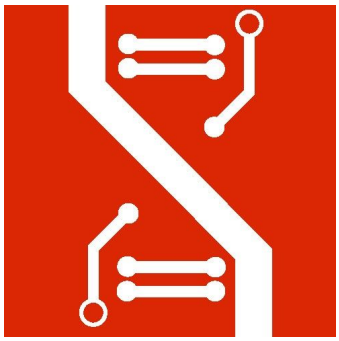


BIOMATH

NC STATE UNIVERSITY

Genomic Sciences & Biomathematics Symposium

Spring 2020

Saturday, February 15th

9:30AM – 5:00PM

James B Hunt Jr. Library – Room 4106

We would like to acknowledge the NC State Graduate Student Association for their generous financial support through the block grant funding program. We also thank Spencer Muse, Dana Ripperton, Jenni Wilson, and Madge Waterman for kindly assisting the Genomic Sciences and Biomathematics GSAs in coordinating this event.

Schedule of Events

9:30AM – 10:00AM	**Bagels, Coffee, & Check-in**
10:00AM – 10:15AM	**Welcome**
	Hayden Brochu, Genomic Sciences GSA President
10:15AM – 11:15AM	**Talks - Bioinformatics Focus**
	Yue Hao, Caizhi (David) Huang, Evan Walsh, Hayden Brochu
11:15AM - 12:15PM	**Morning Keynote Speaker**
	Dr. Paul Magwene
12:15PM – 1:45PM	**Lunch & Poster Session**
	Matthew Nethery, Madison Moore, Avery Roberts, Tao Jiang, Kuncheng Song, Montana Knight, Echo Pan, Dani Joseph
1:45PM – 2:25PM	**Lightning Talks**
	Lenora Kepler, Vaishnavi Venkat, Jun Ma, Ashley Schoonmaker, Ian Huntress, Jackson Parker
2:25PM – 2:55PM	**Talks - Biomathematics Focus**
	Annabel Meade, Maureiq Ojwang'
2:55PM – 3:15PM	**Coffee Break & Networking**
3:15PM – 4:00PM	**Talks - Bioinformatics & Functional Genomics Focus**
	Bryan Ting, Will Kohlway, Yueyang Huang
4:00PM – 5:00PM	**Afternoon Keynote Speaker**
	Dr. Nicolas Buchler
5:00PM – 5:10PM	**Closing Remarks**
	Evan Curcio, Biomathematics GSA President

Morning Keynote Speaker | 11:15AM – 12:15PM

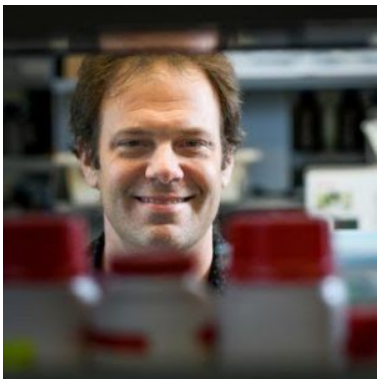
Genetic Architecture, Gene Network Variation, and Evolutionary Hotspots in Yeast

Paul Magwene, Ph.D.
Associate Professor, Department of Biology
Director, Graduate Program in Computational Biology & Bioinformatics
Duke University

I will discuss a growing body of work, based on QTL mapping and comparative genetic analyses, that suggests that natural genetic variation that affects Ras-cAMP-PKA signaling is a major contributor to phenotypic diversity within and between yeasts species of the genus Saccharomyces. Studies employing experimental evolution also indicate that mutations that affect Ras-cAMP-PKA signaling are favored during adaptation to novel nutrient environments. Recent studies from my group suggest that the central importance of this pathway in contributing to phenotypic variation extends to many other fungal clades.

