Project Summary

The main goal of this project is to investigate the fire-adapted, thick bark (TB) phenotype in four, economically-important, species of pines with a range along the eastern and south-eastern United States. The evolution of this complex trait will be studied in populations of slash pine (*Pinus elliottii*), pond pine (*P. serotina*), loblolly pine (*P. taeda*), and long leaf pine (*P. palustris*). Next generation sequencing (NGS) of 25 individuals from 20 natural populations for each species (25 x 20 x 4 = 2000 total individuals) will be used to genotype individuals with an eye toward uncovering shared genetic architecture of these four closely related species. The results of this research, while also providing new genomic resources for non-model species and augmenting exiting resources, will not only address fundamental principles in evolutionary biology, but also serve to inform and improve breeding programs and land management initiatives in light of global climate change.

Fire has played an important role in the evolutionary paths of many adaptive traits in plants, across a wide variety of species and habitats. Extant *Pinaceae* species occupy a variety of habitats across North and Central America, Asia, and Europe. These areas have been strongly influenced by fire, and as such, their plants have developed a range of fire-adapted traits. This study will investigate the underlying genetic architecture associated with the complex TB phenotype, the degree to which this architecture is shared among four closely related species across their natural ranges and multiple evolutionary time scales, and how this variability can inform current economic interests.

Intellectual Merit

By focusing on the dissection of an adaptive trait at both microevolutionary and macroevolutionary time scales, this study will address a fundamental principle in evolutionary biology: the mechanism associated with the vast majority of incomplete lineage sorting among populations related species. The rise of shared phenotypes between related species has been a source of debate for many years, dividing into camps supporting either ancient inherited polymorphisms or multiple instances of convergent evolution. Very recently, Ségurel et al. (2012) showed that the ABO blood group antigens in humans are the result of vertical inheritance of maintained, ancient polymorphisms rather than the commonly-held view of convergent evolution of this multiallelic phenotype. In short, this study seeks to investigate the same phenomenon in a quantitative, economically and environmentally-important, quantitative trait in four species of related plants at multiple evolutionary time scales. Additionally, the genomic resources developed in this project will augment and facilitate knowledge transfer of genomic data in non-model species which, until recently, has been technologically out of reach.

Broader Impacts

The PI is committed to fostering open and transparent scientific research. Toward this end, data will be made available for download and exploration from VCU-hosted FTP and web sites. Second, the PI will develop and teach a new graduate course, Applied Ecological Genomics, to introduce and expose students to the bioinformatic skills necessary to effectively manage terabases of data as well as how to exploit NGS data to infer population-level, evolutionary processes. Finally, this will give the PI additional teaching and research experience needed to jumpstart an academic career, focusing on ecologically-relevant problems in an evolutionary context.

Project Description

The PI will investigate local adaptation of a complex trait in 2000 trees from four, economically-important species of pines with a range along the eastern and south-eastern United States (25 trees from 20 populations for each specie). The PI will study thick bark (TB), as a quantitative phenotypic trait, in populations of slash pine (*Pinus elliottii*), pond pine (*P. serotina*), loblolly pine (*P. taeda*), and long leaf pine (*P. palustris*). Using next-generation, high-throughput sequencing, the PI will dissect TB as a complex adaptive trait across populations of these four closely related species with an eye toward uncovering shared genetic architecture at multiple evolutionary times scales.

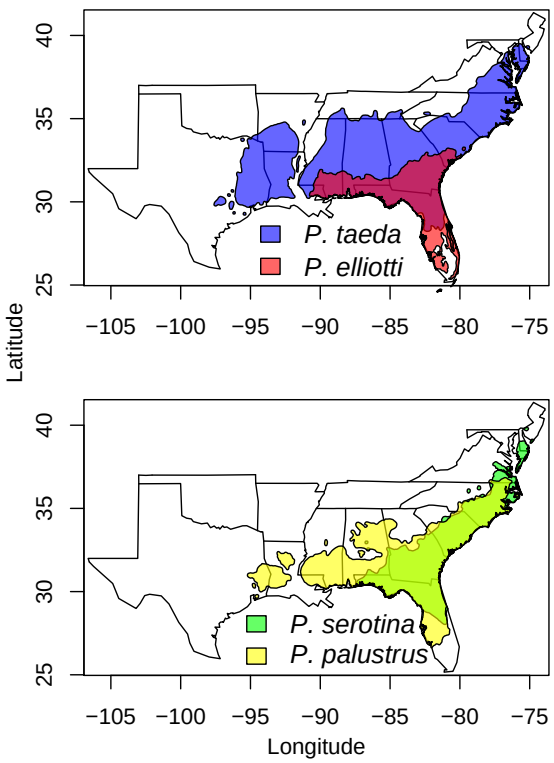

Figure 1: Natural ranges for four study species

Plants display strong patterns of adaptation, especially local adaptation, when populations sizes are large (Leimu and Fischer, 2008). This characteristic is especially important in conifers, which display a lack of population structure and large effective population sizes (Neale and Savolainen, 2004); this is partly why association genetic studies have been so successful in these species (Eckert et al., 2012, 2010; Wegrzyn et al., 2010; Eckert et al., 2009; González-Martínez et al., 2007, 2006; Gupta et al., 2005). Additionally, because linkage disequilibrium is low in conifer populations, it is possible to detect frequency differences in alleles significantly associated with ecologically-relevant phenotypes (Neale and Savolainen, 2004).

