NCSU Genomic Sciences
Functional Genomics | Bioinformatics

NC STATE
Biomathematics Graduate Program

Genomic Sciences & Biomathematics Symposium

Spring 2018

Friday, April 6th
6:00PM – 10:00PM
Museum of Natural Sciences – Environmental Conference Center

Saturday, April 7th
9:00AM – 5:00PM
James B Hunt Jr. Library – Room 4106

We would like to acknowledge the following people for funding (in alphabetical order): Dr. David Bird, Dr. Spencer Muse and Dr. Fred Wright. We would also like to acknowledge the Bioinformatics Research Center and the NC State Graduate Student Association.

Schedule of Events

Friday, April 6th

6:30PM – 7:15PM	Talks
	Cristina Lanzas ,David Rasmussen, Benjamin Callahan
7:15PM – 8:00PM	Keynote Speaker – Sergei Kosakovsky Pond
8:00PM – 9:00PM	Dinner & Poster Session
	Eli Buckner, Marco Hamins-Puertolas, Courtney Klotz, Jun Ma, Matt Nethery, Rocky Patil, Alice Toms
9:00PM – 10:00PM	Social Hour

Saturday, April 7th

9:00AM – 9:45AM	Breakfast & Coffee
10:00AM – 10:15AM	Welcome
10:15AM – 11:45AM	Talks
	Dr. Hamid Ashrafi, Wenbin Zhou, Yue Hao, Annabel Meade, Katelyn Brandt, Sadie Wisotsky
11:45AM – 11:55AM	*Break*
11:55AM – 12:55PM	Lightning Talks
	Dr. Hamed Bostan, Mitchel Colebank, Cody Ellington, Tao Jiang, E. Benjamin Randall, Lauren E. Redpath, Michael Vella, Jamie Nosbich
1:00PM – 2:00PM	Lunch
2:00PM – 3:30PM	Talks
	Brandon Hollingsworth, Neha Murad, Jeremy Ash, Kyle Roell, Hayden Brochu, Amanda Reeder
3:30PM – 3:45PM	*Break*
3:45PM – 5:00PM	Talks
	Kimberly To, Haonan Tong, Sadie Wisotsky, Alexandra Crawley, Samiul Haque, Patrick Perkins
5:00PM	After Hours Social

Sergei Kosakovsky Pond

Keynote Speaker

Originally trained as a computer scientist and a numerical physicist (Kiev State University), Dr. Pond was rescued from a life as a (poor) mathematician by a chance encounter with Dr. Muse while they were both at the University of Missouri, Columbia. This is also where the development of the HyPhy package (hyphy.org) started. Due to this fortuitous turn of events, Sergei was able to join the field of bioinformatics and computational biology on the ground floor, going on to receive a PhD in applied mathematics (statistical methods for sequence analysis) from the University of Arizona in 2003. He spent the next twelve years at UC San Diego, first as a postdoc, then as faculty in the Department of Medicine. As of 2016, he is a Professor of Biology at Temple University (Institute for Genomics and Evolutionary Medicine).

Over the past fifteen years, Dr. Pond has been developing bioinformatics tools and statistical methods for sequence analysis, and applying them, almost always with a large number of interdisciplinary collaborators, to the study of various systems, especially viral pathogens (HIV-1, Influenza A virus, Ebola virus, etc.).

His research is not limited to any particular system, any particular set of evolutionary questions, or any particular methodological or computational approach: rather the objective is to identify and implement a practically useful, and statistically justifiable solution to a particular problem. Dr. Pond's group is focused on delivering high quality, computationally efficient and user friendly scientific software offering solutions to questions like this:

- How does one extract evolutionary imprints left by the action of natural selection, recombination, and other processes on sequence data (www.datamonkey.org)?
- How can sequence data be used to understand the dynamics of pathogen transmission (especially HIV-1, www.hiv-trace.org)?
- How does one mine deep sequencing data to understand within-host evolution of pathogens and the concommitant immune response (www.antibodyo.me)?
- How can one deliver software that allows to test various evolutionary hypotheses (www.hyphy.org)?"

Presentation Session 1 – Friday April 6th 6:30PM-7:15PM

Data-driven and modeling approaches to mitigate antibiotic resistance
Cristina Lanzas, Associate Professor, Population Health and Pathobiology

Antibiotic resistance is one of the foremost challenges in public health. Antibiotic drug use in medicine and agriculture generates selective pressures that select for resistance. Efforts to decrease antibiotic use are central to most antimicrobial resistance mitigation efforts. However, a decrease in use is not always followed by a decrease in resistance. I will discuss how we can use data-driven and mathematical approaches for minding clinical and surveillance data to understand the persistence of resistance and to discover and evaluate mitigation strategies at the hospital and at national level.

Phylodynamic insights into the South African HIV epidemic in the era of universal treatment
David Rasmussen, Assistant Professor, Department of Entomology and Plant Pathology, Bioinformatics Research Center

Despite increasing access to effective antiretroviral drugs, the rate of new HIV infections has not declined in many populations around the world. Recently, clinical trials run specifically to test whether increasing antiretroviral coverage prevents new infections at the population level have also returned negative results. I will present a phylodynamic analysis of one such clinical trial in South Africa where we combined epidemiological modeling with a phylogenetic analysis of HIV sequence data to determine the source of new HIV infections during the trial. Our results strongly suggest that antiretroviral treatment does effectively reduce transmission, but delays in linking infected individuals to care largely undermines the ability of antiretroviral drugs to prevent new infections.

Building a better metagenomics microscope.
Dr. Benjamin Callahan, Assistant Professor, Department of Population Health and Pathobiology, Bioinformatics Research Center

The study of microbiomes has been revolutionized by the application of marker-gene and metagenomics sequencing (MGS) technologies, but MGS measurements of community composition are often imprecise and not quantitatively comparable between studies. I will present some recent computational methods from my lab that improve the precision and quantitative comparability of MGS measurements, including algorithms for the inference of exact amplicon sequence variants (ASVs) and a new statistical technique for identifying and removing contaminants in MGS data.