Afternoon Keynote Speaker | 4:00PM – 5:00PM

Insights into the evolution of biological oscillators using genomics and mathematical modeling



Nicolas Buchler, Ph.D.
Associate Professor, Department of Molecular Biomedical Sciences
North Carolina State University

Few processes in biology are as evolutionarily constrained as the cell division cycle, yet the eukaryotic cell cycle network in Fungi appears to have been rewired by a viral protein. In the first half of my talk, I will show how my lab uses genomics to understand how a horizontally-transferred viral protein (SBF) integrated into the G1/S regulatory network of a fungal ancestor and eventually replaced the original transcription factor (E2F) in the ancestor of most Fungi without disrupting the cell cycle. In the second half of my talk, I will show how we use mathematical modeling of gene network dynamics to understand the evolution of circadian clocks. Gene expression is a biochemical process, where stochastic binding and unbinding events naturally generate fluctuations and cell-to-cell variability in gene dynamics. These fluctuations typically have destructive consequences for proper biological dynamics and function (e.g. loss of timing and synchrony in biological oscillators). Using computer simulations, I will show how noise counter-intuitively accelerates the evolution of a biological oscillator and, thus, could impart a benefit to living organisms.

Talks - Bioinformatics Focus
10:15AM – 11:15AM

Ancient whole genome duplications in early vertebrates
Yue Hao under the advisement of *Dr. Gavin Conant*
Program: Bioinformatics

Two rounds of whole genome duplications (2R-WGD) occurred in ancient vertebrate evolutionary history, followed by intensive genomic reshuffling and gene loss. Among the surviving duplicated genes from this polyploidy event (homeologs), many are sensitive to stoichiometric dosage. These duplicates therefore are associated with diseases in mammals, such as heart disease and cancer, when their dosage is perturbed. We will statistically model the pattern of gene loss/retention after 2R-WGD using a likelihood-based tool, POInT (the Polyploid Orthology Inference Tool), hoping to identify the preferentially retained homeologs after 2R-WGD, and to find out whether the remnants after the polyploidy could be the key entities involved in biological functions, disease processes and evolutionary innovations.

Structure-guided Microbiome OTU-specific Association Test
Caizhi (David) Huang under the advisement of *Dr. Jung-Ying Tzeng and Dr. Ben Callahan*
Program: Bioinformatics

Background: The human microbiome is currently known to be associated with human health. Increasing microbiome research projects such as Integrative Human Microbiome Project (iHMP) produce an enormous amount of microbiome sequencing data. With these data, identification of certain species or operational taxonomic units (OTU) that are associated with human phenotypes can help to understand the mechanism of diseases. However, due to the sparse and large OTU number of the data, traditional association test often suffers from low power. One unique characteristic of microbiome data is that OTUs are evolutionarily related. It has been noted that phylogenetically related species are expected to respond to the environment perturbations in similar manners and statistics methods incorporating phylogeny information tend to be more powerful.
Method: Here we propose a structure-guided microbiome association test (SMAT) under the framework of distance-based kernel machine regression. By borrowing information from neighboring OTUs from the phylogeny tree, SMAT improves the statistical power to detect the associated microbiome features at OTU level.
Result: We illustrate our method through the simulated data and real data. Power and type I error estimates of SMAT are compared with existing OTU-level association test. ROC curve and AUC are used to demonstrate the performance. We have showed that while the phylogenetic tree is informative, SMAT has best performance with largest AUC. Under the scenario that phylogenetic tree is not informative, SMAT can achieve the similar AUC with the other methods. In the real data, we show SMAT can identify more vaccine efficacy associated microbial.

Exploratory analysis of single-cell RNA sequencing in a developing mouse brain
Evan Walsh under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

