly charismatic but highly endangered species. Pollen data seem to indicate that the Bwindi forest was widely replaced by open scrub communities during the dry Last Glacial Maximum (26,000-19,000 years ago; Marchant et al. 1997), leading Tocheri et al. (2016) to suggest that mountain gorillas survived this period only in the Virunga region and dispersed into Bwindi about 10,000 years ago. Whether Bwindi served as a Pleistocene refugium remains to be empirically tested.

2017 Update:
The goal of this project is to explore the suitability of the novel technique sedimentary ancient DNA (sedaDNA) analysis in the reconstruction of past species occurrence in tropical rainforests. In particular, we plan to apply the sedaDNA technique on thousands of year old sediments from a swamp in Bwindi Impenetrable Forest in Uganda to reconstruct for how long mountain gorillas (Gorilla beringei beringei) and African elephants (Loxodonta africana) are present in this unique tropical forest. The Bwindi forest is among the most diverse regions in Africa and harbors one of the two remaining populations of mountain gorillas. Its high diversity suggests that this forest served as a refuge for forest organisms during previous periods of arid climate, but this idea has not been tested empirically. Any DNA preserved in sediments can be extracted with the sedaDNA technique and targeting the DNA of specific taxa can potentially reveal their local presence in the past. As a first step in this project we designed a custom MYBaits (Mycroarray, Inc.) hybridization capture assay to capture ancient DNAs pertaining to mountain gorillas and African elephants that may be preserved in soils. To test the efficacy of the MYBaits capture assay and to prove our approach before sampling old
sediments in Bwindi, we collected 23 soil samples from the Smithsonian National Zoological Park elephant and gorilla enclosures. We included soils from a range of conditions including areas in which the animals were known to defecate, areas in which they traversed and also areas in which they were rarely seen. As positive controls, we included feces from both species collected in the enclosures. We built these into Illumina sequencing libraries and captured them using the MYBaits kit. All soil samples contained DNA and as expected organic-clayey soils had higher concentrations than sandy soils while deeper soil layers generally contained less DNA than near-surface samples. We anticipate to have the sequencing data analyzed by the end of August 2017 and a manuscript ready for submission by December 2017. The field sampling in the Bwindi forest will take place in early August 2017 and the extraction and sequencing of sedimentary ancient DNA together with radiocarbon dating of the sediments is planned for fall 2017. We will start drafting a manuscript on the African samples in December 2017, which should be ready for submission in the following year.

Blog: Gorillas, Elephants and a tropical rainforest in Uganda – in search of their history with sedimentary ancient DNA

By: R. Dommain and M. Campana

It has long been assumed that the exceptional biodiversity of the tropical rainforests is the result of stable climatic conditions that persisted for millions of years. However, there is growing evidence that also the tropics experienced severe climatic fluctuations during the past that included long periods of much drier climate when rainforests in Africa and elsewhere were restricted in their extent. Where did forest animals survive during these unfavorable periods? Did the fragmentation of large forest blocks contribute to diversification? Unfortunately, fossil remains of animals are rarely being found in tropical forests leaving us with an incomplete picture of the past distribution of tropical rainforest species. A new technique - the analysis of so-called sedimentary ancient DNA (sedaDNA) - may soon change the game. By extracting and sequencing old DNA that is preserved in lake or swamp sediments the local presence of certain species during the past could be proven.

The Smithsonian scientists René Dommain, Michael Campana, Molly McDonough, Jesus Maldonado, and Rick Potts received the first GGI grant to study sedimentary...
ancient DNA in an exploratory project that aims at reconstructing the history of mountain gorillas and African elephants in one of the most diverse rainforests of Africa. In particular, we would like to test whether the Bwindi forest in Uganda served as a refuge for these species during previous dry periods. The results could potentially provide better insights into the survival capabilities of these now highly endangered species under future climatic changes.

