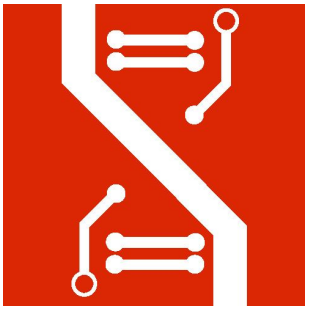
NC STATE
Biomathematics Graduate Program

BASF
We create chemistry



Genomic Sciences & Biomathematics Symposium

Spring 2019

Friday, February 22nd
6:00PM – 9:30PM
JC Raulston Arboretum – York Auditorium

Saturday, February 23rd
10:00AM – 5:00PM
James B Hunt Jr. Library – Room 4106

We would like to acknowledge BASF for their generous contribution to help fund this year's symposium. We would also like to acknowledge the following people for funding (in alphabetical order): Dr. David Bird, Dr. Spencer Muse and Dr. Fred Wright. Lastly, we would like to acknowledge the Bioinformatics Research Center and the NC State Graduate Student Association for their support.

Schedule of Events

Friday, February 22nd

5:30PM – 6:00PM	**Appetizers & Registration**
6:00PM – 6:15PM	**Welcome**
	Alice Toms, Genomic Sciences GSA President
6:15PM – 7:15PM	**Talks**
	Belinda Akpa, Daniel Schrider
7:15PM - 8:00PM	**Keynote Speaker**
	Kirk Francis
8:00PM – 8:30PM	**Dinner**
8:30PM - 9:30PM	**Social Hour & Poster Session**
	Jeremy Ash, Lenora Kepler, Jackson Parker, Patrick Perkins, Kuncheng Song

Saturday, February 23rd

9:30AM – 10:00AM	**Bagels, Coffee, & Registration**
10:00AM – 10:15AM	**Welcome**
	Marco Hamins-Puertolas, Biomathematics GSA President
10:15AM – 12:00PM	**Talks - Bioinformatics Focus**
	Hayden Brochu, Alice Toms, Bryan Ting, Jun Ma, Yue Hao, Tao Jiang
12:00PM – 1:30PM	**Lunch & Industry Roundtable**
1:30PM – 2:30PM	**Talks - Biomathematics Focus**
	Praachi Das, Brandon Hollingsworth, Annabel Meade, Jamie Nosbisch
2:30PM – 3:15PM	**Coffee Break & Networking**
3:15PM – 4:15PM	**Talks - Bioinformatics & Functional Genomics Focus**
	Aiden Jones, Yaxu Wang, Yueyang Huang, Ian Huntress
4:15PM – 4:25PM	**Closing Remarks**
	Hayden Brochu, Genomic Sciences GSA Treasurer
4:25PM	**After Hours Social**

Keynote Speaker

Five Cool Things BASF is Doing With Genomics
Kirk Francis, Ph.D.
Manager
Trait Knowledge & Performance Biologicals KTC
BASF Plant Science
Research Triangle Park, NC

Dr. Kirk Francis grew up on a small family farm in rural Iowa and developed an early interest in science and technology. After studying genetics at Iowa State University, he received graduate degrees in plant breeding from the University of Wisconsin and genetics from North Carolina State University. He continued his research in genetics at the University of North Carolina at Chapel Hill, and later joined BASF. While at BASF, Dr. Francis has led multiple trait discovery projects focused on gene and microbial discovery. Currently, he leads BASF's Performance Biologicals platform.

Presentation Session
Friday February 22nd 6:15PM-7:15PM

Realizing the potential of 'tiny data' – a cell biology case study
Belinda Akpa, Ph.D.
Assistant Professor of Integrated Synthetic and Systems Biology
Department of Molecular Biomedical Sciences, North Carolina State University
http://emrlab.org/

Stomata are the pores on a leaf surface that regulate gas exchange. Each stoma consists of 2 guard cells whose movements regulate pore opening and thereby control CO2 fixation and water loss. Guard cell movements depend critically on the remodeling of cell vacuoles, which have been observed to change morphology from a highly fragmented state to a fused state during stomata opening. The evolution of vacuole morphology requires a membrane fusion mechanism that responds rapidly to environmental signals, allowing plants to respond to diurnal cues or environmental stresses such as drought. With guard cells being both large and responsive to external signals, stomata represent a unique system in which to delineate mechanisms of membrane fusion and fission.