Background. Sequencing mRNAs at the single-cell level (scRNAseq) enables the discovery of specific changes and signatures of heterogeneous populations of cells which cannot be detected by bulk mRNA sequencing. However, limitations in mRNA capture and reverse transcription in droplet-based scRNAseq technologies create a "dropout" phenomenon where genes with low cellular expression often have read counts of zero. The sparsity of scRNAseq data resulting from dropout has made studying gene expression changes a challenge for many.
Materials & Methods. To begin to understand scRNAseq data processing, we sequenced over 18,000 cells corresponding to four single-cell barcoded cDNA libraries as part of a study investigating the role that Epidermal Growth Factor Receptor (EGFR) has on gliogenesis in the developing mouse brain. After sequencing, we projected cells onto a dimensionally reduced space using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) to infer cell-type profiles. To understand the role of EGFR at the transcript level, we have begun to explore the effect of dropout in scRNAseq differential expression tests.
Results. Here we show how conditional deletion of EGFR reduces the number of developing oligodendrocytes during embryogenesis. In addition, we characterize two distinct clusters of developing oligodendrocytes by the pattern of EGFR expression. Finally, we demonstrate how current single-cell differential expression tests create false negatives for low abundant genes affected by dropout. To fairly measure gene expression differences between cell types, we propose a stratified method of differential expression to give more power to these genes expressed at low levels.

Microbial feature selection using a greedy pairwise balance search
Hayden Brochu under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

Background. Differential abundance analysis of microbial communities is a challenging task due to the relative nature and sparsity of the data. Treatment of microbial marker gene data (e.g. 16S rRNA) as compositional has become a promising approach for effective normalization and analysis. Many recent developments use isometric log ratios (balances) to evaluate differences in microbial communities. However, these balances have poor interpretability and do not ensure that all components are associated with the phenotype of interest.
Materials & Methods. Here, we develop a novel compositional strategy for identifying bacteria predictive of a phenotype. We scan microbial features by computing all pairwise balances and assessing how well they segregate samples based on a phenotypic marker. After enumerating pairwise balances above a chosen significance threshold, we identify top overrepresented features within those balances. Next, we normalize these using a set of features enriched among low variance pairwise balances.

Results. We illustrate this approach using 16S rRNA microbiome data from a study of rhesus CMV vector based SIV vaccines, where ~50% of animals successfully controlled infection after SIV challenge. Beginning with 2,400 16S rRNA amplicon sequence variants (ASVs) that represent potential bacterial species/strains, we applied our feature selection procedure using consensus nested cross validation to identify features associated with protection status. We identified a robust set of pre-challenge microbial features that clearly distinguish animals based on protection outcome. This method produces interpretable results in compositional space and can be combined with common cross validation procedures to build predictive models. In the future, this method could be applied to other types of compositional data.

Poster Presentations
12:15PM – 1:45PM

Expanding the bioinformatic toolkit for rapid detection and classification of CRISPR-Cas systems

Matthew Nethery under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Recent advances in computing power have led to extensive accumulation of genomic data, prompting the development of next-generation in silico tools centered around large data set mining and high-throughput analyses. These new tools, coupled with the widespread accessibility and depth of public genome repositories, provide us with valuable data for improved CRISPR-Cas characterization and facilitate efforts in discovery of novel CRISPR-Cas systems. Here, we illustrate the use of a rapid, high-throughput CRISPR-Cas classification tool to systematically identify CRISPR repeat-spacer arrays, flanking cas genes, and canonical classification types in CRISPR-Cas systems across all available prokaryotic genomes in the NCBI RefSeq database. These results were subsequently converted into a local searchable database. During this process, 80,762 CRISPR loci were identified across 89,427 bacterial, archaeal, and viral genomes. Of these loci, only ~59% could be attributed a canonical type through identification of flanking cas genes. Systems that could not be typed with flanking cas genes were successfully classified using Cas proteins identified at a whole-genome level. In total, 398 of 630 archaeal genomes and 33,741 of 81,206 bacterial genomes were found to contain classifiable CRISPR-cas systems. Furthermore, we developed several bioinformatic pipelines on top of this database to expedite identification of self-targeting spacers, to describe conserved CRISPR-linked genes, and developed methods to distinguish between CRISPR-Cas system types based solely on repeat characteristics at each locus, exhibiting > ~96% accuracy for Type I and Type II systems with basic machine learning approaches. Indeed, the growing abundance of genomic data will continue to stimulate the evolution of new bioinformatic tools and will constantly test and expand our knowledge of CRISPR-Cas systems, further refining existing features and classifications as well as promoting the discovery of novel systems.

