Annual Genomic Sciences Symposium

Saturday, March 13th, 2021 10:00 AM Eastern Time

Zoom Meeting Information
https://ncsu.zoom.us/w/98577921926
Meeting ID: 985 7792 1926
Passcode: 800317

Session 1: Opening Remarks, 4 10-min talk, 2 lightning talk

10 - 10:15 AM : Check-in; Welcome talk
10:15 AM : Hayden Brochu
Title : Genetic diversity and transcriptional regulation of MHC-E: Implications for RhCMV-based vaccines
10:25 AM: Dillon Llyod
Title: Exposome Factors Associated with Development of Type 2 Diabetes
10:35 AM: Madison Moore
Title: An in-silico approach for taxonomic assignment of a novel isolate of Lactocaseibacillus (Lactobacillus) rhamnosus NCB 441
10:40 AM: Avery Roberts
Title: Evaluating functionality of transposon-encoded CRISPR-Cas with a cell-free expression platform
10:45 AM: Tanchumin Xu
Title: Relationship between IGC rates and paralog divergence
10:55 AM: Kuncheng song
Title: Systematic Comparisons for Microbiome-Based Disease Predictions

11:05 - 11:20 AM : Coffee break

Session 2: 5 10-min talk, 2 lightning talk

11:20 AM : Jackson Parker
Title: Impact of early-life TCDD exposure on molecular aging in the mouse liver.
11:30 AM : Caizhi Huang
Title: A Meta-Analysis of the Vaginal Microbiome and Preterm Birth
11:40 AM: William Kohlway
Title: Metagenomic analysis of Phytophthora-infected Trojan Fir roots
11:45 AM: Matt Nethery
Title: CRISPR classify: repeat-based classification of CRISPR loci

11:50 AM: Yueyang Huang

Title: Chromosome Structure Guided Rare Variant Association Test

12 noon: Meichen Pan

Title: Abundant Occurrence of CRISPR-Cas systems in Bifidobacterium Genomes

12:10 PM: Nathan Harry

Title: Genetic and regulatory architecture of a developmental dimorphism in the marine polychaete Streblospio benedicti

12:20 - 12:50 pm : Lunch Break

12:50 - 1:30 Poster session

Session 3: 6 10-min talk, 2 lightning talk

1:30 PM: Ian Huntress

Title: Computational Prioritization of long noncoding RNA in respiratory infection

1:40 PM: Vaishnavi Venkat

Title: Obtaining Polygenic Risk Score using Random Forest

1:50 PM: Lenora Kepler

Title: Decomposing the sources of SARS-CoV-2 fitness variation in the United States

2:00 PM: Ashley Schoonmaker

Title: A Whole-Genome Assembly of St. Augustinegrass and Detecting Resistance to Gray Leaf Spot

2:05 PM: Margot Ruffieux

Title: Understanding introgression and mito-nuclear incompatibilities in Saccharomyces cerevisiae

2:10 PM: Evan Walsh

Title: Development of single-cell based immune repertoire analysis in rhesus macaques

2:20 PM: Preethi Thunga

Title: High-throughput chemical hazard identification using behavioral assessments in zebrafish

2:30 PM: Montana Knight

Title: Assessing the Nucleotide-Level Impact of Spaceflight Stress using RNA-Sequencing Data

2:40 - 3 PM: Closing Remarks, Coffee break

3 - 4 PM: Informal chat with incoming students + Virtual Happy Hour!

ABSTRACTS:

Nnamdi Osakwe
PI: Dr. David Reif

Title of Poster: Building Comprehensive Models of Environmental Integrity

Currently, many environmental modeling tools function as information silos with little to no correspondence with other environmental domains and features. Current tools lack multi-faceted functionality and the ability to scale from local principalities. The potentially adverse environmental effects that these scores have on surrounding communities is often ignored and poorly emphasized. There is a need for a new framework that incorporates novel deep learning methods to integrate individual environmental domain indices via an ensemble learner to predict environmental quality as a dynamic quantity. Incorporating water, air, land, sociodemographic data from an array of publicly available environmental resources can be used to generate environmental quality integrity indices, forecast environmental integrity scores and comparing its relationship to health risks in a specific location. In addition, prioritizing translational accessibility and availability of this method would ensure that community stakeholders and risk assessors have an effective way to select problem sites for monitoring affected communities.