To resolve a counter-intuitive observation regarding the role of the HOPS protein complex in regulating vacuole morphology, we derived a quantitative model of vacuole fusion dynamics and used it to generate testable predictions about HOPS-SNARE interactions. We derived our model from limited - and, initially, qualitative - data by integrating statistical inference with fluorescence imaging and mechanistic modeling. The dynamic model predicted the evolution of vacuole morphology as it arises from intracellular signaling events that include: cytosol-to-membrane recruitment, chaperoned protein complexation,

and complex disassembly.

By constraining the model parameters to yield the emergent outcomes observed for stoma opening (as induced by two distinct signals), we predicted a dual role for HOPS and identified a stalled form of the SNARE complex that differs from phenomena reported in yeast. We predict that HOPS has apparently contradictory actions at different points in the fusion signaling pathway, promoting the formation of SNARE complexes, but limiting their activity.

Supervised Machine Learning for Population Genetic Inference

Daniel Schrider, Ph.D.
Assistant Professor
Department of Genetics, University of North Carolina at Chapel Hill
https://www.schriderlab.org/

Population-scale genomic datasets have given researchers incredible amounts of information from which to infer evolutionary histories. Concomitant with this flood of data, theoretical and methodological advances have sought to extract information from genomic sequences to infer demographic events such as population size changes and gene flow among closely related populations/species, construct recombination maps, and uncover loci underlying recent adaptation. To date most methods make use of only one or a few summaries of the input sequences and therefore ignore much of the information encoded in the data. In this talk we will demonstrate the enormous potential of adapting techniques from the field of machine learning to population genetic inference. We find that supervised machine learning tools typically outperform more conventional methods in part because of their ability to simultaneously examine of a variety of different facets of genetic variation measured by summary statistics. Modern deep learning tools also have the ability of operating directly on DNA sequence alignments as input, and can perform both evolutionary model selection and parameter estimation. Because these methods do not require explicit formulae for summarizing sequence data, they are particularly well suited for problems that have not received detailed theoretical treatments. Thus, when applied to population genetic data, machine learning approaches outperform expert-derived statistical methods, and offer a new path forward in cases where no theory-based approaches exist.

Poster Presentations
Friday April 6th 8:30PM-9:30PM

Molecular modeling of differential ERK1/2-ligand dynamic interactions and the development of ERK1/2 inhibitor resistance

Jeremy Ash under the advisement of *Dr. Jacqueline Hughes-Oliver and Dr. Denis Fourches*
Program: Bioinformatics

The development of small molecule inhibitors for the ERK1 and ERK2 kinases is a highly active area of research, as these two isoforms are key drivers of cancer cell proliferation.

However, drug resistance is the "Achilles heel" of kinases inhibitors. As members of signaling pathways that are critical to an array of cellular processes, kinase inhibition gives rise to strong selection pressure for drug resistance conferring mutations in patients. Because ERK1 and ERK2 are highly similar in both sequence and structure, it is unclear what selection pressure differences occur at the ERK1/2 binding site, and whether this contributes to drug resistance events in certain patients. In particular, there has not been a thorough exploration of how preferentially certain molecules can bind to ERK1 or ERK2, and whether there are key structural and/or dynamic differences in their binding modes. Recently, we conducted a study that encompassed (i) the structure-based docking of a series of inhibitors in the binding site of ERK2, (ii) the independent molecular dynamics (MD) simulations of each ERK2-inhibitor complex, and (iii) the computation of novel "MD descriptors" to characterize the dynamic non-covalent ERK2-inhibitor interactions. We discovered these MD descriptors could distinguish ERK2 binders from non-binders. In this project, we conduct this modeling workflow for a curated and refined set of small molecule inhibitors that have been tested against both ERK1 and ERK2 kinases. Our chemocentric analysis identifies significant differences between the dynamic intermolecular interactions of highly potent ERK2 inhibitors versus ERK1 inhibitors. This analysis deepens the understanding of why certain inhibitors interact differently with these two isoforms. We identify residues where the selection pressure driving drug resistance may be different, and inform chemists of strategies to design new ERK inhibitors with greater efficacy. At last, our analysis of ERK-ligand dynamic interactions that confer specificity to ERK1, ERK2, or both will aide in the development of new chemical probes capable of more selective inhibition.