Developing a strategy for assembling a microbial synthetic community
Madison Moore under the advisement of *Dr. Jose Bruno-Barcena*
Program: Functional Genomics

Assembling synthetic communities for experimental normalization and testing may provide insights into the complex nature of interactions occurring between diverse microbes. Every discipline is choosing to assemble microbiota adapted to specific environments such as the human gut. Afterwards, bioinformatics and other functional genomic approaches facilitate the elimination of presence absence and relative strain abundances as the first step to evaluate in depth interactions and unknown ecological rules. Thus, to ensure reproducible and reliable data for analysis, maintaining quality control measures during cell culturing and preservation is of the utmost importance. Here, master cell banks (MCB) were established for multiple gram-positive bacterial strains by growing and harvesting cells in the late logarithmic phase of growth to maximize the number of cells concentrated. Multiple working cell banks (WCB) were then created from the MCB to later obtain reliable inoculum pools that can be exposed to detailed and thorough analysis including DNA extraction and subsequent 16S sequencing for taxonomic identification. In doing so, the data produced here can be replicated and deemed reliable due to the samples being established from a single source. A downstream synthetic community will be assembled by multiplexing only microbial strains that have been fully sequenced to validate the structure of the community.

Repurposing endogenous CRISPR-Cas systems in *Lactobacillus*
Avery Roberts under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Lactobacillus is a genus of Gram-positive, rod-shaped lactic acid bacteria that inhabit a diverse range of environments, including the human gastrointestinal tract and dairy products. Lactobacilli are important players in food manufacturing settings as starter cultures that drive dairy and other food fermentation, where they inhibit the growth of spoilage agents and contribute to product nutrition, flavor, and texture. CRISPR-Cas systems are present within approximately 63% of *Lactobacillus* strains and act as adaptive immune systems that defend against foreign nucleic acids. Endogenous CRISPR-Cas systems can be functionally characterized and then harnessed for specific genome editing applications. Currently, only endogenous type I and type II CRISPR-Cas systems have been repurposed in *Lactobacillus*. Here we provide an overview of the endogenous CRISPR-Cas system repurposing process as well as experimental data from select *Lactobacillus* species.

Evaluation of the Efficacy of K-mers in Host Trait Prediction
Kuncheng Song under the advisement of *Dr. Fred Wright and Dr. Yihui Zhou*
Program: Bioinformatics

In recent years, 16s rRNA amplicon sequencing data have been widely used regarding the ecology of the intestinal microbiome. In the analysis of host trait phenotypes such as age,

diabetes, obesity, etc., the existing literature mainly focuses on the analysis of a de novo clustered OTUs. There is a lack of investigation into how different OTU grouping methods could potentially impact the prediction of the host phenotypes.
The goal of our study is to get the insights of the best host trait prediction under different clustering methods, including open reference, close reference, de novo, and ASV (amplicon sequencing variants). In addition to the species level count table, we constructed count tables of higher taxonomic order. We also compare the prediction accuracy between the normalized count data with the raw abundance. A variety of standard machine learning methodologies are applied to the host trait prediction by using a Crohn's Disease 16s rRNA dataset.

Higher Criticism Tuned Sparse Group Lasso for Weak and Sparse Signals in GWAS
Tao Jiang under the advisement of *Dr. Alison Motsinger-Reif*
Program: Bioinformatics

In the current study, we propose an extension of least absolute shrinkage and selection operator (LASSO) regression to address variable selection and modeling when sample sizes are limited compared to the data dimension. Our method is motivated by high-throughput biological data, such as genome-wide association studies (GWAS). We propose a new upper bound of the regularization parameter λ in sparse group LASSO based on an estimated lower bound of the proportion of false null hypotheses with confidence ($1-\alpha$). The bound is estimated by applying the empirical distribution of dependent or independent *p*-values from single marker/variable analysis, where a second-level significance testing, the higher criticism statistic is used. An upper bound of tuning parameter in LASSO, λ, is decided corresponding to the lower bound of the proportion of false null hypotheses. Thus, the tuning range is narrow since the upper bound of λ is lower. The final decision of non-zero estimates (e.g., significant loci in GWAS) will contain more variables so that the power of modified GWAS is higher than or equal to the original sparse group Lasso. Different correlation levels among variables in true regression models are also studied. We demonstrate the performance of our method using both simulation experiments and a real data application in lipid trait genetics from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial.