In a first step we had to design a so-called RNA probe set pertaining to mountain gorillas and African elephants that will allow us to capture the ancient DNAs of these species from the bulk of the extracted sedimentary DNA. Before sampling old sediments in tropical Africa we have to test the efficacy of the probe set, however. For our test study we collected over 20 soil samples from the Smithsonian National Zoological Park elephant and gorilla enclosures with the help of elephant specialist Natalia Prado-Oviedo and the curator of primates at the National Zoo Meredith Bastian. We included soils from a range of conditions including areas in which the animals were known to defecate, areas in which they traversed and also areas in which they were rarely seen. We built these soil samples into Illumina sequencing libraries and captured them using the probe set. All soil samples contained DNA and we will soon have the sequencing data analyzed. Molly, Michael, and René will then travel to Uganda to collect several thousand year old sediments on which they will perform the sedaDNA technique in the ancient DNA lab of the Smithsonian’s Center for Conservation Genomics. The ancient DNA will eventually be permanently archived in the GGI Biorepository of the Smithsonian Institution’s National Museum of Natural History. Stay tuned to hear about our upcoming expedition to Uganda.
The Center for Conservation Genomics is part of the Smithsonian National Zoological Park and Conservation Biology Institute. We use genomic data, particularly non-invasive and ancient DNA data, to inform animal conservation. Michael G. Campana is a computational genomicist and ancient DNA specialist. Jesus Maldonado is a mammologist with a particular focus on the applications of non-invasive sources for DNA data. Molly McDonough is a molecular systematists focusing on ancient DNA techniques and African mammals.

René Dommain is a paleoecologist at the Human Origins Program of the National Museum of Natural History and a Geo.X fellow at the University of Potsdam, Germany with a focus on studying ecosystem dynamics in the Old World tropics. Rick Potts is a paleoanthropologist and the director of the Smithsonian's Human Origins Program with over three decades of experience in excavating fossils in Africa.

Above left: The collected soil samples await DNA extraction. Soil samples ranged from organic rich clays to pure sands. All contained DNA. Above right: A soil core collected in the gorilla enclosure. Bottom: Meredith Bastian, Natalia Prado-Oviedo, Michael Campana, and Jesus Maldonado after successful sampling of the gorilla enclosure (in the background).
The Caddisfly Genome: Psychomyia flavida

Principal Investigator: Paul B. Frandsen (Vanessa González, Rebecca Dikow, Matt Kweskin, Karl Kjer, Xin Zhou, co-PIs)

Projected Timeline: June 2016 start date, 8 month duration

Budget detail: Budget costs for this project include PacBio SMRT cells (75), library prep (2), Illumina sequencing and library prep (for genome polishing), and Illumina RNAseq run for annotation.

Project Synopsis:
This project will focus on generating a genome for a taxon for which a previous attempt using solely Illumina data resulted in a highly fragmented genome. We will use long read PacBio data to assemble the first high quality reference genome for this important insect order. We will also sequence low error Illumina data for genome polishing. We will use the lessons that we learn during this study to improve our ability to conduct genome assembly and annotation at the Smithsonian and to advise future genome projects by Smithsonian researchers.

2017 Update:
The Trichoptera genome project is in full swing. Though initial efforts at collection did not yield enough DNA for a high-quality genome (for the initial targeted species—Psychomyia flavida), we were recently able to successfully extract high-quality DNA from a larger, fresh caddisfly. Attached is a photo of the pulse field gel electrophoresis run, which shows that we successfully extracted high molecular weight DNA, which is important for PacBio sequencing. We are currently preparing the libraries, which will be used for the sequencing. A transcriptome (all expressed RNA) has already been sequenced, which is going to be vital for genome annotation once we have an assembled genome.

Timeline: Library prep and sequencing will be completed by mid-September. The bioinformatics analysis will be completed by the end of October with a publication describing the genome slated to be submitted by January 2018.

Figure: Pulse field gel electrophoresis (PFGE) run, showing that we successfully extracted high molecular weight DNA. Size standards (wells labelled “Ladder”) are used to estimate size of DNA (labelled “Trichoptera”) based on relative locations on the gel (entire figure). Size standard indicators in the “Ladder” wells are of known size, shown by the yellow bands. DNA present as the white horizontal band, is measured in relative position on the gel to the known size standards.
Blog: The caddisfly genome project