Using phylodynamics and machine learning to estimate effects of mutations on viral transmission rate

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

While viruses are characterized by high mutation rates, mutations at a given loci have uncertain impacts on the fitness of a virus. The ability to identify the mutations that increase a pathogen's transmission rate would enable enhanced public health planning; targeted research into mechanistic interactions; and identification of potential drug resistance. One approach is to use phylodynamic methods, which enable population-level inference about pathogen characteristics from relatively small sample sizes by taking into account viral and population structure. The aim of this work is two-fold. First, to create flexible and extensible software to simulate a viral outbreak with specified mutation and population parameters, including sequence data and reconstructed transmission tree. We use this software with a fixed-population SIR model to generate sample sequence and phylogenetic data for viruses with loci of differing mutational cost. Second, we aim to create methods to infer the effects of viral mutations on transmission rate using this simulated data. This is done with numeric integration of a single-type birth death model, and then in the future with machine learning methods, which will be utilized to infer mutational cost even within complex gene interactions. In the future, we hope to apply

methods used on the simulated data to infer information about mutations' effect on viral transmission rate from real-world datasets.

Early-life TCDD exposure shapes gene expression across the life course of mice
Jackson Parker under the advisement of *Dr. David Aylor*
Program: Functional Genomics

2,3,7,8-Tetracholorodibenzodioxin (TCDD) is a potent environmental toxin that is generated as a byproduct of industrial operations involving high temperature processing of organic material. It enters into environmental systems as a constituent of solid waste and flue gas. In vertebrate systems, TCDD activates the AhR-mediated xenobiotic response which modulates transcription of numerous genes responsible for metabolizing toxic compounds. The World Health Organization recognizes links between early-life exposure to TCDD and late-onset pathologies including neurological disability, reproductive impairment, and increased cancer risk.

Our goal is to understand the consequences of early-life TCDD exposure on the molecular state of multiple tissues. Mice were exposed to TCDD from preconception through gestation and lactation. Tissue samples were taken three weeks, five weeks, twenty weeks, and forty weeks after birth. From our measurements of transcriptional profiles, we show that gestational TCDD exposure shapes gene expression both in the short-term and in the long-term. The effects of exposure differed dramatically between males and females. Furthermore, there is no overlap between the response in liver and in blood. Though there were clear gene expression signatures of TCDD exposure at all ages, the changes observed at three weeks did not persist into adulthood. We conclude that a complex cascade of gene regulatory events are set in motion by early-life TCDD exposure that result in long-term gene expression differences in adult mice.

A Negative Selection Inspired Anomaly Detection Algorithm for Identification of Corrupted Ribo-seq Samples
Patrick Perkins under the advisement of *Dr. Steffen Heber*
Program: Bioinformatics

Ribo-seq is a technique used for quantifying protein synthesis by determining the positions of active ribosomes along transcripts. Despite its recent rise in popularity, quality control and analysis of Ribo-seq data remain challenging, mainly due to its complicated experimental protocol and demanding preprocessing requirements. Because of this, a method with the ability to accurately determine the quality and reliability of Ribo-seq data is greatly needed. Here we describe a novel negative selection inspired anomaly detection algorithm, which we call Boundary Detection Using Nearest Neighbors, or BDUNN. The algorithm attempts to define the boundary between the self and non-self spaces using a set of detectors, and subsequently employs a nearest neighbor algorithm to classify unseen points based on their proximity to the detectors and a subset of training points. We also establish methods for improving the performance of the algorithm through detector set optimization. We show that BDUNN is capable of accurately and efficiently detects

anomalies when compared to other popular negative selection and one-class classification algorithms. Furthermore, we demonstrate that this method is not only capable of identifying low quality Ribo-Seq samples, but can also highlight the features which are the sources of the quality irregularities.