How to Grow Plants in Space: a community project with Exploris elementary school
Montana Knight under the advisement of *Dr. Dahlia Nielsen and Dr. Colleen Doherty*
Program: Bioinformatics

Hands-on science education has tremendous benefits on early childhood development and can influence lifelong scientific interest. According to a recent study, the more time spent studying and doing science in those early years relates to later scientific achievement (Curran & Kitchin). As part of North Carolina State University's science community; it is up to us to go into Raleigh to inform and excite our local area about the research we are doing. That is what we chose to do in the Fall of 2019 at Exploris Elementary School. Related to our current research, we chose to teach 11 second and third graders about growing plants in space. Students were divided into three groups to grow a variety of plants in different

light conditions. Through the five week program, they monitored the plants and recorded their findings while learning about space, plants, and scientific methodology. At the end of the program students created blue prints for their own lunar growth chamber by incorporating everything they had learned about what plants would need while in space. For us, the project was a fun way to reach out to the community and show young students the kind of exciting research done at NCSU. For the students of Exploris, the project planted a seed for scientific interest that they can carry with them through their education.

Host and body site-specific adaptation of *Lactobacillus crispatus* genomes
Echo Pan under the advisement of *Dr. Rodolphe Barrangou*
Program: Functional Genomics

Lactobacillus crispatus is a common inhabitant of both healthy poultry gut and human vaginal tract. A higher risk of developing infectious diseases has been associated with the absence of this species. In this study, we performed comparative genomic analysis on 105 *L. crispatus* genomes isolated from a variety of ecological niches including the human vaginal tract, human gut, chicken gut and turkey gut to shed light on the genetic and functional features that drive evolution and adaptation of this important species. We performed *in silico* analyses to identify the pan and core genome of *L. crispatus*, and to reveal the genomic differences and similarities associated with their origins of isolation. Our results demonstrated that, although a significant portion of the genomic content is conserved, human and poultry *L. crispatus* isolates evolved to encompass different genomic features (e.g. carbohydrate usage, CRISPR-Cas immune systems, prophage occurrence) in order to thrive in different environmental niches. We also observed that chicken and turkey *L. crispatus* isolates can be differentiated based on their genomic information, suggesting significant differences may exist between these two poultry gut niches. These results provide insights into host and niche-specific adaptation patterns in species of human and animal importance.

A quarter Century of re-annotation
Dani Joseph under the advisement of *Dr. David Bird*
Program: Functional Genomics

Reliable gene annotation is prerequisite to deducing gene function from sequencing data. In the past, the relative small number of entries in GenBank also served to limit gene discovery. This was certainly the case when we attempted to isolate and annotate the transcripts specifically expressed in specialized feeding cells called giant cells (GC). Subtractive cDNA libraries were made from individual single, hand-dissected GC (Phytopath. 84: 299-303, **1994**). Approximately 200 clones were sequenced (MPMI: 964-967, **1994**), revealing fifty-seven RNA sequences from *Solanum lycopersicum* (tomato) able to be identified; these were re-annotated using the *S. lycopersicum* reference genome. The mRNA sequences and deduced amino acid sequences of these genes are also identified for further functional analysis. I determined the similarity of reannotated tomato genes with the genes identified in tomato-nematode gene interaction network in tomato plant.

Among the fifty-seven re-annotated genes, I identified one gene which is similar to one of the network identified (Genetics 207: **2017**).

Lightning Talks
1:45PM – 2:25PM

Using phylodynamics and decision trees to estimate the effects of mutations on viral transmission rates

Lenora Kepler under the advisement of *Dr. David Rasmussen*
Program: Bioinformatics

In the context of rapidly evolving pathogens such as Ebola and HIV, it is crucial to be able to track mutations in a microbial population, and the effects of these mutations on the transmissibility of a pathogen.