My name is Paul Frandsen and I am a Research Data Scientist in the Data Science Lab housed in the Smithsonian’s Research Computing department. I earned my Ph.D. in entomology, during which my research was focused on the phylogenomics of caddisflies and the development of algorithms to automatically select partitioned phylogenetic models. At the Smithsonian, my work as a Research Data Scientist is focused on generating insights from large digital data sets, including genome sequence data and digitized museum collections. I also maintain my interest in the "little creatures who run the world," especially members of the insect order Trichoptera, commonly known as caddisflies.
Comprising almost 15,000 named species, caddisflies are the most speciose group of strictly freshwater insects. Their larvae inhabit lakes, rivers, streams, and even marine environments (an uncommon feature for an insect). They are differentially sensitive to pollution levels in streams, which allows them to be reliable indicators of freshwater quality and they have been used in biomonitoring efforts around the globe. While the adults resemble small brown moths and can be seen fluttering around aquatic habitats, the larvae are only found under the water and are well-known for their case making and fixed-retreat making behavior. The larvae have glands in their mouthparts that secrete silk, which they use in a variety of ways to build tube cases and intricate underwater homes. Indeed, they are often referred to as “nature’s underwater architects.” Though the caddisflies are fascinating in a variety of ways, little is known about their evolutionary history, including insights concerning their development of silk and how they have adapted to life in aquatic environments. Genome sequencing is one way that we might illuminate the answers to some of these mysteries.

Until now, only a single caddisfly genome has been sequenced and it consists of so many small pieces that it isn’t very useful for any meaningful analysis. The focus of the present study funded by the Global Genome Initiative is to sequence a high-quality trichopteran genome that we might use to further our understanding of the group. While the first trichopteran genome was sequenced using solely Illumina short reads, this project will take advantage of PacBio sequencing technology to produce long reads (on average 100x times longer than the reads used in the first genome), which will result in...
much more contiguous sequences. A more contiguous genome will enable analyses that offer better insights into the natural history of the caddisflies. The genome will become especially powerful when compared against the genomes of its closest relatives, the butterflies and moths (insect order Lepidoptera), which primarily inhabit terrestrial environments.

(Embedded YouTube link (https://www.youtube.com/watch?v=Z3BHzDHoYo), caption: The video was produced by PBS to extoll the virtues of the caddisfly.)

Through the analysis of the newly sequenced genome, we hope to discover genes that play a role in the transition to aquatic environments and to better understand the genetic basis of silk production. Since it will be the first high quality genome sequenced from Trichoptera, it will fill an important piece in the insect evolution puzzle and enable other researchers to also gain insights into broader questions of insect genome evolution.
The Atlantic awning clam genome: *Solemya velum*

**Principal Investigator:** Vanessa L. González (Colleen Cavanaugh, Rebecca B. Dikow, Paul B. Frandsen, Matthew Kweskin, Shelbi Russell, Ellen Strong, co-PIs)

**Projected Timeline:** April 2016 start date, 10 month duration

**Budget detail:** Budget costs for this project include DNA Extraction and QA/QC, PacBio SMRT Cells (50), and library prep (2).

**Project Synopsis:**
This project will focus on improving genome assemblies in groups where deep coverage Illumina sequencing has produced sub-par assemblies. We will pursue long read sequencing (Pacific Biosciences) to aid in genome assembly for this taxon. We will use existing genomic resources for genome polishing (100bp and 150bp Illumina sequence information) and existing transcriptomic resources to aid in annotation. We pursue a variety of assembly strategies, including the utilization of software that can combine the two types of data (hybrid assemblies), which has been shown to improve genome assembly (fewer, longer contigs).

**2017 Update:**
Among the poorest known molluscan groups is the subclass Protobranchia, a bivalve lineage that has diversified and colonized the deepest oceans, with numerous cosmopolitan species at abyssal depths. Of the ca. 750 protobranch species, most are deposit feeders in soft sediments, but two lineages host chemoautotrophic, sulfide-oxidizing bacteria, with concomitant reductions of the hosts’ alimentary system, including *Solemya velum*. All symbiotic bacterial genomes have been sequenced and identified in *Solemya velum* (Shelbi Russell, doctoral thesis). Studies corroborating symbiotic bacteria diversity and adaptation can be greatly aided by genomic resources. By generating a reference genome, information can be provided on fundamental questions about the evolution of shell formation and the organismal responses to changes in ocean chemistry. Existing high quality, alcohol preserved lots for 17 populations of *Solemya velum* across its Atlantic distribution, used in preliminary genomic data generation, have been deposited and cataloged in the Harvard MCZ. These specimens have been requested and have been shipped to NMNH for DNA sequencing. DNA quality and quantity assessments have been performed and are of adequate size to continue with genome assembly and are in the queue for library preparation and long read sequencing on the Pacific Biosciences, newest sequencing platform, the Sequel.
Blog: The Atlantic awning clam genome: *Solemya velum*