Gene expression data extraction from microscopy images in the *A. thaliana*

Eli Buckner under the advisement of *Dr. Cranos Williams*
Program: Electrical and Computer Engineering

Understanding biological low-level metrics such as spatiotemporal gene expression measurements give great insight into the mechanics of developmental plant biology. Recent technological advancements such as digital imaging systems, fluorescent microscopy, and gene tagging with Fluorescent Proteins (FP) make it possible for us to observe gene expression in living plant organisms in 3D digital images over time-course experiments (4D). Automated software for analyzing these images is necessary for high throughput data extraction of spatiotemporal gene expression measurements. A common practice is to tag a gene localized to the cell walls or the cell nuclei with a FP so that each individual cell is distinguishable in the images while also tagging a gene-of-interest with a different FP. This allows for already existing automated software solutions to measure and distinguish gene expression in relative locations within an organism; however, the process of tagging a gene localized to the cell walls or the cell nuclei in plants is time consuming and creates a bottleneck to data collection because experimentation is reliant on this step. Thus, an automated solution that does not require the tagging of genes localized to all cell nuclei or cell walls is desirable. We propose an automated image analysis pipeline that can extract spatiotemporal gene expression data without the need for a fluorescent tag in cell walls or cell nuclei for the *Arabidopsis thaliana* root. Through computer vision and pattern recognition techniques, we analyze a non-fluorescent image (bright field) to create a custom coordinate system inside the root in addition to a fluorescent image which contains the tagged gene-of-interest. The regions of gene expression found from the fluorescent image is automatically located, segmented, and measured and mapped into the organism through the use of our custom coordinate system. We run this analysis pipeline and show its usefulness in evaluating the effects of heat shock on the plant's cell cycle by analysis gene expression of the CCNB1 gene during and after heat application.

Modeling stochastic viral evolution: A multiscale Wright Fisher model

Marco Hamins-Puértolas under the advisement of *Dr. Ruian Ke*
Program: Biomathematics

Viral evolution is influenced by both demographic changes during infection at the in-host level and transmission at the between-host level. Here we implement a Wright-Fisher model at both the host and the population level. We analyze how key parameters including bottleneck size, time between transmission events, and selection impact probability of fixation and time to fixation. Both host population and bottleneck size have a role in determining probability of fixation. As the time between transmission events increases, generations to fixation is practically entirely dependent on host population size. This model provides a quantitative framework to study how population dynamics alter evolutionary dynamics in viruses across scales.

Investigating the effect of growth phase on the surface-layer associated proteome of *Lactobacillus acidophilus* using quantitative proteomics
Courtney Klotz under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Bacterial surface-layers (S-layers) are semi-porous crystalline arrays that self-assemble to form the outermost layer of some cell envelopes. S-layers have been shown to act as scaffolding structures for the display of auxiliary proteins externally. These S-layer associated proteins have recently gained attention in probiotics due to their direct physical contact with the intestinal mucosa and potential role in cell proliferation, adhesion, and immunomodulation. A number of studies have attempted to catalog the S-layer associated proteome of Lactobacillus acidophilus NCFM under a single condition. However, due to the versatility of the cell surface, we chose to employ a multiplexing-based approach with the intention of accurately contrasting multiple conditions. In this study, a previously described lithium chloride isolation protocol was used to release proteins bound to the L. acidophilus S-layer during logarithmic and early stationary growth phases. Protein quantification values were obtained via TMT (tandem mass tag) labeling combined with a triple-stage mass spectrometry (MS3) method. Results showed significant growth stage-dependent alterations to the surface-associated proteome while simultaneously highlighting the sensitivity and reproducibility of the technology. Thus, this study establishes a framework for quantifying condition-dependent changes to cell surface proteins that can easily be applied to other S-layer forming bacteria.

Modeling nonlinear dose-response relationships using evolutionary computation
Jun Ma under the advisement of *Dr. Alison Motsinger-Reif*
Program: Bioinformatics

Nonlinear dose-response relationships exist extensively in the cellular, biochemical, and physiologic processes that are affected by differing levels of biological, chemical or radiation stress. The traditional model fitting methods like nonlinear least squares are very sensitive to initial parameter values and often suffer from convergence failure. Therefore, we propose the use of an evolutionary algorithm for dose-response modeling. This new method can not only fit the most commonly used nonlinear dose-response models, such as exponential models, 3-parameter logistic models, 4-parameter logistic models and 5-parameter logistic models, but can also select the best model if no model assumption is made. Compared with nonlinear least squares, the new method provides stable and robust solutions without sensitivity to initial values.

Rapid identification and visualization of CRISPR loci via automated high-throughput processing pipeline
Matt Nethery under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Clustered regularly interspaced short palindromic repeats (CRISPR) and associated sequences (cas) comprise an adaptive immune system widespread in bacteria and archaea. Located between CRISPR repeats is a short segment of DNA collected from an invading mobile genetic element, called a spacer. Visualizing iterative spacer acquisitions representing unique evolutionary tracks has proven useful for genotyping, especially for comparative analysis of closely related organisms, and even clonal lineages. Current spacer visualization methods are tedious and typically require manual data manipulation and curation, including spacer extraction at each CRISPR locus from a genome of interest. Once spacers have been isolated, information regarding their length and content must be laboriously distilled and summarized into a format suited for comparative analysis. Here, we present a high-throughput processing pipeline and web-based visualization tool, facilitating spacer extraction, rapid visualization, graphical comparison, and alignment. The analysis pipeline automates the extraction of spacers from a large number of genomes simultaneously, then feeds the resulting spacer files into a visualization engine for comparison of spacer length and content. Additional manipulation, including multiple sequence alignment, can be performed from the graphical user interface. Due to its ability to process unannotated genome files with minimal preparation, this pipeline can be implemented promptly. This efficient high-throughput solution supports rapid analysis of large data sets and will enable and expedite large-scale genotyping efforts based on CRISPR loci.