Development of New Paired-read Merging Criteria for ITS Amplicon Sequencing
Kuncheng Song under the advisement of *Dr. Ben Callahan*
Program: Bioinformatics

ITS amplicon sequencing is the most widely used method for elucidating the diversity and structure of fungal communities. The ITS region is the preferred genetic barcode for fungi, but it presents specific challenges related to the high levels of length variation at this locus. As a result, current bioinformatics methods for analyzing ITS amplicon data systematically exclude taxa that with lengths outside the expected range, particularly taxa with long ITS regions that cause paired-end sequences to fail to overlap. My rotation project at Dr. Ben Callahan's laboratory focuses on improving the read- merging algorithm in the Divisive Amplicon Denoising Algorithm (DADA2) package that will retain real ITS sequences that are currently lost due to their lengths. More specifically, we would like to keep legitimate read pairs where the forward and reverse reads did not overlap. The goal of the project is to keep the pairs that we have sufficient evidence to identify as reads from the same template and evaluate how this affects the overall performance of fungal community profiling. The process of this project include understanding the DADA2 method, development of new criteria to identify legitimate non-overlapping read pairs, implementation of those criteria in the R package, and rigorous evaluation of the new functionality on samples from synthetic and real fungal communities.

Talks - Bioinformatics Focus
Saturday February 23rd 10:15AM-12:00PM

Simultaneous metatranscriptomic and host transcriptional profiling to elucidate signatures of host-pathogen-microbiome interactions
Hayden Brochu under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

During infection, the interplay between microbiota, pathogen and host response in a clinical setting can significantly improve our understanding of disease progression and pathogenesis. With sequencing ribosomal RNA-depleted total RNAs (Total RNA-Seq), there is an under-explored potential to investigate these complex interactions through parallel analysis of microbial community activity and host transcriptional response. Here we aimed to explore the feasibility of this strategy with a pilot study of cervical samples obtained from women at high risk for sexually transmitted infections (STIs). These clinical samples were analyzed using Total RNA-Seq, resulting in approximately 260 million paired-end sequencing reads. We performed separate analysis of infected pathogens, host transcriptome, and metatranscriptome in the same samples, using sequential alignments to the human reference genome followed by a collection of thirteen thousand bacterial

genomes and a parasite genome of interest (*Trichomonas vaginalis*). We computationally determined the taxonomic profiles of the microbial communities, the transcriptional profiles of the host response, as well as the transcriptomes of STI pathogens detected in the microbial communities. After integrating these data, we detected clustering of host transcriptional profiles that reflected microbiome differences and STI infection. We also coupled the detection of overrepresented host regulatory pathways with pathogenic transcriptional activity to infer potentially novel biomarkers. This study demonstrated a unique bioinformatics opportunity for integrating orthogonal omics data to improve the understanding of host-pathogen-microbiome interactions. We expect broader applications of Total RNA-Seq analysis of clinical samples with advanced analysis can enable researchers to uncover novel, outcome-relevant biomarkers.