Fire has been crucial influence in global ecosystem processes, driving not only locally adapted phenotypes in plant populations (Lamont et al., 1991; Vega et al., 2008; Midgley and Bond, 2011; Keeley et al., 2011; He et al., 2012; Parchman et al., 2012), but also impacting carbon storage and climate (Bowman et al., 2009). Understanding the underlying genetic architecture associated with fire-associated traits is therefore critical to dealing with environmental impacts of increased carbon inputs and changing climate. He et al. (2012) recently investigated five fire-adapted traits for 101 species of *Pinus* and found evidence for a strong influence of fire in the Cretaceous on trait evolution. However, they considered only presence or absence the traits in their study. **There have been no studies to date which capture, as in this proposal, the quantitative nature of fire-associated trait evolution in these species.**.

The application of next-generation sequencing technology to conifer genomes, as proposed in this study, holds the promise of being able to provide a way to dissect the genetic architecture of an adaptive, and quantitative trait. Until recently, techniques to consider adaptation on a genomic scale have been no match for the size and complexity of conifer genomes (Mackay et al., 2012).

However, advances in genomic enrichment techniques such as those by Parchman et al. (2012) and Willing et al. (2011) and the ever-decreasing cost of DNA sequencing have facilitated studies to use genotype-by-sequencing (GBS) approaches to answering fundamental questions in evolutionary biology in non-model species. The PI is well-positioned to utilize the sophistication of the sequencing facilities at VCU coupled with his backgrounds in evolutionary, molecular, and computational biology to study how thick bark has evolved across *Pinus* and how this knowledge may serve to inform ecologically-relevant and economically-important decisions related this important genus.

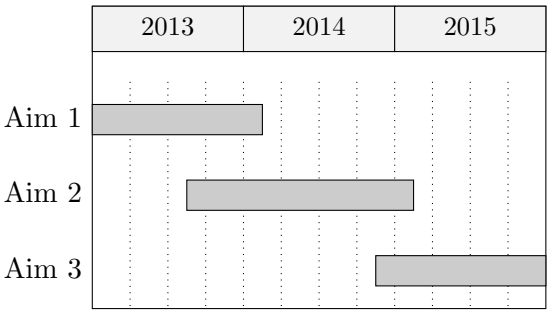

Figure 2: Project timeline

This study is laid out below in three general aims: (1) field sampling, (2) DNA sequencing and bioinformatics, and (3) hypothesis testing. A projected timeline for this project is shown in Figure 2.

Aim 1: Collect field samples, extract genomic DNA, and prepare sequencing libraries

Field sampling The PI will sample four species in their natural ranges, with 25 individuals from 20 natural populations (Figure 1). The locations will be determined randomly. Within each population, a representative of the focal species will be chosen and all other members will be chosen from with 50 m of the first sampling location meeting the requirement of having DBH of at least 15 cm. Three to five needle fascicles will be collected at each from each tree for use in DNA extraction. Needles, with a desiccant, will be stored in 15 ml Falcon™ tubes at ambient temperatures in the field and in −80 °C, long-term in the laboratory.

Thick bark will be treated as a quantitative character and will be estimated for each tree using a regression model based upon tree height and diameter at breast height (DBH) (Cao and Pepper, 1986; Li and Weiskittel, 2010). Measurements of tree height will be obtained using a clinometer. Additionally, measurements of bark thickness will be taken from a representative set of five trees from each population. This will allow for the application of a multiple regression model to non-invasively estimate the bark thickness from the other 80 % of the individuals in each population. Briefly, for each of the sampling trees chosen for invasive sampling, a Haglöff® bark gauge will be used to measure the bark thickness. These measurements will be used as a covariate in the regression equation from Cao and Pepper, which has been shown to significantly decrease bias by up to 43 % in some species (Li and Weiskittel, 2010). At the end of sampling, 2000 individual tubes of needles (20–25 needles/tube) will be banked along with bark thickness measurements and estimates for all trees for all species.

DNA extraction Genomic DNA will be extracted from leaf tissue Qiagen DNEasy extraction kits in a 96-well format using an existing tissue preparation protocol in the Eckert lab. A single DNA extraction from multiple leaves per individual will be performed resulting in 2000 individual genomic DNA extracts.

Criteria for completion Aim 1 will be designated complete when genomic DNA has been extracted for all 2000 individuals. It is estimated that this will be complete by June 2014.