Poor feed efficiency in sheep is associated with several structural abnormalities in the community metabolic network of their ruminal microbes

Rocky Patil under the advisement of *Dr. Gavin Conant*
Program: Animal Sciences

Ruminant animals have a symbiotic relationship with the microorganisms in their rumens. In this relationship, the host protects the microbes while they degrade complex plant-derived compounds that cannot be metabolized by any enzyme the animal encodes. The resulting simpler metabolites can then be absorbed by the host and converted into other compounds with its own enzymes. We used a microbial metabolic network inferred from shotgun metagenomics data to assess how this metabolic system differs between animals that are able to turn ingested feedstuffs into body mass with high efficiency and those that are not. We conducted shotgun sequencing of microbial DNA from the rumens of sixteen sheep that differed in their residual feed intake (**RFI**), a measure of feed efficiency. Metagenomic reads from each sheep were mapped onto a database-derived microbial metabolic network, which was linked to the sheep metabolic network by interface metabolites (metabolites transferred from microbes to host). Somewhat surprising, no single enzyme was identified as being significantly different in abundance between the low and high RFI animals ($P > 0.05$, Wilcoxon Test). However, when we analyzed the metabolic network as a whole, we found several differences between efficient and inefficient animals. Microbes from low RFI (efficient) animals use a suite of enzymes closer in network space to the host's reactions than those of the high RFI (inefficient) animals. Similarly, low RFI animals have microbial metabolic networks that, on average, contain reactions using shorter carbon chains than do those of high RFI animals, potentially allowing the host animals to extract metabolites more efficiently. Finally, the efficient animals possess community networks with greater Shannon diversity among their enzymes than do inefficient ones. Using a systems approach, we were able to discern differences attributable to feed efficiency in the structure of the ruminal metabolic network that were undetectable solely at the level of individual microbial taxa or reactions.

On the soil mycobiome associated with orchids in Sweden

Alice Toms under the advisement of *Dr. Ignazio Carbone*
Program: Bioinformatics

While microbiomes continue to garner an ever-increasing amount of attention, mycobiomes that focus on fungi are still relatively understudied. There have been only a few large-scale fungal biodiversity studies that have elucidated the genetic composition and relationships of fungi and their surrounding plants. In this study, we collected soil samples associated with Swedish orchids. Operational taxonomic units (OTUs) were identified and used to characterize members of the soil fungal community using QIIME and the UNITE fungal ITS r-DNA reference database. Sequence data were further analyzed using phylogeny-based placement methods implemented in the Tree-Based Alignment Selector (T-BAS) toolkit version 2.1. We classified unknowns at high taxonomic levels (i.e phylum) and we conducted categorical analyses to determine if an association exists between these orchid species and any beneficial fungi on the phylum level. Additionally, by partitioning the taxonomic layers into subtrees of higher phylogenetic resolution we observed a higher than expected presence of *Sebacinales* in our soil samples, an order of fungi which is reported to be of relatively higher abundance for orchid species. Our study showed that although many different fungi comprise the soil mycobiome of orchid species, there are some that are recovered more frequently and may be important components of the microbiome of Swedish orchid species.

Developing genomics and bioinformatics resources for blueberry breeding
Dr. Hamid Ashrafi, Assistant Professor, Department of Horticultural Science, Bioinformatics Research Center

Blueberries are members of the Ericaceae family and include several subgenera or sections. Vaccinium section Cyanococcus is native to North America and all species of this section have contributed to the genetic background of most commercially important cultivars. Traditional breeding efforts to develop superior blueberry cultivars began in 1908 and, as a result, many of today's cultivars are the product of interspecific hybridization followed by backcrossing. Consequently, modern cultivars are segmental allopolyploids, which share a complex ancestry resulting from the intercrossing of different wild accessions and cultivated varieties. The history of blueberry breeding at NC State dates back to 70 years ago. However, up until recently all breeding efforts at NC State and other public breeding institutes were focused on traditional breeding. There was much needed molecular resources for breeding blueberries and other small fruits such as blackberry and red raspberry at NC State. Our lab is focused on genome and transcriptome sequencing of blueberry, blackberry and raspberry and their pathogens. The haploid genome size of blueberry is 670 Mb and that of diploid blackberry and raspberry is ~280 Mb. With relatively small genomes compared to the other plant species, nowadays, it is possible to sequence many blueberry or blackberry genomes at a low cost with Illumina short read sequencing. However, in order to have a high quality reference genome it is preferable to use long read sequencing. Align with this objective, an effort started in our lab in 2015 to sequence the genome of a diploid blueberry, and more recently we finished genome assembly of a tetraploid blueberry, a fungal pathogen as well as two diploid blackberries. We used Single Molecule Real Time (SMRT) sequencing technology aka PacBio for all of our sequencing projects. The assembly and the required resources to make the assembly will be discussed. In addition to genomic sequences we have generated an immense amount of data for transcriptomes and genes that are differentially expressed in various tissues of blueberry and different developmental stages. These genomic and transcriptome resources will be used for genome annotation and trait discovery, which collectively enhance our ability to breed for higher quality blueberries and other small fruits in future.