Analysis of mycobiomes to uncover biodiversity: a case study between soil fungi and orchid species in Sweden

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

An analytical approach and framework were developed to examine the soil mycobiome associated with orchids in Sweden as a case study and proxy for improved taxonomic placement and understanding of fungal biodiversity. DNA was extracted, amplified, and sequenced using the entire Internal Transcribed Spacer (ITS1-5.8-ITS2) region from soil sampled from eight locations in central Sweden. Sequence data was analyzed using the DADA2 pipeline in order to use the resulting amplicon sequence variants (ASVs) for higher resolution of fungi present in the soil. Initial taxonomic assignments were performed using the UNITE database and association between fungal order and orchid species was subsequently measured using Chi-square tests. This analysis showed an apparent enrichment of Sebacinales, which are fungi commonly associated with roots of orchids and other plant species. Further taxonomic resolution and insight into trophic behavior of sampled Sebacinales across different host plants was examined using the Tree-Based Alignment Selector (T-BAS) toolkit version 2.1. This analysis was based on a comprehensive Sebacinales phylogeny of the ITS1-5.8-ITS2 region developed by Oberwinkler and colleagues. Phylogenetic placement showed a clear affinity of putatively sampled Sebacina taxa to other Sebacinales reference taxa that represent beneficial endomycorrhizal symbionts of orchids, ectomycorrhizae, and saprobic lifestyles.

Fast Multivariate Probit Estimation via Composite Pairwise Likelihood & Differential Evolution

Bryan Ting under the advisement of *Dr. Fred Wright and Dr. Yihui Zhou*
Program: Bioinformatics

As has been commonly observed, decreasing sequencing costs and increasing computational power in recent years have provided an increasing availability of genomic data. Such increased availability of data naturally invites greater demand for methodological sophistication in modeling, testing, and inference. However, at the same time, striving for such methodological improvements can face considerable headwind in

the form of computational challenges and/or lack of clear statistical frameworks. Thus, we look to meet this demand by presenting some novel methods for genomics and precision medicine, but can also be applied in other settings. High priority concerns include computation time and statistical efficiency. We introduce a composite pairwise likelihood approach for multivariate probit estimation for responses with binary components, using differential evolution to perform likelihood estimations. This approach is designed to be fast, lending itself well to settings such as genomewide association studies for detecting single nucleotide polymorphism (SNP) associations with mean changes in phenotype. Next, we extend this approach by incorporating consideration for SNP-associated heteroskedasticity (i.e. variance changes in phenotype) for where the multivariate responses can include both binary and continuous components.

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, three-parameter logistic models, four-parameter logistic models and five-parameter logistic models but can also select the best model if no model assumption is made, which is especially useful in case of high throughput curve fitting. Compared with nonlinear least squares, the new method provides stable and robust solutions without sensitivity to initial values.

Patterns of population variation in two paleopolyploid eudicot lineages suggest that dosage-based selection on homeologs is long-lived.
Yue Hao under the advisement of *Dr. Gavin C. Conant*
Program: Bioinformatics

Genes that are inherently subject to strong selective constraints tend to be over-retained in duplicate after polyploidy. They also continue to experience similar, but somewhat relaxed, constraints after that polyploidy event. We sought to assess for how long the influence of polyploidy is felt on these genes' selective pressures. We analyzed two nested polyploidy events in Brassicaceae: the At-α genome duplication that is the most recent polyploidy in the model plant *Arabidopsis thaliana* and a more recent hexaploidy shared by the genus *Brassica* and its relatives. By comparing the strength and direction of the natural selection acting at the population and at the species level, we find evidence for continued intensified purifying selection acting on retained duplicates from both polyploidies even down to the present. The constraint observed in preferentially retained genes is not a result of the polyploidy event: the orthologs of such genes experience even stronger constraint in non-polyploid outgroup genomes. In both the *Arabidopsis* and *Brassica* lineages, we further

find evidence for segregating mildly deleterious variants, confirming that the population-level data uncover patterns not visible with between-species comparisons. Using the *A. thaliana* metabolic network, we also explored whether network position was correlated with the measured selective constraint. At both the population and species level, nodes/genes tended to show similar constraints to their neighbors. Our results paint a picture of the long-lived effects of polyploidy on plant genomes, suggesting that even yesterday's polyploids still have distinct evolutionary trajectories.