Aim 2: Perform sequencing of 2000 individuals and bioinformatic analysis

The PI will perform genotype-by-sequencing (GBS) utilizing the highly-multiplexed, next-generation sequencing capability of the Illumina HiSeq 2000 platform. Genomic enrichment will be performed using the approach approach outlined in Parchman et al. (2012), in order to ensure high-coverage of the library given the complexity and size of conifer genomes (Mackay et al., 2012). Additionally, each individual will be associated with a computationally-correctable sequencing barcode (454 Life Sciences Corp., 2009). Briefly, the genomic DNA from an individual is digested with two different restriction enzymes coupled with Illumina-specific adaptor ligation. The product is amplified by PCR and size selected on an agarose gel, which produces a pool of fragments for sequencing. Estimates for sequencing include approximately 100 individuals per lane on an eight-lane flow cell, with a per-lane yield of 200 million reads in a paired-end reaction. All costs for sequencing will be absorbed by the Eckert lab. The result of this aim will be individual FASTQ files, one for each lane in the flow cell containing reads for each multiplexed sample in that lane, for each flow cell that is used. So, if 100 trees are multiplexed per lane, 2.5 flow cells will be needed for 20 lanes of sequencing. These 20 lanes will produce four billion reads in total (approximately 8 terabases).

The Nucleic Acids Research Facility (NARF) at VCU is currently producing paired-end reads of 150 bases, but due to the specific error profile of Illumina data, not all of the data is immediately useful. Using software that the PI has already written (Friedline, 2008–2012) coupled with existing libraries (e.g., BioPython (Cock et al., 2009), NGS QC Toolkit (Patel and Jain, 2012)), the processed sequencing data computationally divided into individuals and will be purged of low-quality reads. Additionally, reads that meet acceptance criteria globally, but possess bases (most often at the 3' end) that have low quality scores, will be trimmed and retained. The processed sequence data will be stored in both compressed FASTA data for distribution via FTP as well as in a relational database to enable both sample tracking and metadata association.

Once quality-controlled reads are obtained for each individual (and each population), they will be aligned (mapped) to the most current draft genome of *P. taeda*, available from `dendrome.ucdavis.edu`, using the Burroughs-Wheeler aligner (BWA) (Li and Durbin, 2009), and SNPs will be called. Because loblolly pine is in both our experimental sample as well as the source of the reference genome, we have a unique opportunity to get a measure of confidence in the read mapping algorithms such as BWA as well as in the overall sequencing quality of our data. A SNP is called at particular alignment positions if there are enough high quality reads to indicate a difference between the sample and the reference genome given a model of error. Sequence data from NGS technology requires the use of new SNP calling algorithms that properly model all the sources of errors (e.g, base calling errors, mapping errors, and sample preparation errors, particularly PCR errors can have a significant impact on the analysis). Several tools have been developed to perform this task such as as SAMtools (Li et al., 2009), GATK (McKenna et al., 2010), and, more recently, GeMS (You et al., 2012). The result of this aim will be a sample (n = 2000) by SNP matrix (n = thousands) for each species (n = 4).

Criteria for completion Aim 2 will be complete when all libraries have been sequenced and processed. Processing will include demultiplexing of reads according to individual, quality control

processing of the reads, and storage both on disk and in a relational database system. The sequence data will be made available to the scientific community at this point. Additionally, mapping to the loblolly genome and SNP calling will result in a sample by SNP matrix. This aim is the most computationally complex, and should be complete in early summer 2015.

Aim 3: Perform statistical analysis and hypothesis testing

At the conclusion of the significant bioinformatic effort needed, ecologically-relevant hypotheses can be tested. These null hypotheses are outlined below.

Question 1: What is the genetic architecture of the thick bark phenotype for the four focal pine species that are adapted to fire?

Null hypothesis (H_0) There is no association between genotypic variation at surveyed SNPs and thick bark phenotypes in any of the four species.

Prediction under H_0 Statistical significance after correcting for multiple tests will be lacking for genotype effects in linear models relating SNP genotypes to thick bark for each SNP in each species.

Methods The PI will employ standard linear models that correct for kinship and population structure that are commonly used in genome-wide association studies (GWAS; Yu et al. (2006)). Association analysis has been widely used in and has for dissecting the genetic basis of phenotypes for conifers (Neale and Kremer, 2011; Ingvarsson and Street, 2011). In addition, the PI will explore the use of the Bayesian models presented by Parchman et al. (2012), which was the first GWAS for a conifer, and regression tree approaches presented by Holliday et al. (2012). Linear models will be used to explicitly test the effects of genotypic class of each discovered SNP on quantitative measures of bark thickness for each species. Thus, the number of tests for genotypic effects for each species will be equal to the number of discovered SNPs, which is expected to be in the range of 50 000 to 100 000. Parchman et al. (2012). Multiple tests will be accounted for using a permutation approach that adjusts the global significance level (e.g., $\alpha = 0.05$). In brief, this procedure randomizes genotypic vectors relative to the phenotypic vector to produce a large number of randomized data sets (e.g. 10 000). For each randomized data set, the association analysis is conducted again and the minimum P-value associated with a genotypic effect is retained. Finally, the multiple-test corrected $\alpha = 0.05$ is then the 5 % quantile of the distribution of minimum P-value.