Resolving relationship and phylogeographic history of the *Nyssa sylvatica* complex using data from RAD-seq and species distribution modeling
Wenbin Zhou under the advisement of *Dr. Jenny Xiang*
Program: Plant & Microbial Biology

Nyssa sylvatica complex consists of three tree species, and four varieties occurring in eastern North America. Due to high morphological similarities and complexity of morphological variation, classification and delineation of taxa in the group have been difficult and the subject of multiple taxonomic debates. Here we employed data from RAD-seq to elucidate the genetic structure and phylogenetic relationships within the group and evaluate previous classification schemes. We also employed Species Distribution Modeling (SDM) to predict distribution ranges of the taxa to evaluate impacts of climatic changes and to gain insights into the refugia of trees in eastern North America. Results from Molecular

Variance Analysis (AMOVA), Structure, phylogenetic analyses using Maximum likelihood, Bayesian Inference and Splittree methods of RAD-seq data strongly supported a two-clade pattern, largely separating samples of *N. sylvatica* from those of *N. biflora-N. ursina*. Divergence time analysis with BEAST suggested the two clades diverged in the mid Miocene and the ancestor of the present trees of *N. sylvatica* in the Pliocene and that of *N. biflora-N. ursina* in the end of Miocene. Results from SDM predicted a smaller range in the southern part of the species present range of each clade during the Last Glacial Maximum (LGM) and northward expansion of ranges at interglacial periods, as well as a northward shift of the range in the future under the model of global warming. Our results support recognition of two species in the complex and an *N. ursina*-like ecotype within *N. biflora* due to its dwarf habit associated with frequent fire habitat. Our results further support movements of trees in eastern North America in responding to climatic changes and that RAD-seq data and a combination of population genomics and SDM are valuable in resolving relationship and biogeographic history of closely related species that are taxonomically difficult.

Preferential retention of homeologs from a single parental subgenome after polyploidy is shaped by functional interactions and dosage-based intrinsic selective constraint

Yue Hao under the advisement of *Dr. Gavin Conant*
Program: Bioinformatics

Polyploidy is seen as a driver of both evolutionary innovation and ecological success. When the contributors to the polyploidy are distinct species, it has been claimed that the post-polyploidy retention of genes often favors one of the two subgenomes. However, most of the analyses of this pattern of "biased fractionation" are limited to single or pairwise genome comparisons, potentially giving rise to artifactual estimates of such bias. Using our likelihood-based tool POInT (the Polyploid Orthology Inference Tool), we model the resolution of the At-α allopolyploidy event in the *Arabidopsis thaliana* and its relatives by phasing the syntenic regions of six plant genomes with respect to each other. We find statistically robust evidence for the existence of *biased fractionation*. More importantly, we show that this bias was not confined to the earliest phases of post-WGD evolution. We also show that a driver of this pattern of biased losses from one subgenome relative to the other is the co-retention of genes that are members of co-evolved functional complexes from the same parental genome. Meanwhile, to better understand the evolutionary pressures acting on the surviving duplicates from recent plant polyploidies (namely the At-α duplication and the more recent Brassica hexaploidy Br-α), we compare the strength and direction of the natural selection acting at the species and at the population level, we show that genes retained after polyploidy are intrinsically more constrained; even though genetic redundancy appears to relax the selective pressure to some degree (which is probably due to functional biases). Importantly, the intensified purifying selection acting on retained duplicates are still detectable in populations at the present day. We conclude that biased fractionation and preferential retention are long-term forces shaping the evolution of paleopolyploid genomes, suggesting that even yesterday's polyploids still have distinct evolutionary trajectories.

Population model for the pest *Homalodisca vitripennis* and its natural enemy *Gonatocerus ashmeadi*
Annabel Meade under the advisement of *Dr. H.T. banks*
Program: Biomathematics

The glassy-winged sharpshooter, *Homalodisca vitripennis*, is an invasive pest which presents a major economic threat to the grape industries in California by spreading a disease-causing bacteria, *Xylella fastidiosa*. Recently a common enemy of *H. vitripennis*, certain mymarid parasitoid species including *Gonatocerus ashmeadi* and *Gonatocerus morrilli*, have been studied to use in place of insecticides as a control method. We create a time and temperature dependent mathematical model to analyze data and answer the question: Does the implementation of *G. ashmeadi* as a biological control method cause a significant decrease in the population of *H. vitripennis*?

CRISPR Characterization and Diversity in *Lactobacillus fermentum*
Katelyn Brandt under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Lactic Acid Bacteria are gaining notoriety not only for their historical food benefits, but most recently for their rising potential in the microbiome, as probiotics, and development of CRISPR technology. Lactic Acid Bacteria are enriched for CRISPR systems, especially the Type II-A CRISPR-Cas9 system. With a growing desire to utilize CRISPR technologies for genome editing, it has become evident that the CRISPR toolbox will have to expand. Part of the Lactic Acid Bacteria complex, *Lactobacillus fermentum* is typically found in food products and human metagenome studies. It has recently increased in popularity due to its potential probiotic and antimicrobial effects. Due to these reasons, we investigated and characterized the CRISPR-Cas systems of *Lactobacillus fermentum*. Mining for CRISPR systems in 36 publicly available genomes, revealed a rich diversity of systems within this species. We determined the occurrence of CRISPR across genomes, and assigned Class and Type for each candidate. 78% of the genomes contain a putative CRISPR-Cas system, of which 71% contain a Type I CRISPR-Cas system and 57% encode a Type II CRISPR-Cas system. Our findings showed that while most systems did have a CRISPR locus, the number and type could differ greatly across genomes within this species. Next we looked at the similarity of Cas1 and Cas2 sequences between strains. Predictably, the groupings were consistent based on the type of system, rather than the strain. Next, we examined the spacer profile. Contrary to previous studies, we found that the spacer profile of each system varied more than expected, with no conserved ancestral spacers. Actually, only two groups shared common ancestors, while the others showed unique spacer acquisition tracks. Finally, we investigated the Cas9 from the model strain ATCC 14931, a distant homolog of Sth and Spy canonical Cas9 proteins sharing only 32% AA identify, as a potential orthogonal systems for the development of novel CRISPR-based molecular tools.