Same-Species Contamination Detection in Next Generation Sequencing
Tao Jiang under the advisement of *Dr. Alison Motsinger-Reif*
Program: Bioinformatics

Same-species contamination detection is an important quality control step in genetic data analysis in human genetic sequencing as samples might be contaminated by lab technicians or samples from other contributors. Compared with widely discussed cross-species contamination, same-species contamination is more challenging to detect and few methods have been published to address this issue. This article introduces a novel machine learning algorithm to detect same species contamination using support vector machines. Our approach uniquely detects such contamination using variant calling information stored in the variant call format (VCF) files (either DNA or RNA), and can differentiate between same species contamination and mixtures of tumor and normal cells.

In the first stage of our approach, a change-point detection method is used to identify copy number variations or copy number aberrations (CNVs or CNAs) for filtering prior to testing for contamination. Next, single nucleotide polymorphism (SNP) data is used to test for same species contamination using a support vector machine model. Based on the assumption that alternative allele frequencies in next generation sequencing follow the beta-binomial distribution, the deviation parameter is estimated by maximum likelihood method. All features of a radial basis function (RBF) kernel support vector machine (SVM) are generated using training data from Q2 Solutions and publicly available source, and then applied in the test data to detect contamination. If training data is not available, a default RBF kernel SVM model is used.

We demonstrate the potential of our approach using simulation experiments with varying levels of contamination. Sequence data in the form of fastq files from two publicly available cell lines (NA12878 and NA10855 from the 1000Genomes project) was used to create synthetically contaminated samples with various levels of contamination. VCF files were generated, and the power and false positive rate of our approach to detect same species contamination was evaluated. Our simulation experiments show that our method can detect levels of contamination as low as 5%, with reasonable false positive rates. We provide an R software implementation of our approach using the det_ct() function in the CTN R package.

Talks - Biomathematics Focus
Saturday February 23rd 1:30PM-2:30PM

Pathogen Evolution and the Probability of Mosquito-borne Disease Emergence: a mathematical host-vector transmission model

Praachi Das under the advisement of *Dr. Alun Lloyd*
Program: Biomathematics

Recent years have seen vector-borne viral diseases such as dengue, chikungunya and Zika become rising sources of global public health concern. The worldwide incidence of dengue has been estimated to have doubled in the last 20 years and chikungunya incidence has expanded to over 50 countries leading to millions of cases of the disease. Prior research has linked pathogen mutation and subsequent increase in efficiency of transmission via a secondary vector of the disease, such as *Ae. albopictus*, as a significant contributing factor. This change in disease vector brings about the possibility of disease spread to areas where the primary vector, *Ae. aegypti*, population is either very low or absent – a phenomenon which was previously less likely. For instance, there are many regions of eastern USA and the Midwest where *Ae. albopictus* is predominant which may now be at a higher risk for such diseases. The effects of pathogen mutation on the probability of disease emergence has previously been mathematically analyzed in a model which indicated that occasional mutation events could lead to the evolution of the original pathogen strain to have an $R_0 >$ the threshold value of 1. This could result in successful emergence of a disease even when the original pathogen is incapable of causing an outbreak. Here, we implement these findings in a branching process theory-based mathematical model for the two-step transmission characteristic of vector-borne diseases to investigate the influence of pathogen mutation on the probability of disease emergence. This model is a step towards a better understanding of these dynamics of disease emergence with regard to pathogen evolution, and our results have the potential to influence preventative and control measures that are currently taken against mosquito-borne diseases.

Spill-Over Effects of Yard-Scale Mosquito Control

Brandon Hollingsworth under the advisement of *Dr. Alun Lloyd and Dr. Michael Reiskind*
Program: Biomathematics