Expected outcomes and relevance The null hypothesis will be rejected when at least one SNP in each of the four pine species is significantly associated to quantitative variation in bark thickness. Previous association studies in conifers typically identify five to 20 plus SNPs associated with the trait under consideration Eckert et al. (2012). These results will classify SNPs discovered into each species into two classes: those associated to bark thickness (which rejects the null hypothesis for this question) and those unassociated with bark thickness. The latter will be used as controls for questions two through four. The former establishes the genetic architecture of bark thickness for each focal species.

Question 2: Is the identified genetic architecture (e.g. the set of SNPs for each species) from Question 1 shared across the four focal species, and if so, is it shared to a greater extent than random variation within the genomes of these four closely related species?

Null hypothesis 1 ($H_0{}^1$) There is no shared genetic architecture for bark thickness among the four focal species.

Null hypothesis 2 ($H_0{}^2$) The degree of allele sharing is greater for loci associated to bark thickness than randomly selected loci

Prediction under ($H_0{}^1$) There are no SNPs associated to bark thickness that are shared across the four focal pine species.

Prediction under ($H_0{}^2$) The fraction of SNPs associated to bark thickness that is shared across species is the same as fraction of SNPs for a random sample of loci.

Methods To evaluate $H_0{}^1$, the PI will leverage the underlying, queryable data storage infrastructure to produce the intersection of SNPs associated with bark thickness in all four focal species. From Aim 2, confidence in SNP calls can be quantified because the genome of the reference organism is represented in the set of focal species. To evaluate $H_0{}^2$, the union of all SNPs associated with bark thickness will be created. The fraction of this list that is shared across all four species will be used as a test statistic, the magnitude of which will be tested via a simple permutation analysis. The PI will construct a null distribution for the fraction of alleles shared across all four species for sets of SNPs randomly selected from those unassociated to bark thickness (see Question 1).

Allele sharing across species is common among conifer species and has been attributed largely to incomplete lineage sorting resulting from recent divergence times and large effective population sizes (Syring et al., 2007; Willyard et al., 2009), ancient admixture (Liston et al., 2007) and long-term gene flow (Zhou et al., 2010). For example, an analysis of levels of long-term gene flow between loblolly and slash pines using multiple random sets of 50 nuclear genes and an isolation-with-migration model (Becquet and Przeworski, 2007) detected low but significant levels of gene flow over the divergence history of these species (unpublished data, $4N_em = 2.5$ with a 95 % confidence interval of 0.52–5.14). The sets of loci randomly sampled from the unassociated loci will be assumed to largely reflect these processes, so that the enrichment of allele sharing at loci associated to a trait such as bark thickness with clear adaptive relevance (He et al., 2012) would imply additional processes such as natural selection, as in Ségurel et al. (2012), in promoting allele sharing among species.

Expected outcomes and relevance $H_0{}^1$ will be rejected when multiple SNPs (i.e., > 5) associated with bark thickness are shared across species. The relevance of this result would be to establish the shared genetic architecture of an adaptive trait. $H_0{}^2$ will be rejected when the fraction of SNPs associated to bark thickness that are shared across the four focal species lies in the upper 1 % tail (i.e., $\alpha = 0.01$ for a one-tailed test) of the null distribution based upon random sets of

unassociated SNPs. The relevance of this result would be to establish that a process beyond incomplete lineage sorting, admixture and gene flow is contributing to the maintenance of trans-species polymorphisms across the four focal species.

Question 3:

Question 4: To what extent is natural selection maintaining the fire-associated, thick bark phenotype in these focal species?

Null hypothesis (H_0) There is no evidence of selection in genes containing SNPs that are significantly associated with thick bark.

Prediction under H_0 In genes containing SNPs associated with thick bark, there will be evidence of natural selection.

Methods As the loblolly pine genome continue to improve, establishing the genomic context of thick bark-associated SNPs will enable the PI to test for evidence of selection in polymorphic, protein coding genes. The PI will select 10 to 20 SNPs from a subset of associated loci from each population for each species. Using genomic position of SNPs and the loblolly genome annotations, the PI will design PCR primers to amplify the protein coding regions from each tree, and each amplicon will be sent for Sanger sequencing. Once the sequences have been obtained and translated to amino acids, they will be aligned at the amino acid level using standard tools such as MAFFT (Katoh et al., 2005) or MUSCLE (Edgar, 2004). From the alignment (back-translated to DNA), d_N/d_S will be estimated using established maximum likelihood methods (Yang, 2007).