Synonymous Rate Variation Impacts Inferences of Positive Selection
Sadie Wisotsky under the advisement of *Dr. Spencer Muse*
Program: Bioinformatics

Most methods of inferring positive selection assume the site to site synonymous substitution rate remains constant. A growing body of literature does not support this

assumption from a biological point of view and previous work has shown the synonymous rate does vary from site to site at levels similar to the nonsynonymous rate. Here, we introduce a new method of gene wide episodic selection detection that incorporates site to site synonymous substitution rate variability (SRV), BUSTED+SRV. We compare BUSTED+SRV to the previous method BUSTED using both empirical data from the Selectome database and data simulated using the BUSTED+SRV framework. The empirical data shows the BUSTED+SRV is a better fit for the majority of the data. While the majority (74%) of the datasets from Selectome show no evidence of positive selection according to both methods, BUSTED and BUSTED+SRV find that 11% of datasets do show evidence of positive selection according to both methods. BUSTED finds an additional 11% of datasets have evidence of positive selection and BUSTED+SRV only finds an additional 3%. We also find that BUSTED has a high false positive rate when SRV is present. BUSTED+SRV, however, loses power at moderate levels of selection.

Phaser v.1: An integrated Bioinformatics suite towards chromosome scale haplotype-resolved genome assembly

Dr. Hamed Bostan
Plants for Human Health Institute, North Carolina State University, Kannapolis, NC

The interest in reconstructing phased chromosome haplotypes is increasing since it is crucial in many bioinformatics workflows such as genetic association studies and genomic imputation. Several tools have been developed to reconstruct phased haplotypes relying on the polymorphism information encoded on the target genome sequences. These algorithms produce haplotig blocks by phasing the adjacent polymorphic sites that can be reconstructed by overlapping reads. Since, multiple factors including the length of the reads, coverage and polymorphism rate, are limiting the development of chromosome scale haplotypes, these tools usually produce non-overlapping haplotypes within each chromosome and are not able to phase these blocks vs each other. *HAPCUT2* gets the advantage of proximity ligation (Hi-C) pair-end sequencing data to phase blocks vs each other up to chromosome level. Nonetheless, *HAPCUT2* phases only the SNPs excluding structural variants (SV, insertions/deletions), works only on diploid genomes and outputs only the phased SNP blocks and does not produce phased genomic sequences. An alternative approach recently presented by *CANU* developers use F1 parental resequencing data to cluster parent-specific k-mers for phasing F1 diploid genome. This algorithm works only on the diploid genomes and require that both parents are known, their resequencing data is available and that their genomes are diverged enough to enable effective separation of the parental haplotypes using k-mer specific sequences.

Here we present the current challenges and opportunities in reconstructing a chromosomal scale haplotype-resolved genome. We present *Phaser v.1*, a bioinformatics suite that can phase all types of the polymorphism sites (SNP and SV) on the genome, produce a phased pseudo-chromosome sequence in the output and that is independent from the divergence nature of the target genome. We also discuss the expansion of the *Phaserv.1* for phasing genomes with any ploidy level.

The rarity of data in physiological systems: perturbations for populations

Mitchel Colebank under the advisement of *Dr. Mette S. Olufsen*
Program: Biomathematics

Advances in engineering have paved the way for more abundant and more accurate sources of physical data. From a computational science perspective, this increase in available data allows mathematical models to be calibrated to measured data, leading to parameter estimation and realistic model prediction. Obtaining patient specific data can often be considered a rare occurrence in the biomedical sciences, as techniques for obtaining quantities of interest can be invasive. However, model predictions and scientific conclusions can be changed in a drastic manner if the data is subject to measurement uncertainty. Due to the lack of abundant medical measurement data, patient specific data is normally taken “as is” without accounting for population variability or medical device accuracy. This (brief) talk will focus on modeling physiological fluid flow in a network of blood vessels. In particular, I will highlight the importance of understanding how data is collected and show the outcomes that follow from small perturbations in medical image segmentation parameters. Additionally, I will highlight a technique for introducing uncertainty in 3D medical imaging segmentation, and how that uncertainty can help understand population variability through model outputs.

Modeling Inositol Phosphate Signaling in Arabidopsis thaliana

Cody Ellington under the advisement of *Dr. Cranos Williams* and *Dr. Joel Ducoste*
Program: Electrical Engineering

Myo-inositol phosphates (InsPs) are molecules that are critically important in a number of developmental, metabolic and signaling processes in eukaryotes. Given the immediate need to understand and manipulate plant bioenergy, the long-term goal of this project is to understand how InsP6, InsP7 and InsP8 convey signaling information within the cell. InsPs are synthesized in a complex pathway that uses various kinases to add phosphates at specific positions on the inositol ring. The number and position of phosphates on the inositol ring make up a type of chemical signaling language. The fully phosphorylated form, InsP6, accumulates to high levels in seeds and has been linked to regulation of stress and pathogen responses, phosphate sensing, and mRNA processing and export in vegetative tissues. A kinetic model is being developed which will help provide insight into how other InsPs in the signaling pathway impact the concentration of InsP6 in the cell. This simulation-based model utilizes MATLAB and ordinary differential equations defined by Michalis-Menten rate equations. State of the art modeling techniques allow for non-intuitive relationships to be investigated between metabolites in the pathway. Metabolic Control Analysis is used to determine the influence of specific kinetic parameters on the projected steady state metabolite concentrations. Understanding which parameters are more influential can help guide biochemical researchers as they utilize their resources to build kinetic profiles for the enzymes in this signaling pathway.

CTN: an R package for Variant Calling Format data

Tao Jiang under the advisement of *Dr. Alison Motsinger-Reif*
Program: Bioinformatics

Next generation sequencing is a new technique that can generate informative sequencing data efficiently and accurately. Variant calling format (VCF) files are important output files for next generation sequencing. Unfortunately, there are not many R functions that can deal with VCF files currently. The R package CTN contains functions which can read in variant calling format files; show the information contained in the file; detect sample contamination from another contributor; and estimate tumor-normal ratio for any single cancer sample.

Parameter subset selection and identifiability for a baroreflex model

E. Benjamin Randall under the advisement of *Dr. Mette S. Olufsen*
Program: Applied Mathematics

Mathematical models of biological phenomena tend to be large systems with many parameters. Some model parameters can be motivated empirically, while others are unknown. Determining the values for these unknown parameters is not trivial, especially since some parameters cannot be identified individually. Furthermore, optimization (fitting a model output to available data) can only be performed effectively on a subset of parameters to ensure (a) biological relevance and (b) avoidance of "overfitting" and interactions of correlated parameters. Subset selection and identifiability of parameter subsets is an open area of research. This talk will focus on the methods used to find an identifiable subset of parameters for a mathematical model of the baroreceptor reflex (baroreflex). We will discuss the importance of performing a sensitivity analysis on the model and construction of a covariance matrix that elucidates parameter correlations.