Aedes albopictus is an invasive, peridomestic, container-breading mosquito commonly found throughout the southeastern United States and world-wide. Its large range, preference for artificial containers, and a propensity for biting humans makes it one of the most prolific nuisance mosquitoes in the US. However, *Ae. albopictus* is also known to spread the arboviruses dengue, Zika, and chikungunya, making it a public health threat. Traditionally, responses to an outbreak of any of these diseases would utilize ultra-low volume (ULV) applications of insecticides to large areas using vehicles. However, evidence for the epidemiological efficacy of these treatments is lacking and there are a large number of off-target effects. Recently, though, there has been significant growth in private

mosquito control companies in the US that utilize yard-scale control measures to reduce mosquito nuisance around homes. While these measures are commonly used, the short-term efficacy of these controls and their effects on neighboring yards has not yet been quantified. Here, we describe a field experiment in which 16 pairs of neighboring houses were monitored following one yard being treated using either a bifenthrin barrier spray, larval habitat management (LHM), a combination of the two, or neither. We see that while the treatments each have different dynamics, they all show evidence for reduced mosquito numbers in neighboring yards. In light of this evidence, it may be possible that using these yard-scale control measures could take advantage of the spatial heterogeneity of *Ae. albopictus* and allow for more efficient responses to mosquito-borne disease outbreaks.

Delayed differential population models for the decline of *Homalodisca vitripennis* (Hemiptera: Cicadellidae) densities over a ten-year period

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 grape industries in California, because it spreads a disease-causing bacterium, *Xylella fastidiosa*. We continue an earlier investigation into a long-term phenological decline of *H. vitripennis* densities by studying a system of delayed differential equations (DDEs) and analyzing aggregate population data for *H. vitripennis* from a 10-year study in which bi-weekly monitoring of *H. vitripennis* populations decreased significantly. These data present several challenges for modelers. First, they involve truly *aggregate* population level sampling and hence cannot properly be treated as ordinary longitudinal time series data corresponding to individual level models. The appropriate modeling involves estimation of probability distributions for parameters rather than estimation of the dynamic parameters themselves. Moreover, our analysis reveals that the correct corresponding statistical models involve errors that are observation size dependent (e.g., relative errors should be employed in statistical models).

Feedback Loops at the Level of Lipid Metabolism Enhance Sensitivity and Robustness in Models of Chemotactic Gradient Sensing

Jamie Nosbisch 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, whereas pathways that collaborate to activate the Arp2/3 complex were found to be dispensable. PKC is activated through its binding to the lipid second messenger diacylglycerol (DAG), which is formed from hydrolysis of phosphatidylinositol (4,5)-bisphosphate (PIP2) 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, characteristic of an internal amplification mechanism. In previous work, we

developed a reaction-diffusion model of the PLC/PKC signaling pathway and implicated phosphorylation of myristoylated alanine-rich C kinase substrate (MARCKS) by membrane-localized PKC as a positive feedback mechanism sufficient for local amplification of DAG and active PKC. However, by itself, the MARCKS feedback only weakly amplifies the signal in shallow PDGF gradients.

Our new model includes phosphatidic acid (PA), a lipid intermediate in the metabolism of DAG. It has been shown that PA binds PLC and that active PKC can enhance the activity of phospholipase D (another enzyme that produces PA), implicating additional feedback loops. Model simulations show that the MARCKS feedback mechanism synergizes with these new feedback loops for increased amplification even at shallow PDGF gradients and over an appreciable range of midpoint PDGF concentrations. Simulations with variations of parameter values or cell geometry further indicate that this signaling network is a highly sensitive and robust gradient sensing circuit. We are currently integrating this model with 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.

Talks - Bioinformatics & Functional Genomics Focus
Saturday February 23rd 3:15PM-4:30PM

LncRNAs associated with immune response using CRISPR
Aiden Jones under the advisement of *Dr. Xinxia Peng*
Program: Genetics

Long non-coding RNAs (lncRNAs) are transcripts of length greater than 200 nucleotides that are not translated into proteins. Previously viewed as transcriptional errors, the function of the majority of lncRNAs is still unknown, thus providing a potential source for novel targets for many applications. Within our lab, we use RNA-Seq to analyze gene expression of cells with/out infection. In doing this, we identify individual lncRNAs for further functional analysis such as MALAT1. Alongside our experiments targeting selected lncRNAs, we also use the CRISPR-dCas9 system that can interfere with gene expression of a library of target lncRNAs in parallel. This CRISPRi screen is being used to identify additional lncRNAs that may have immune functions involved in viral infection.