Expected outcome and relevance Knowing whether or not genes containing SNPs are under positive ($d_N/d_S > 1$) or purifying selection ($d_N/d_S < 1$) addresses a fundamental question surrounding maintenance of these polymorphisms in natural populations.

Broader Impacts

This project will benefit both the local and global scientific communities, as well as help train the next generation of multi-discipline scientists in the world of big data. First, as soon as the data is curated, it will be made available for download and exploration from FTP and web sites hosted here at VCU. Second, the PI will develop and teach a new three-credit course, BIOL 591 Applied Ecological Genomics, which will train advanced undergraduates and established graduate students, using the data generated in this project, in skills necessary to both understand the technology available at VCU (e.g., Illumina HiSeq/MiSeq, 454, IonTorrent), and to process (via programming), manage (via databases), and analyze the deluge of data generated from current genomics projects. This course will run in conjunction with an existing course, BIOL 693 Ecological Genomics taught by sponsor Eckert, and has received approval from the Biology department chair.

Given the diverse student body at VCU, the proposed research and teaching activities will increase research opportunities for underrepresented groups. VCU is among the top 20 universities in the nation for Biology degrees awarded to ethnic and racial minority students. In 2009, 18.3% of baccalaureate Biology degrees were conferred to African American students, exceeding the national average in Biology by 11.5% (NSF, 2010). A high percentage of VCU Biology (39.4%) baccalaureate

degrees were also earned by Asian/Pacific Islanders, a value that far exceeds the national average in Biology of 6.2%. Caucasian students were awarded 34.9% of VCU baccalaureate Biology degrees as compared with the national average in Biology of 66.4% (NSF, 2010).

Results From Prior NSF Support

PI Friedline has yet to receive funding from the National Science Foundation.

References Cited

Laure Ségurel, Emma E Thompson, Timothée Flutre, Jessica Lovstad, Aarti Venkat, Susan W Margulis, Jill Moyse, Steve Ross, Kathryn Gamble, Guy Sella, Carole Ober, and Molly Przeworski. The ABO blood group is a trans-species polymorphism in primates. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, October 2012.

Roosa Leimu and Markus Fischer. A meta-analysis of local adaptation in plants. *Plos One*, 3(12): e4010, 2008.

David B Neale and Outi Savolainen. Association genetics of complex traits in conifers. *Trends in Plant Science*, 9(7):325–330, July 2004.

Andrew J Eckert, Jill L Wegrzyn, W Patrick Cumbie, Barry Goldfarb, Dudley A Huber, Vladimir Tolstikov, Oliver Fiehn, and David B Neale. Association genetics of the loblolly pine (Pinus taeda, Pinaceae) metabolome. *New Phytologist*, 193(4):890–902, March 2012.

Andrew J Eckert, Joost van Heerwaarden, Jill L Wegrzyn, C Dana Nelson, Jeffrey Ross-Ibarra, Santiago C González-Martínez, and David B Neale. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). *Genetics*, 185(3):969–982, July 2010.

Jill L Wegrzyn, Andrew J Eckert, Minyoung Choi, Jennifer M Lee, Brian J Stanton, Robert Sykes, Mark F Davis, Chung-Jui Tsai, and David B Neale. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem. *New Phytologist*, 188(2):515–532, October 2010.

Andrew J Eckert, Andrew D Bower, Jill L Wegrzyn, Barnaly Pande, Kathleen D Jermstad, Konstantin V Krutovsky, J Bradley St Clair, and David B Neale. Association genetics of coastal Douglas fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-hardiness related traits. *Genetics*, 182(4):1289–1302, August 2009.

Santiago C González-Martínez, Nicholas C Wheeler, Elhan Ersoz, C Dana Nelson, and David B Neale. Association genetics in Pinus taeda L. I. Wood property traits. *Genetics*, 175(1):399–409, January 2007.

Santiago C González-Martínez, Konstantin V Krutovsky, and David B Neale. Forest-tree population genomics and adaptive evolution. *New Phytologist*, 170(2):227–238, 2006.

Pushpendra K PK Gupta, Sachin S Rustgi, and Pawan L PL Kulwal. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology*, 57(4):461–485, March 2005.

Byron B Lamont, D C Maitre, R M Cowling, and N J Enright. Canopy seed storage in woody plants. *The Botanical Review*, 57(4):277–317, October 1991.

J A Vega, C Fernández, P Pérez-Gorostiaga, and T Fonturbel. The influence of fire severity, serotiny, and post-fire management on¡ i¿ Pinus pinaster¡/i¿ Ait. recruitment in three burnt areas in Galicia (NW Spain). *Forest Ecology and Management*, 256(9):1596–1603, 2008.

Jeremy Midgley and William Bond. Pushing back in time: the role of fire in plant evolution. *New Phytologist*, 191(1):5–7, July 2011.