While phylodynamic methods exist to estimate the fitness of a given viral genotype, the current methods are too computationally intensive to explore complex genotype to fitness maps, including epistatic interactions between sites.

We propose a methodology that integrates and extends existing maximum likelihood inferences based on birth-death models with decision trees. These methods are validated using simulation studies, and then are applied to the Ebola outbreak that occurred in Guinea, Sierra Leone, and Liberia between 2014 and 2016 to estimate the fitness effects of nine observed mutations in the virus's glycoprotein.

Exploring random forest approaches for phenotype prediction using GWAS data for predicting Asthma.

Vaishnavi Venkat under the advisement of *Dr. Jung-Ying Tzeng*
Program: Bioinformatics

Asthma is a chronic disease involving the airways in the lungs. All risk factors for asthma come down to some form of environmental exposure and genetic predisposition that cannot be captured by standard single-SNP GWASs. We propose the use of random forests classifiers to select SNPs that would result in an improved predictive model of asthma exacerbations. We explored different methods to identify potentially significant variants related to asthma and extracted them to build a random forest model. The effects of different parameters of a random forest model were studied. The performance of a LASSO regression model was used as a baseline to see if the random forest approach was doing better. Finally, we developed a generalized random forest implementation for phenotypic prediction from GWAS data.

Prediction of Synergistic Drug Combinations with Deep Learning
Jun Ma under the advisement of *Dr. Alison Motsinger-Reif*
Program: Bioinformatics

Cancer is one of the main causes of death worldwide. Combination drug therapy has been a mainstay of cancer treatment for decades, which has been shown to reduce host toxicity and prevent the development of acquired drug resistance. However, the immense space of possible drug combinations makes it infeasible to experimentally screen all effective drug pairs. Therefore, it is crucial to develop computational approaches for predicting drug synergy and guide experimental design for the discovery of rational combination therapy. In this work, we present a new deep learning approach to predict synergistic drug combinations by integrating gene expression profiles from cell lines and chemical structure data. Specifically, principal component analysis was used to reduce the dimensionality of the chemical descriptor data and gene expression data. Then the low-dimensional data were propagated through a neural network to predict drug synergy values. We applied our method to O'Neil's high-throughput drug combination screening data. We demonstrate the effectiveness of the approach, and compare its performance with three state-of-the-art machine learning methods: Random Forests, Elastic Net and XGBoost.

Detecting Resistance Alleles in Upland Cotton (*G. hirsutum*) to the Cotton Leaf Curl Virus
Ashley Schoonmaker under the advisement of *Dr. Amanda Hulse-Kemp*
Program: Bioinformatics

Cotton Leaf Curl Virus (CLCuV), the causative agent for Cotton Leaf Curl Disease (CLCuD), first appeared in Multan, Pakistan in the 1980s. The Multan strain of the virus was mitigated with resistant varieties until the early 2000s, when the new Burewala strain overcame resistance previously bred into cotton varieties. The disease stunts growth of the plant and prevents the production of flowers and therefore cotton fibers. CLCV causes an estimated loss of 2 to 3 million bales of lint in Pakistan. Whitefly is the CLCuV vector and is a pest on many crops and ornamentals. The disease has also been reported in neighboring countries, i.e. India and China, and it is the potential threat in all countries where whitefly is prevalent causing concern that the virus will move into unaffected countries before resistance can be bred. Currently development of resistant lines depends on screening each generation in Pakistan. DNA markers would allow development of resistant varieties without field screening in Pakistan each generation. By making F2 mapping populations from crosses between a resistant line from one of two different sources and a non-resistant line, we used quantitative trait loci (QTL) mapping to identify single nucleotide polymorphism (SNP) markers associated with CLCuD resistance trait. The study used SNP markers obtained using CottonSNP63K array data and phenotypic data from the F2 populations. Genetic linkage mapping and QTL mapping identified SNP markers associated with resistance. SNP markers were identified for each F2 population, but the SNPs for each were from different chromosomes, indicating the resistance may be due to different alleles in the two resistance sources. Scripts were developed in R to

further streamline the process of gathering info from the array through to analysis. Future directions will strive to validate the QTLs through additional F2 populations and SNP assays created from validated SNPs. SNP assays will be used for marker assisted selection and eliminate the need to field screen every generation.