My name is Vanessa L. Gonzalez and I am a Computational Genomics Scientist for the Global Genome Initiative (GGI). In my current position, I’m using genomics to understand how species are related across the tree of life. By peering into the genetic makeup of organisms, we can figure out how, where, when, and why biodiversity happens. I work on a breadth of the organisms from plants, spiders, to Narwhals, but have focused the majority of my work on Bivalve mollusks.

Bivalves are clams, oysters, mussels, scallops and other species that have a shell made of two parts. These species are important for many aspects of human life, including fishing and food sources, ecosystem clean-up, and even biomedical applications. They are also a huge part of the fossil record and can tell us a lot about evolutionary processes. Bivalves are an ancient and ubiquitous group of aquatic invertebrates with an estimated 10,000 – 20,000 living species. They are economically significant as a human food source, and ecologically important given their biomass and effects on communities. Their inter-relationships have been studied for decades, and their unparalleled fossil record extends from the Cambrian to the recent.

One of the poorest known bivalve groups is the subclass Protobranchia, a bivalve lineage that has diversified and colonized the deepest oceans, with numerous cosmopolitan species at abyssal depths. Of the ca. 750 protobranch species, most are deposit feeders in soft sediments, but two lineages host chemoautotrophic, sulfide-oxidizing bacteria, with concomitant reductions of the hosts’ alimentary system, including *Solemya velum*. All symbiotic bacterial genomes have been sequenced and identified in *Solemya velum*. Studies corroborating symbiotic bacteria diversity and adaptation can be greatly aided by genomic resources. By generating a reference genome, information can be provided on fundamental questions about the evolution of shell formation and the organismal responses to changes in ocean chemistry.

Here, we will be generating a high-quality genome for *Solemya velum* (Atlantic awning clam - Protobranchia, Bivalvia, Mollusca), whose genome size we have estimated to be around 2 gigabases, about two-thirds the size of a human genome. This project will focus on improving...
genome assemblies in groups where deep coverage Illumina sequencing has produced sub-par assemblies. We are pursuing long read sequencing (Pacific Biosciences) to aid in genome assembly for this taxon. We will use existing genomic resources for genome polishing (100bp and 150bp Illumina sequence information) and existing transcriptomic resources to aid in annotation. We pursue a variety of assembly strategies, including the utilization of software that can combine the two types of data (hybrid assemblies), which has been shown to improve genome assembly (fewer, longer contigs).

This would be the first bivalve genome produced at SI and would strengthen the reputation of the Smithsonian Institute for Biodiversity Genomics and pave the way to tackling increasingly complicated genome projects.
The Harvestman Genome: *Phalangium opilio*

Principal Investigator: Vanessa L. González (Jonathan Coddington, Rebecca B. Dikow, Paul B. Frandsen, Matthew Kweskin, Prashant Sharma, co-PIs)

Projected Timeline: April 2016 start date, 10 month duration

Budget detail: Budget costs for this project include DNA extraction and QA/QC, PacBio SMRT Cells (50), and library prep (2).

Project Synopsis:
This project will focus on improving genome assemblies in groups where deep coverage Illumina sequencing has produced sub-par assemblies. We will pursue long read sequencing (Pacific Biosciences) to aid in genome assembly for this taxon. We will also sequence low error Illumina data for genome polishing. We pursue a variety of assembly strategies, including the utilization of software that can combine the two types of data (hybrid assemblies), which has been shown to improve genome assembly (fewer, longer contigs).