Exploring the association between pre-birth maternal factors, cord blood methylation and newborns' growth curve
Yaxu Wang under the advisement of *Dr. Jung-Ying Tzeng and Dr. Cathrine Hoyo*
Program: Bioinformatics

Growth curve is a useful marker of general health. Irregular growth in early ages poses a potential risk factor for adulthood metabolic disorders. It has been reported that maternal factors such as pre-pregnant body mass index (BMI), smoking, parity and weight gain during pregnancy are associated childhood obesity. However, most of the studies focus only on categorized indicators of growth status at particular ages, instead of fully exploring

the entire growth curve. In addition, the epigenetic mechanism underlying those associations still remain unclear. CpG methylation in cord blood may act as a mediator between pre-birth maternal factors and newborns growth. In this study, we explored the effectiveness of functional principal component analysis (FPCA) in summarizing newborns growth curves. We found that the FPCA is powerful for dimension reduction of newborns' growth curve. We further examined the association between pre-birth maternal factors and first principal component score of newborns' weight curve. We also investigated the epigenome-wide association between cord blood methylation and pre-birth maternal factors. Our results support the mediator effect of methylation on the association between pre-birth factors and newborns' growth curve.

Effects of Peripartum Exposure to Anesthetics on Longer-term Chronic Diseases
Yueyang Huang under the advisement of *Dr. Jung-Ying Tzeng*
Program: Bioinformatics

Neuraxial anesthesia is widely used to provide analgesia for labor and anesthesia for cesarean section. Despite the widespread use of epidural and spinal anesthesia for delivery, data on the effects of these anesthetic techniques on longer-term chronic diseases and conditions in children are limited. Anesthesia logs and postnatal medical records were used to abstract anesthesia administration during parturition, and postnatal asthma and weight and height at age 5 years. Logistic regression models were used to evaluate associations between anesthesia during parturition and outcomes in male and female children. We present early data demonstrating that peripartum exposure to anesthetics/opioid analgesics increases the risk of asthma and childhood obesity at age 5, and associations are likely sex-specific. Larger studies are required to confirm these findings.

HIV infection and Total RNA-seq of PolyA depleted human SUPT1 cells reveals noncoding transcription and suggests extra-coding loci spanning known genes
Ian Huntress under the advisement of *Dr. Xinxia Peng*
Program: Bioinformatics

Infection can reveal differentially expressed coding and non-coding genes in host cells. However, RNA-Seq may only sample a subset of host response due to chemical selection of the full RNA transcriptome and limited sequencing depth. Comparison of selection by PolyA depletion, therefore represents an opportunity to reveal therapeutic targets, chromatin-silencing RNA, and fundamental protein-RNA regulatory networks promised by the growing non-coding RNA field. To that end, we used oligoDT beads to deplete polyadenylated (PolyA+) transcripts from a range of HIV infected and uninfected SUPT1 cell transcriptome libraries. Just as standard removal of rRNA dedicates more sequencing power to mRNA, depletion of PolyA+ RNA (primarily mRNA) dedicates more sequencing power to PolyA- non-coding RNA. We compute global genome coverage and find increased coverage in PolyA- samples, specifically within introns and flanking known coding regions. Building on the work of the Day Lab at UAB, we suggest these transcriptional patterns are explained by extra-coding RNA (ecRNA), which are unspliced, non-polyadenylated, non-coding transcripts that span over and beyond known genes. By definition, we expected

genes with unannotated ecRNA to show a trend toward an equal ratio of intron to exon reads as PolyA+ transcripts were depleted. By computing this trend over our samples, we produce a measure of ecRNA-like transcription in the context of HIV infection. Future work will use this measure to update our annotation with more certain ecRNA boundaries and allow differential expression analysis of these novel targets.

An electronic version of this document is available at:

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