Differential Gene Expression of 'O'Neal' Blueberry Floral Buds in Response to Freeze Treatment and Recovery Periods

Lauren E. Redpath under the advisement of *Hamid Ashrafi*
Program: Horticultural Science

Cold hardiness and chill hour requirement in blueberries is a function of germplasm composition. Southern highbush blueberries (SHB) are less cold tolerant than northern highbush blueberries and have a lower chill hour requirement, causing buds to deacclimate and break earlier. Deacclimation and budswell heightens susceptibility to late spring freezes, a recurring event in the southeast of the U.S. The objective of this study was to determine the differentially expressed genes of SHB cv. 'O'Neal' floral buds prior to budswell and post-budbreak. Additional treatment conditions included temperature conditions of non-freezing (4 °C) or freezing conditions (-12 °C), as well as bud recovery period of either 1 day or 1 week after temperature treatments. Treated buds were flash frozen in liquid nitrogen. A total of 24 stranded mRNA-Seq libraries (8 treatments x 3 biological replications) were paired end sequenced to generate 160 Gbp of raw reads. After trimming ~150 Gbp of data was retained. CLC genomics workbench (V.11) was used to make a transcriptome assembly with 273,702 contigs, and with N50=916. BLAST2GO (v. 5) was used to functionally annotate all assembly contigs. CLC Genomics WB (V.11) was used to map the trimmed reads to the assembly and to identify differentially expressed genes (DEGs) between tissue, temperature, and recovery periods. Identification of DEGs produced 8 upregulated of 11 DEGs between tissue types ($\log_2$ fold change >|2|; $p \leq 0.05$), 12 upregulated of 34 DEGs between the temperature treatments, and 83 upregulated of 126 DEGs between recovery periods. Temperature expression had greater effect on downregulation of genes (65%), than either tissue type or recovery. Downregulated recovery contigs were largely unannotated in either BLAST or Interpro Scan. More contigs were downregulated in response to temperature treatment in floral buds at tight cluster than those at budswell; similarly, more contigs were upregulated between tissue types in non-frozen buds with 1-day recovery than buds treated to freezing temperatures and 1-day recovery. Overall, more contigs were differentially expressed in recovery compared to tissue type or temperature treatments. Future work involves mapping and analyzing the assembly against the 'O'Neal' genome and its PacBio Iso-Seq data.

Modeling CRISPR gene drives and reversals

Michael Vella under the advisement of *Alun Lloyd*
Program: Biomathematics

A gene drive biases inheritance of a gene so that it increases in frequency within a population even when the gene confers no fitness benefit. There has been renewed interest in environmental releases of engineered gene drives due to recent proof of principle experiments with the CRISPR-Cas system as a drive mechanism. Release of modified organisms, however, is controversial, especially when the drive mechanism could theoretically alter all individuals of a species. Thus, it is desirable to have countermeasures to reverse a drive if a problem arises. Several genetic mechanisms for limiting or eliminating gene drives have been proposed and/or developed, including reversal drives. While predictions about efficacy of these mechanisms have been optimistic, there has been little detailed analyses of their expected dynamics. We develop a discrete time model for population genetics of a drive and proposed genetic countermeasures. For some parameter values, the model predicts unexpected behavior including polymorphic equilibria and oscillatory dynamics. The timing and number of released individuals containing a genetic countermeasure can substantially impact outcomes. The choice among countermeasures by researchers and regulators will depend on specific goals and population parameters of target populations.

A reaction-diffusion model explains amplification of the phospholipase C/ protein kinase C pathway in fibroblast chemotaxis

Jamie Nosbich under the advisement of *Dr. Jason Haugh*
Program: Biomathematics

In fibroblasts responding to gradients of platelet-derived growth factor (PDGF), an important chemoattractant in development and wound healing, signaling through the phospholipase C (PLC)/protein kinase C (PKC) pathway proved necessary for chemotaxis. PKC is activated through its binding to the lipid second messenger diacylglycerol (DAG), which is formed from hydrolysis of phosphatidylinositol (4,5)-bisphosphate (PIP_2) by PLC. Strikingly, in fibroblasts exposed to a shallow PDGF gradient, the density of DAG in the plasma membrane is focally enriched at the up-gradient leading edge, suggesting an internal amplification mechanism that has yet to be explored. We have developed and analyzed multiple mechanistic, reaction-diffusion models of the PLC/PKC signaling pathway activated in a PDGF gradient. The models include the major proteins (PDGF receptor, PLC, and PKC) and lipids (PIP_2 and DAG) in the canonical pathway, as well as other signaling molecules that we implicate in various positive feedback loops. Model simulations suggest that the synergy of at least two of the putative feedback loops is needed to drive order-of-magnitude enrichment in a shallow PDGF gradient. Experiments will need to be performed, in concert with refinement of our modeling framework, to validate the source(s) of nonlinearity in the signaling pathway. Current work involves exploring the effects of the cell's geometry on the polarization of the signaling network and assessing the effects of stochasticity on the performance of this system. In the future, this model will be linked to models describing the organization of the actin cytoskeleton and directionality of cell migration for a more comprehensive understanding of how fibroblast chemotaxis proceeds during physiological processes such as wound healing.