2017 Update:
In order to pursue the study of development in non-model arachnids, establishment of a reference genome key step toward testing hypotheses in evolutionary and developmental biology. The eupnoid harvestman *Phalangium opilio* is a relatively large species with a broad temperate distribution and lays multiple, large clutches of synchronously developing eggs. These qualities make this species an excellent choice for study of development. Active areas of research into this system include, the investigation into the patterning of Hox genes, evolution of cheliceral segmentation, RNAi, and specification of the deutocerebral appendage identity. The *Phalangium opilio* (Harvestman, Daddy long-leg) genome poses several bioinformatics challenges to genome assembly (Genome size: 1.125) as it is highly heterozygosity, highly repetitive. Throughout the course of this project, PI: Prashant Sharma, has cultivated an inbred line of Phalangium opilio, this helps reduce the amount of heterozygosity in the genome. We have currently extracted high molecular weight DNA from 2 individuals of third generation, inbred sibling back crosses and successfully prepped these for genome sequencing. Genome data will be generated using long read sequencing on the Pacific Biosciences newest sequencing platform, the Sequel, which generates up to 50 Kilobase reads and should be completed in the next few weeks.

Blog: The Harvestman Genome: *Phalangium opilio*

My name is Vanessa L. Gonzalez and I am a Computational Genomics Scientist for the Global Genome Initiative (GGI). In my current position, I’m using genomics to understand how species are related across the tree of life. By peering into the genetic makeup of organisms, we can figure out how, where, when, and why biodiversity
happens. I work on a breadth of the organisms from clams, snails, plants, spiders, to Narwhals.

Along with collaborator, Dr. Prashant P. Sharma, Assistant Professor of Biology, University of Wisconsin-Madison, we are interested in using genomic tools to aid in the understanding of how genes shape the body patterning in certain groups of arachnids, or spiders. The spiders in particular we are focusing on are known as “harvestman” or the daddy longlegs.

Harvestmen are important have an extraordinary life history as they carry a record of hundreds of millions of years of geological history; daddy longlegs fossils have been found in 400 million-year-old rocks. This long fossil record has allowed for insights into the change in Earth’s landmasses as the relationships of these animals provide clues into the paths that these landmasses have undertaken.

Additionally, these harvestmen have also recently be use to study how genes shape the patterning of the body during development. But in order to pursue the study of development in these arachnids, establishment of a high-quality genome is key toward testing hypotheses in evolutionary and developmental biology. The harvestman *Phalangium opilio* is a relatively large species with a broad temperate distribution and lays multiple, large clutches of developing eggs. These qualities make this species an excellent choice for study of development. One active areas of research into this system includes, the investigation into the patterning of Hox genes – a group of genes that control the body plan of a developing embryo – similar to those found in humans. The estimated genome size for *Phalangium opilio* is about one-third a human genome, or around 1.1 gigabases.
Our GGI funded project to sequence a high-quality genome for *Phalangium opilio* is part of a group of projects focusing on improving genome assemblies in groups where deep coverage Illumina sequencing has produced sub-par assemblies. We pursued long read sequencing (Pacific Biosciences) to aid in genome assembly for this taxon. We will also have sequenced low error Illumina data for genome polishing. We then will pursue a variety of assembly strategies, including the utilization of software that can combine the two types of data (hybrid assemblies), which has been shown to improve genome assembly (fewer, longer contigs). This would be the first harvestman genome produced at SI and would strengthen the reputation of the Smithsonian Institute for Biodiversity Genomics and pave the way to tackling increasingly complicated genome projects.
Building A Reference Genome For *Centrapalus pauciflorus*, An African Oilseed Crop (*Compositae*)

**Principal Investigator:** Rebecca Dikow (Vanessa González, Paul Frandsen, Matthew Kweskin, Vicki Funk, Sterling Keeley, Morgan Gostel, Jennifer Mandel, Co-Pis)

**Projected Timeline:** April 2016 start date, 9 month duration

**Budget detail:** Budget costs for this project include DNA extraction and quantification, Illumina library prep, Illumina HiSeq lanes (2), and PacBio smart cells (80) plus library prep.

**Project Synopsis:** Plant genomes pose significant bioinformatics challenges because they are often very large, highly repetitive, and polyploid (contain more than one set of chromosomes). There are very few high-quality plant reference genomes for the above reasons and those that do exist are over-represented by domesticated species, which have undergone significant artificial selection. We propose to sequence the genome for a species of the family Compositae, *Centrapalus pauciflorus*, an undomesticated oilseed crop with economic importance. *C. pauciflorus* is rich in vernolic acid, an epoxy fatty acid, which is used in plastics and coatings similar to rubber (Perdue et al., 1986). We have genome-quality tissue growing in a greenhouse and partial funding in hand from the University of Hawaii (to S. Keeley) to obtain long-read DNA sequence data to produce this genome. We are seeking additional funding to augment these sequence data with additional short-read and long-read DNA sequence data with the goal of producing a reference-quality genome. The *C. pauciflorus* genome will provide resources that will allow us and *Compositae* researchers across the world to begin asking and answering evolutionary questions about the success of the family and the evolutionary basis for traits of interest. Moreover, access to a high-quality genome sequence has become a crucial piece of plant breeding programs and efforts aimed at domestication and crop improvement. Our genome could play a role in the improvement of this crop and its economic importance in Africa. In addition, this project would showcase the Smithsonian’s ability to tackle difficult genomes.