Jon E JE Keeley, Juli G JG Pausas, Philip W PW Rundel, William J WJ Bond, and Ross A RA Bradstock. Fire as an evolutionary pressure shaping plant traits. *Trends in Plant Science*, 16(8): 406–411, August 2011.

Tianhua He, Juli G Pausas, Claire M Belcher, Dylan W Schwilk, and Byron B Lamont. Fire-adapted traits of Pinus arose in the fiery Cretaceous. *New Phytologist*, 194(3):751–759, May 2012.

Thomas L Parchman, Zachariah Z Gompert, Joann J Mudge, Faye D Schilkey, Craig W Benkman, and C Alex Buerkle. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, 21(12):2991–3005, June 2012.

David M J S Bowman, Jennifer K Balch, Paulo Artaxo, William J Bond, Jean M Carlson, Mark A Cochrane, Carla M D'Antonio, Ruth S Defries, John C Doyle, Sandy P Harrison, Fay H Johnston, Jon E Keeley, Meg A Krawchuk, Christian A Kull, J Brad Marston, Max A Moritz, I Colin Prentice, Christopher I Roos, Andrew C Scott, Thomas W Swetnam, Guido R van der Werf, and Stephen J Pyne. Fire in the Earth system. *Science*, 324(5926):481–484, April 2009.

John Mackay, Jeffrey F D Dean, Christophe Plomion, Daniel G Peterson, Francisco M Cánovas, Nathalie Pavy, Pär K Ingvarsson, Outi Savolainen, M Ángeles Guevara, Silvia Fluch, Barbara Vinceti, Dolores Abarca, Carmen Díaz-Sala, and María-Teresa Cervera. Towards decoding the conifer giga-genome. *Plant Molecular Biology*, pages –, September 2012.

Eva-Maria Willing, Margarete Hoffmann, Juliane D Klein, Detlef Weigel, and Christine Dreyer. Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, 27(16):2187–2193, August 2011.

Quang V Cao and William D Pepper. Predicting Inside Bark Diameter for Shortleaf, Loblolly, and Longleaf Pines. *Southern Journal of Applied Forestry*, 10:220–224, 1986.

Rongxia Li and Aaron R Weiskittel. Estimating and predicting bark thickness for seven conifer species in the Acadian Region of North America using a mixed-effects modeling approach: comparison of model forms and subsampling strategies. *European Journal of Forest Research*, 130 (2):219–233, August 2010.

454 Life Sciences Corp. Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Extended MID Set . *Technical Bulletin*, TCB 005-2009:1–7, April 2009.

Christopher J. Friedline. Code repository, 2008–2012. URL `http://code.friedline.net`.

Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009.

Ravi K Patel and Mukesh Jain. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *Plos One*, 7(2):e30619–e30619, 2012.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

Heng Li, Bob Handsaker, 1000 Genome Project Data Processing Subgroup, and 10. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010.

N You, G Murillo, X Cui, and 9. SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, pages –, January 2012.

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S. Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, February 2006.

David B Neale and Antoine Kremer. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, 12(2):111–122, February 2011.

Pär K Ingvarsson and Nathaniel R Street. Association genetics of complex traits in plants. *New Phytologist*, 189(4):909–922, March 2011.

Jason A JA Holliday, Tongli T Wang, and Sally S Aitken. Predicting Adaptive Phenotypes From Multilocus Genotypes in Sitka Spruce (Picea sitchensis) Using Random Forest. *G3: Genes, Genomes, Genetics*, 2(9):1085–1093, September 2012.

John Syring, Kathleen Farrell, Roman Businský, Richard Cronn, and Aaron Liston. Widespread genealogical nonmonophyly in species of Pinus subgenus Strobus. *Systematic biology*, 56(2): 163–181, April 2007.

Ann Willyard, Richard Cronn, and Aaron Liston. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. *Molecular phylogenetics and evolution*, 52(2):498–511, August 2009.

Aaron A Liston, Mariah M Parker-Defeniks, John V JV Syring, Ann A Willyard, and Richard R Cronn. Interspecific phylogenetic analysis enhances intraspecific phylogeographical inference: a case study in Pinus lambertiana. *Molecular Ecology*, 16(18):3926–3937, September 2007.

Yong Feng Zhou, Richard J Abbott, Zu Yao Jiang, Fang K Du, Richard I Milne, and Jian Quan Liu. Gene flow and species delimitation: a case study of two pine species with overlapping distributions in southeast China. *Evolution; international journal of organic evolution*, 64(8): 2342–2352, August 2010.

Celine Becquet and Molly Przeworski. A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17(10):1505–1519, October 2007.

Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, 2005.

Robert C Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics*, 5:113, August 2004.

Ziheng Z Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, August 2007.

C J Friedline, R B Franklin, S L McCallister, and M C Rivera. Bacterial assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal biogeographic signatures. *Biogeosciences*, 9 (6):2177–2193, 2012.