After the Honeymoon, the Divorce: Unexpected Outcomes of Disease Control Measures

Brandon Hollingsworth under the advisement of *Alun Lloyd*
Program: Biomathematics

For many endemic human diseases vaccinations can be difficult to develop and prohibitively expensive to employ. Control of these diseases often works by minimizing transmission of the infection; e.g. minimizing contact with infectious individuals, treatments that cause people to be less infective, or targeting some vector of the disease. These control programs require ongoing effort and are often subject to sudden interruptions. While the dynamics of disease control programs are well studied, the period following the end of a control is not well understood. To address this, we simulated the dynamics of a disease system in which a successful control is suddenly stopped and compared the results to what would be expected in the endemic setting. We show the non-intuitive result that over time, there are periods in which the population experiencing the control measure will see more cases of the disease than if no control was implemented. We show that this result – which we term the divorce effect – is caused by the buildup of susceptible individuals in the population during the application of certain control measures, leading to spikes of disease outbreaks after control. Further, we find this effect is present in many disease systems that allow for a buildup of susceptible individuals. Our current work is focused on better understanding the conditions that allow for this divorce effect and to find control plans that would be able to mitigate the risk of such an effect. These non-intuitive post-control disease dynamics, their causes, and possible solutions are becoming an ever more pertinent subject as many of the disease controls currently employed are beginning to fail due, amongst other things, to the rise of antibiotic and insecticide resistance.

Immunosuppressant treatment dynamics in renal transplant recipients: An iterative modeling approach.

Neha Murad under the advisement of *Dr. H.T. Banks*
Program: Biomathematics

Finding the optimal balance between over-suppression and under-suppression of the immune response is difficult to achieve in renal transplant patients, all of whom require lifelong immunosuppression. Our ultimate goal is to apply control theory to adaptively predict the optimal amount of immunosuppression; however, we first need to formulate a biologically realistic model. The process of quantitatively modeling biological processes is iterative and often leads to new insights with every iteration. We illustrate this iterative process of modeling for renal transplant recipients infected by BK virus. We analyze and improve on the current mathematical model by modifying it to be more biologically realistic and amenable for designing an adaptive treatment strategy.

Characterizing the Chemical Space of Kinase Inhibitors Using Molecular Descriptors Computed from Molecular Dynamics Trajectories

Jeremy R. Ash under the advisement of *Dr. Denis Fourches*
Program: Bioinformatics

2D and 3D descriptors are typically used to characterize the structure of chemicals and then establish Quantitative Structure-Activity Relationships (QSAR). In a previous study, we showed that ranking a set of small molecule kinase inhibitors could be very challenging for certain targets like the ERK2 kinase. In order to build more reliable models for this challenging datasets of relevance for lead optimization, we have developed a new workflow that encompasses (i) the structure-based docking of a series of known inhibitors in the binding site of a particular kinase, (ii) the independent molecular dynamics (MD) simulations of each protein-ligand complex (15 ns, NVT, 300K, TIP3P, 1fs), (iii) the computation of novel "MD descriptors", which measure the variations of ligands' 3D shape over the course of MD simulations as well as their dynamic protein-ligand (PL) interactions. We have constructed cross-validated QSAR models using several machine learning methods and various sets of MD-based descriptors. Some of these models are highly interpretable, identifying features of ERK2 inhibitor MD conformation shape and PL interactions that are key determinants of activity. These features may help medicinal chemists design and identify new analogs likely to be more potent. These results demonstrate how MD simulations can provide an additional layer of information on ligands' dynamic characteristics that are of importance for further focused structural optimization.

Synergistic Chemotherapy Drug Response is a Genetic Trait in Lymphoblastoid Cell Lines

Kyle Roell under the advisement of *Dr. Alison Motsinger-Reif and Dr. David Reif*
Program: Bioinformatics

Combination therapy is quite common in chemotherapy nowadays since drugs work synergistically. According to studies, the LCL model has proven to be highly successful in understanding the genetic etiology of drugs' responses to cancer. We demonstrate that synergy occurs in LCLs across a wide range of drug combinations. LCLs have been commonly employed in association mapping in drug response. In order to determine if this would be a useful model for understanding the etiology of synergy, we evaluate whether variation in synergy is heritable. Most importantly, we demonstrate that there is a substantial heritable component to variation in does response in LCLs. This demonstration supports the premise of expanding the use of LCL model to perform association mapping for combination therapies.

Resolving Alternative Splicing Patterns in Rhesus Macaque Transcriptomes using Full-Length Transcriptome Sequencing Analysis

Hayden Brochu under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

Motivation: Complex alternative splicing is one of the key mechanisms for transcriptional regulation. Despite the importance of rhesus macaque in biomedical research, the annotation of its alternative splicing remains largely incomplete. It can be challenging to resolve this transcriptional complexity with only short-read sequencing data, as individual loci can generate large numbers of alternatively spliced transcripts. Here we sought to utilize long-read sequencing to improve isoform annotation for rhesus macaque, one of the most widely used non-human primate models.

Results: Using PacBio Iso-Seq, we generated over 2.8 million transcript sequencing reads (Circular Consensus Sequence reads, CCSs), ranging from 300 to 45,549 nucleotides, from four different macaque tissues. We obtained 74,006 high quality full-length transcripts, 96% of which were aligned accurately to the current rhesus macaque genome assembly. Among detected coding genes, 50% had at least 1 novel isoform when compared to the reference macaque annotation. We showed that multiple mechanisms, including intron retention and skipping, alternative transcription start and termination sites, fusion transcripts, and macaque-specific transcripts, contributed to these novel splicing patterns. With this full-length transcriptome sequencing analysis, we present major improvements to the rhesus macaque genome annotation and demonstrate unique benefits of long-read sequencing in elucidating transcriptional complexity.