**2017 Update:**
The *Centrapalus pauciflorus* genome project is making great progress. We have generated long-read data from the PacBio RSII, RNAseq data on the Illumina HiSeq platform, and short-read Illumina data on a HiSeq. All the data are very high quality. The longest PacBio read we have seen so far was > 87,000 bp. So far, we have completed assemblies for half of the PacBio data and the Illumina DNA data and the results are quite promising: we have a contig N50 of 104,000 bp and an estimated genome size of 1.25Gb. Preliminary work was presented at the Global Biodiversity Genomics Conference in February, 2017 -- poster is shown below. We plan to complete bioinformatics analyses in Fall 2017 and submit the paper for publication by the end of 2017.
Blog: Tackling plant genomics at the Smithsonian

I am Rebecca Dikow, a Research Data Scientist at the Data Science Lab, part of the Smithsonian Office of the Chief Information Officer. We are living in an age of “big data.” Insights from big data touch almost every part of our lives—from the way we navigate in our cars to the way we shop. Big data has also arrived in biodiversity research due to rapid change in the types and volume of data that researchers can use to ask and answer their scientific questions.

For many Smithsonian scientists, genomics is a tool that can be used to broaden their research. Vicki Funk, a Curator and Senior Scientist at the National Museum of Natural History, Jennifer Mandel, an Assistant Professor at the University of Memphis, and I have been working on the phylogenomics (deciphering the evolutionary relationships with genome-scale data) of Compositae (a family of flowering plants that comprises 25,000 species including sunflowers, daisies, lettuce, and thistles) for the past four years. We decided that generating genomes for species of Compositae would be the next step in understanding evolution—particularly gene duplication and polyploidy—across this diverse family.

Plant genomes pose challenges because they are often very large, highly repetitive, and polyploid (containing more than two sets of chromosomes). These characteristics also vary quite a bit across very closely related taxa. In order to begin tackling Compositae genomes, we assembled a team of experts in bioinformatics (Vanessa González, Paul Frandsen, Matt Kveskin, and me), genome sequencing (Marc Allard and Tim Muruvanda), and plant systematics and biology (Sterling Keeley, Jennifer Mandel, Vicki Funk, Gabriel Johnson, and Morgan Gostel).

For our first genome, we chose *Centrapalus pauciflorus*, a diploid, undomesticated African oilseed crop. *C. pauciflorus* is rich in vernolic acid, an epoxy fatty acid, which is used in plastics and coatings similar to rubber (Perdue et al., 1986). We were also able to grow it from seed in a greenhouse at the University of Memphis. This point was key, because we are using a combination of PacBio long-read sequencing and Illumina short read sequencing to build a high-quality reference genome. Long-read sequencing is essential because Compositae genomes are highly repetitive and previous attempts at short-read only genomes produced highly-fragmented assemblies. The long reads produced by PacBio sequencing can span the repetitive regions that are so hard to assemble. In order to have success with long-read sequencing, however, we needed a lot of high-quality DNA. Having a live plant in the greenhouse until we were ready to
extract DNA was a huge benefit. Pictured here is the *C. pauciflorus* voucher as received by the University of Memphis Herbarium.

We have finished sequencing both DNA and RNA for *Centrapalus* and are now working on the genome assembly and annotation. RNAseq was performed to help us annotate the genome. Preliminary hybrid assemblies that included 15X PacBio data and 80X Illumina data have been quite promising, with a contig N50 greater than 100,000 bp. With the success of this first genome, we’re hoping to extend these techniques to many more plant genomes, so we hope you stay tuned for more Smithsonian plant genomics!

We would like to thank GGI, the University of Hawaii, Amazon Web Services and the Intel Corporation for sponsoring this research.