James A Lake and Maria C Rivera. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular Biology and Evolution*, 21(4):681–690, April 2004.

Maria C Rivera and James A Lake. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005):152–155, September 2004.

Biographical Sketch: Christopher J. Friedline

1 Professional preparation

Virginia Commonwealth University	Integrative Life Sciences	Ph.D.	5/2013
Virginia Commonwealth University	Bioinformatics	P.S.M.	2008
Gettysburg College	Biology	B.A.	1999

2 Appointments

- Teaching Assistant, Department of Biology, Virginia Commonwealth University, 2008, 2011
  - Biological Concepts Laboratory (BIOZ 101), Fundamentals of Molecular Genetics (BIOL 540), Plant and Animal Cell Biology (BIOL 676)
- Teaching Assistant, Department of Life Sciences, Virginia Commonwealth University, 2008
  - Introduction to Life Sciences (LFSC 101), Integrated Bioinformatics (BNFO 601)
- Senior Network Analyst, VCU Health System, Richmond, VA, 2006–2007
- IT Systems Manager, Infralogics, Inc., Richmond, VA, 2005–2006
- IT Systems Administrator, Infineon Technologies, Sandston, VA, 2005
- Senior Systems Administrator, Circuit City Stores, Inc., Richmond, VA, 2002–2005
- NT Systems Administrator, Virginia Interactive, Richmond, VA, 2002
- Systems Administrator/Senior Systems Analyst/Systems Analyst, Bently Systems, Inc., Exton, PA, 1999-2002

3 Publications

- **Friedline, C. J.** & Rivera, M. C. Bayesian Inference: a more powerful approach to recovering the true microbial community structure. In preparation (2012).
- Gilbreath, J. J., Semino-Mora, C., **Friedline, C. J.**, et al. A Core Microbiome Associated with the Peritoneal Tumors of Pseudomyxoma Peritonei. In preparation (2012).
- **Friedline, C. J.**, Franklin, R. B., McCallister, S. L. & Rivera, M. C. Bacterial assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal biogeographic signatures. Biogeosciences 9, 2177-2193 (2012).
- Fettweis, J. M., Alves, J. P., Borzelleca, J. F., **Friedline, C. J.**, et al. The Vaginal Microbiome: Disease, Genetics and the Environment. Nature Precedings (2011).
- **Friedline, C. J.**, Zhang, X., Zehner, Z. & Zhao, Z. FindSUMO: A PSSM-Based Method for Sumoylation Site Prediction. Advanced Intelligent Computing Theories and Applications With Aspects of Artificial Intelligence 1004–1011 (2008).

4 Synergistic activities

- Invited to participate in two 2011 MBL courses: Workshop on Molecular Evolution and Strategies and Techniques for Analyzing Microbial Population Structures
- Designed and implemented informatic pipeline and data acquisition system for the VCU Vaginal Human Microbiome Project, currently supporting 1,962 samples of 53 million 454 sequencing reads (23 billion base pairs). Data are housed in a PostgreSQL database and made available via a Java-based web portal.
- Developed quality control metrics and the genomes of three antibiotic resistant *S. aureus* strains using Illumina reads. Comparative genomic analyses identified several SNPs associated with the resistant phenotype.
- Created custom software to visually and statistically analyze microarray data. `http://sledride.csbc.vcu.edu`
- Responsible for the design and implementation of multi-million dollar, enterprise-class systems involving integrated technology stacks, multi-terabyte, n-tier architecture, and high-availability across many industry sectors including academic, medical, manufacturing, and retail.

5 Collaborators and other affiliations

Collaborators and Co-Editors

- Rima B. Franklin, Department of Biology, Virginia Commonwealth University
- S. Leigh McCallister, Department of Biology/Center for Environmental Studies, Virginia Commonwealth University

Graduate Advisors

- Maria C. Rivera, Ph.D. Advisor, Center for the Study of Biological Complexity/Department of Biology, Virginia Commonwealth University
- Paul M. Fawcett, P.S.M. Advisor, Center for the Study of Biological Complexity/Department of Internal Medicine, Virginia Commonwealth University

Dissertation Abstract

Using Molecular Sequence Data to Unravel the Relationships between Bacteria and their Environment