Computational prediction of long noncoding RNA related to HIV infection
Ian Huntress under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

Long noncoding RNA (lncRNA) are transcripts longer than 200 base pairs that do not code for protein. Several lncRNA have been associated with HIV infection including Nuclear Enriched Abundant Transcript 1 (NEAT1) and HIV-1-enhanced lncRNA (HEAL) through experimental validation. To further characterize HIV-lncRNA association, Peng Lab is conducting lncRNA CRISPRi screens in CD4+ SUP-T1 cell lines. Using high throughput sequencing of lncRNA-targeting guide RNA we can assess lncRNA HIV-relevance in vitro. However, it is unknown whether in vitro screening results of HIV-relevant lncRNA targets will remain the same in vivo. Therefore, we propose a simple linear model for prediction of lncRNA related to HIV infection using CRISPRi screen data which will serve as a foundation for future modeling in vivo. LncRNA features were selected as model inputs based on their potential to capture the mechanisms of lncRNA function during HIV infection. For each lncRNA, calculated features include: sequence k-mer counts, HIV associated SNPs, RNA-binding protein motifs, and lncRNA transcript abundance. These features access potentially thousands of interactions that can cause the model to narrowly fit the noise of the screen data. To minimize this overfitting, we restrict these features to the subset of known immune and HIV-related pathways and use cross validation to eliminate uninformative features. We anticipate that lncRNA predictive modeling could represent a critical first step toward data-driven exploration of lncRNA function.

Early-life TCDD Exposure Shapes Gene Expression and Chromatin Profiles Across the Life-Course of Mice.
Jackson Parker under the advisement of *Dr. David Aylor*
Program: Functional Genomics

Early-life exposure to environmental toxicants alters molecular profiles of affected tissues. In turn, exposed individuals may experience increased susceptibility to disease. In order to test how early-life toxicant exposure is predictive of molecular profiles over time, we exposed mice to a low dose of TCDD, a potent environmental toxicant, from preconception through lactation. We then measured gene expression and chromatin accessibility in liver tissue collected from four age points across the life course with sexes equally represented. From our measurements, we show that early-life exposure shapes gene expression and open chromatin profiles throughout life. However, signatures of exposure were distinct between sexes and across ages. We conclude that a complex cascade of gene regulatory events is set in motion by early-life TCDD exposure which results in long-term gene expression and chromatin accessibility differences in adult mice.

Talks - Biomathematics Focus
2:25PM – 2:55PM

Population model for the invasive insect *Homalodisca vitripennis* and the egg parasitoid *Cosmocomoidea ashmeadi*

Annabel Meade under the advisement of *Dr. HT Banks and Dr. Hien Tran*
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 *Cosmocomoidea ashmeadi* and *Cosmocomoidea 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 *C. ashmeadi* as a biological control method cause a significant decrease in the population of *H. vitripennis*?

Network modeling of plant diseases: the case of cucurbit Downy mildew in Eastern United States

Maureiq Ojwang' under the advisement of *Dr. Alun Lloyd and Dr. Peter Ojiambo*
Program: Biomathematics

Over the last 20 years, systematic research has been conducted to develop and validate the prediction of cucurbit downy mildew (CDM) in space and time. These efforts have resulted in a prediction framework to guide growers and policymakers in making the relevant decisions in managing CDM. The prediction framework relies on CDM reports from an extensive network of sentinel plots (disease monitoring locations that are strategically placed within specific states). The data about the current disease locations can be used to model the future spread of CDM. We used a dynamic network model for CDM epidemics, with sentinel plots as nodes and edge weights as a function of host density, wind speed, and direction. The model incorporates a power-law function for dispersal. We used the network model and centrality measures to select the most important sentinel plots in terms of node strength, network stability, disease monitoring, and transmission. This information can be used to reduce the resources required to scout and predict CDM outbreak and invasion progress.