Blue Crab Spawning Stock Index using a VAST model

Amanda Reeder under the advisement of *Dr. Kevin Gross* and *Dr. Dave Eggleston*
Program: Biomathematics

Blue crab (*Callinectes sapidus*) fisheries are one of the most valuable crab fisheries in the world and the most valuable fishery in North Carolina in terms of weight and income. They are also of high importance to the estuarine community structure as they are the dominant predator in various benthic environments. Due to the high importance of the blue crabs, it is imperative for fishery management to know an accurate spawning stock index. Currently, the method of estimating a spawning stock index is to randomly sample different locations and average the biomass of mature females and average across all samples. The issue with this method is that blue crabs are highly mobile both during their life cycle and in response to rapid environmental change, such as a hurricane. For a more accurate spawning stock index, environmental covariates such as temperature and salinity need to be taken into account. Another issue is that the data collected has many sites that have zero captures of mature females while other sites have high catches of mature females. This is defined as zero-inflated data and it is dealt with by using a delta generalized linear model. It is proposed to use a vector-autoregressive spatial-temporal (VAST) model because it can handle the environmental covariates (salinity and temperature) as well as zero-inflated data.

Developmental toxicity of engineered nanomaterials in zebrafish embryos
Kimberly To under the advisement of *Dr. David Reif*
Program: Bioinformatics

Nanomaterials are defined as particles with at least one dimension on the nanoscale (1-100nm). Although their toxicity is not yet fully understood, nanomaterials present themselves in a wide variety of consumer products, such as sunscreens or laser printers. Nanomaterials have been proposed as drug delivery systems and as improvements to in vivo molecular imaging techniques. Despite the promising applications of nanomaterials, understanding of their health impact remains elusive. Further, because nanomaterials are already present in consumer products, there is a sense of urgency in understanding nanomaterial toxicity as a preventative measure. We generated morphological data from zebrafish embryos exposed to varying types and concentrations of engineered nanoparticles to explore the effects of physicochemical characteristics on morphological abnormalities.

Identification of Patterns Among Differentially Expressed Temporal Profiles by Hidden Markov Model Based Computational Architecture
Haonan Tong under the advisement of *Dr. Cranos Williams*
Program: Electrical and Computer Engineering

Plants react to stress by regulation of the expression of thousands of genes, known as gene regulatory networks. One way to gain understanding of these networks is to use gene expression data and explore differences in expression to find genes that regulate and genes that are regulated. Differentially expressed genes (DEGs) under a specific stress are identified, and this subset can then be analyzed to unravel genetic regulatory mechanisms. It is generally accepted that DEGs with similar expression patterns are co-expressed or co-regulated. For this reason, identifying genes with similar expression patterns is key to understanding the underlying regulatory mechanisms. This understanding will aid in the development of methods and tools to improve modern agriculture. In this talk, I am introducing a HMM-based clustering computational architecture to identify genes with similar expression patterns based on RNAseq data extracted from etiolated Arabidopsis thaliana seedlings .

CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas systems
Alexandra Crawley under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

CRISPR-Cas adaptive immune systems of bacteria and archaea have catapulted into the scientific spotlight as genome editing tools. To aid researchers in the field, we have developed an automated pipeline, named CRISPRdisco (CRISPR discovery), to identify CRISPR repeats and cas genes in genome assemblies, determine Type and Subtype, and describe system completeness. All six major and 23 currently recognized subtypes and novel putative V-U types are detected. Here, we use the pipeline to identify and classify putative CRISPR-Cas systems in 2,777 complete genomes from the NCBI RefSeq database.

This allows comparison to previous publications and investigation of the occurrence and size of CRISPR-Cas systems. Software available at http://github.com/crisprlab/CRISPRdisco provides reproducible, standardized, accessible, transparent, and high-throughput analysis methods available to all researchers in and beyond the CRISPR-Cas research community. This tool opens new avenues to enable classification within a complex nomenclature and provides analytical methods in a field which has evolved rapidly.

Uncertainty Analysis and Parameter Estimation in Dynamic Model of Gene Regulatory Network

Samiul Haque under the advisement of *Dr. Cranos Williams*
Program: Electrical and Computer Engineering

Mathematical model describing the regulatory relationship between different genes is critical for genetically engineered plants and other organisms. Recent studies have proposed different strategies to develop dynamic models for describing the genetic regulation. However, no one method address all the critical issues that we face during parameter estimation for these models. In this work, we address four common problems often overlooked during model development for GRN; different data sources, lack of data to estimate the model parameters, non-influential parameters in the model, and choosing correct experimental design. We demonstrate through an example (GRN model of Arabidopsis thaliana under iron deprivation), that the use of identifiability analysis, sensitivity analysis, and parameter estimation in conjunction can yield correct model predictions. We also address the heteroscedasticity in experimental data. We employed profile likelihood method for identifiability analysis and used eFAST and PRCC for sensitivity analysis. We incorporated separate experimental data in the form of prior distributions in an MCMC based Bayesian algorithm DREAM. We found that correct prediction can be made from sparse and noisy data by using the proposed strategy. As a result of this research, it is recommended that this strategy could be used in future model-based systems-biology research.

riboStreamR: A Web Application for Quality Control, Analysis, and Visualization of Ribo-Seq Data

Patrick Perkins under the advisement of *Dr. Steffen Heber*
Program: Bioinformatics

Ribo-seq is a popular technique for studying translation and its regulation. Various software tools for data preprocessing, quality assessment, analysis, and visualization of Ribo-seq data have been developed. However, many of them are inaccessible to users without a thorough practical knowledge of software applications, and the use of multiple different tools is often necessary to complete a full analysis. Here, we present riboStreamR, a comprehensive Ribo-seq quality control (QC) platform in the form of an R Shiny web application. RiboStreamR provides visualization and analysis tools for various Ribo-seq QC metrics, including read length distribution, read periodicity, and translational efficiency. The platform's environment is centered on providing a streamlined experience, and includes numerous options for graphical customization and report generation. In practice, Ribo-seq data analysis can be sensitive to data quality issues such as read length variation, low read periodicities, and contaminations with ribosomal and transfer RNA. What constitutes ‘high quality’ data is often unclear, and therefore one of our primary goals is to develop novel functionality to automatically highlight quality issues and anomalies in the data. This NSF-supported project is performed in collaboration with Jose Alonso, Anna Stepanova, Serina Mazzoni-Putman, and Cranos Williams.

An electronic version of this document is available at:
https://gsbmasymposium2018.wordpress.ncsu.edu/program/