Microbial communities are recognized as major drivers of global biogeochemical processes. However, the genetic diversity and composition, as well as processes leading to the origin and diversification of these communities in space and time, are poorly understood. Characterization of microbial communities using high-throughput sequencing of 16S tags shows that OTU abundances can be approximated by a gamma distribution, which suggests structuring around small numbers of highly abundant OTUs and a large proportion of low abundant, rare, OTUs. The current methods used to characterize how communities are structured rely on multivariate statistics, which operate on pair-wise distance matrices. My analyses demonstrate that use of these methods, by reducing a highly-dimensional dataset (tens of samples, thousands of OTUs), result in a significant loss of information (Friedline et al., 2012). I demonstrate that, in some cases, up to 80% of the most abundant OTUs may be removed while still recovering the same community relationships; this indicates these metrics are biased toward the highly abundant OTUs. I will demonstrate that the observed patterns of OTU abundance detected from microbial communities can be properly modeled using techniques similar to those we used to model the presence and absence of genes in genome evolution (Lake and Rivera, 2004; Rivera and Lake, 2004). Using simulation studies, I demonstrate that general Markov models in a Bayesian inference framework outperform traditional, multivariate ecological methods in recovering true community structure. Applying this new methodology to Atlantic Ocean microbial communities allowed us to uncover a distance-decay effect, which was not revealed by the traditional methods (Friedline et al., 2012). Although the ocean dataset operated on a much larger, continental scale, characterization of the sequence data generated from the a nutrient poor soil on Hog Island, a barrier island off the Virginia Coast, allows for a better characterization of the processes affecting these communities on a much smaller scale. Finally, using 16S data from the Vaginal Human Microbiome Project, generated here at VCU under the umbrella of the overall NIH HMP initiative, I show that quality filtering has a profound effect on the reliability of downstream analysis. In conclusion, my analyses of the metagenomic sequence data from three types of bacterial communities demonstrate that the proper identification of the biological process influencing these communities requires the development and implementation of new statistical and computational methodology that takes advantage of the extensive amount of information generated by high-throughput sequencing.

Data Management Plan

The PI is committed to fostering an open and transparent process of data generation, analysis, and publication. As required in the NSF PGRP guidelines, genome level sequence data will be made available as soon as data quality and integrity have been verified via simple FTP sites hosted by sponsor Eckert. In addition, we will deposit data and computer code in standard public repositories prior to, and in all cases regardless of, publication (e.g., GenBank, Dryad, TreeGenes).

Data Types

Genetic Genetic sequence data from Illumina and 454 platforms will be stored as compressed FASTA/FASTQ files on disk. The quality-filtered, working set of sequence data will be stored in a relational database management system (RDBMS) built on PostgreSQL following an existing Rivera lab data storage model currently in production for the Vaginal Human Microbiome Project. Reads will be tracked through a barcoding system relating sequence data uniquely with individuals, populations, and additional metadata, as appropriate (e.g., sampling location, SNP genotypes, annotation, etc.). A copy of the processed, quality-controlled data will be extracted automatically on a schedule and compressed for distribution via FTP. Additionally, a web interface will be provided for dynamic data exploration.

Spatial and phenotypic All metadata will be stored in the same relational database system. This includes latitude and longitude for all sampling locations as well as phenotypic measurements and observations (e.g., percent serotiny, diameter a breast height (DBH), height).

Data Standards

All data, including data generated in downstream analysis, will be made available in a plain-text format. Working sequence data will be stored as text in the RDBMS, as well as associated environmental and phenotypic data. These data will be made available in two ways. First, as downloadable, compressed files available via FTP. Second, as dynamically generated tables available from a web interface. Downloadable data will be in tab-delimited format, with informative headers as appropriate. This format was chosen as it conforms to a standard used in programs such as R, as well as to any current or new custom Java and Python code developed in this project.

Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

Genomic data will be released following the Bermuda/Ft. Lauderdale agreement and will also be deposited in GenBank once data quality is sufficient for release. The remaining data will be released after publication in the text and web formats as described above, along with a copy of the publication, and will be posted on web servers in both the Eckert and Rivera laboratories. All relevant code and schemas used to process, store, and analyze these data will be made available under and open source license and distributed via the web (e.g., bitbucket.org). There will be no charge for these data. There are no known privacy or ethical issues associated with the data collected on this grant. There are no known copyright, licensing, or other intellectual property issues associated with these data.

Policies and Provisions for Re-use and Re-distribution

The data will not contain any permission restrictions beyond citation of our original publication(s) where the data were introduced. The genetic data will be of potential interest to the landscape genomic community (both professional and academic) for subsequent meta analysis, statistical model development, and other summary uses. The PI does not foresee any reason to restrict the re-use of these data to any entity or group.

Plans for Archiving and Preservation of Access

Raw DNA template will be stored at VCU in the Eckert Laboratory in 1.5ml Eppendorf tubes at -80 °C for least as long as it takes to publish all the materials associated with this project. Individual primers will not be made available, though published sequences can be used to create them de novo if necessary. Textual genetic data will be made available, via the web or some other publicly available medium as described above, for as long as the PI is engaged in academic research. Unprocessed leaf tissue samples will be similarly archived until publication, with raw physiological data (e.g. in the case of parameters derived from light-response curves) archived and available to the public following publication. For the foreseeable future, these data will be housed on the servers located within the Eckert Laboratory. All final data will be automatically backed up to VCU common spaces. There are no known issues associated with storing textual genetic data or anonymizing the identification numbers on these data sets.