Talks - Bioinformatics & Functional Genomics Focus
3:15PM – 4:00PM

Fast Multivariate Estimation for Heterogeneous Traits
Bryan Ting under the advisement of *Dr. Fred Wright and Dr. Yihui Zhou*
Program: Bioinformatics

We introduce a Two-Stage Composite Likelihood approach for multivariate probit estimation for responses with binary components. This approach is designed to be computationally fast and statistically efficient, lending itself well to settings such as Genome Wide Association Studies (GWAS) for detecting Single Nucleotide Polymorphism (SNP) associations with mean changes in phenotype.

Historically, multivariate probit estimation has been slow and/or computationally intensive. As the number of components and predictors increases, the number of coefficient and correlation parameters tends to grow quickly, adding to the difficulty of numerical optimization in finding estimates. This can be especially challenging in the genomics space, due to the number of predictors involved. Thus, we address this by separating estimation into two stages, where coefficients are estimated in the first stage via univariate probits and correlation parameters in the second stage via bivariate probits.

Next, we extend this approach to incorporate heterogeneous multivariate responses, i.e. where the response can include both binary and continuous components. Now instead of having only bivariate probits as associated likelihoods, there can also be bivariate normal densities for continuous-continuous pairs, as well as associated likelihoods for binary-continuous pairs.

Further considerations could include heteroskedasticity in the GWAS setting, i.e. non-constant variance across allele counts. There could be possible tie-ins here with gene-gene and/or gene-environment interactions. SNPs associated with variance may be indicative of the presence of interactive effects.

Differential Gene Expression in Trojan fir roots in response to *Phytophthora* Infection
Will Kohlway under the advisement of *Dr. Ross Whetten*
Program: Functional Genomics

The oomycete, *Phytophthora cinnamomi Rands,* causes root rot disease on a broad range of fir and pine species used as Christmas trees. One of the most valuable Christmas tree species, Fraser fir (*Abies fraseri* [Pursch] Poir.) has no innate immunity to *Phytophthora*. However an exotic fir species, Trojan (*Abies equi-trojani*) fir has previously shown varying amounts of resistance to Phytophthora root rot.

Two sets of seedlings from an open-pollinated family of Trojan fir was inoculated to with a single strain of *Phytophthora cinnamomi.* After 48 and 96 hours of incubation with

Phytophthora, the root tip from each seedling within the set was harvested and the seedling was transplanted to sterile medium. Mortality of the transplanted seedlings was observed over 16 weeks, after which RNA was extracted from each of the inoculated root tips. For each exposure length, the extracted RNA were pooled and sequenced in groups of five based experimental phenotype of mortality due to Phytophthora root rot along with non-inoculated control seedlings. A total of 28 of libraries were made. The RNA sequences were assembled and the different phenotypic groups were used to identify genes differentially expressed in resistant or susceptible individuals. This study will help to identify the genetic basis of *Phytophthtora* resistance in Fir species.

Incorporating DNA High Dimensional Structure into Identification of Risk Rare Variants

Yueyang Huang under the advisement of *Dr. Jung-Ying Tzeng*

Program: Bioinformatics

The 3D chromosome structure contains constructive information on how variants or genes interact and function together, and can thus guide the assessment of the joint, interactive effects of multiple variants that are close in 3D structural space but are apart in sequence location. To achieve this goal, we propose to develop and apply association tests that integrate 3D chromosome structure in Hi-C data to evaluate the genetic effects of a variant set and to identify other variants that potentially interact with the target variant set to affect the trait. Instead of annotating identified variants with known structural/functional information that is not trait-specific, our method incorporates the chromatin interactions as prior information to encourage similar effects of the nearby SNPs in the 3D space only if there appears to be sufficient support from the data to do so. In addition, structural supervision can help to dramatically reduce the search space for the potentially interactive variants affecting the traits.

An electronic version of this document is available at:

https://gsbmasymposium.wordpress.ncsu.edu/program/