Jonathan Fleming
PI: Dr. David Reif

Title of Poster: Production of ToxPi Geospatial Feature Layers for an Interactive Visual Analytic

Painting a comprehensive picture of place-based, geographic data comprised of multiple factors is an inherently integrative undertaking. Furthermore, the visualization of this data in an interactive form is essential for public sharing and geographic analysis. This integration is made available via the ToxPi model, and interactive geographic mapping is possible through ArcGIS; however, production of ToxPi figures in ArcGIS maps is not available. We propose a method for drawing these interactive ToxPi figures in ArcGIS, as well as a pipeline for map sharing consisting of the ToxPi GUI for data integration, a custom python script for automated map layer production, and ArcGIS for map sharing. Although this method will require ArcGIS for map production, it does not require the user to have previous knowledge of ArcGIS software, as the usual manual steps for map production have been automated by the python script. This widens the user-base to anyone with access to ArcGIS licensing. Furthermore, map viewing and interaction is easily accessible to the general public via a web url. We illustrate this process with an example using current Covid-19 data to determine risk factors of counties across the United States. The script and its use instructions are available at https://github.com/Jonathon-Fleming/ToxPi-GIS.

Dani Joseph
PI: Dr. Santosh Mishra

Title of Poster: Neuro-immune Interactions linkage to Pruritus and Pain

Pruritus and pain are highly related sensations that are detected at the periphery by the sensory neurons, which peripheral afferents innervated into the skin. Individuals with chronic pruritus and pain experience intense detriment to their ability to function, unbalanced and distracted sleep, and many other symptoms that lead to a significant impact on their quality of life. Inflammation is the common link between pruritus and pain. Until recently, inflammation was associated with the local release of mediators and vasoactive substances by immune/non-immune cells but the role of neuropeptides released by the peripheral afferents and how they activate cell types in the skin remains elusive. Here, we have focused on one of the immune cells, mast cells, and their role in pruritus and pain due to the complex nature of the underlying interactions that may be the causation of the pruritus and pain sensations. To investigate these neuro-immune interactions, we used mouse molecular genetics, behavioral, cellular, and molecular approaches to provide mechanistic insights for these complex sensations.

Nirwan Tandukar
PI: Troy Ghashghaei

Title of Poster: Brain development and repair

The human brain is arguably the most complex of all biological systems. The different layers of cortex contain different types of neurons. Early developmental stage progenitor cells could produce any type of neurons but with age, they become more restricted. The mechanisms that control their proliferation and fate specification are important to understand and ameliorate a wide range of central nervous system (CNS) disorders including neurodegenerative diseases like Alzheimer's disease, Parkinson's, epilepsy, etc. The Ghashghaei lab is interested in Epidermal Growth Factor Receptor (EGFR) which is a regulator for neural stem cells. The lab has shown that EGFR is important for gliogenesis, gray matter astrocyte generation and oligodendrocyte generation. Evidences suggests that EGFR play important roles in psychiatric and cancer disorders. The brain of schizophrenic and depressed patients shows a decrease in EGF levels. Using Mosaic analysis with double markers (MADM), we made a population of cells with EGRF knockouts and wildtype in the same mouse brain. EGFR deletion in the ventral and dorsal region has shown to cause a transient defect in the development of the brain but the brain gets repaired as the mice matures. However, the olfactory bulb remains small. The lab is working intensively to characterize the defects and their mechanism

---

10:15 AM : Hayden Brochu

Title : Genetic diversity and transcriptional regulation of MHC-E: Implications for RhCMV-based vaccines

Rhesus macaques (RM) are an essential biomedical model for the study of human diseases and for vaccine development. Recently, rhesus cytomegalovirus (RhCMV)-based vaccine vectors have shown tremendous promise, programming highly diverse, nonclassical MHC-E-restricted CD8+ T cell responses and providing unprecedented protection against simian immunodeficiency virus (SIV). Despite this crucial importance of MHC-E, there remains a need for further investigation of its genetics and alternative splicing. Here, we use long-read sequencing to interrogate the RM MHC-E (Mamu-E) spliceosome. Using a combination of PacBio Iso-Seq, PCR, and Sanger sequencing, we show that Mamu-E transcribes at least 16 unique isoforms. These isoforms largely differ by splicing patterns that alter the 3' UTR and that lead to the loss or retention of the transmembrane and/or cytoplasmic domains. Using an independent Iso-Seq study from human, we further show that all Mamu-E splicing patterns were identified among human MHC-E (HLA-E) isoforms. Separately, we used PacBio Long Amplicon Analysis (LAA) to investigate the Mamu-E genetics of 58 RM from a RhCMV/SIV vaccine study and for the first time, to our knowledge, show evidence of Mamu-E copy number variation. Three groups of Mamu-E alleles were identified: two ~5% divergent full-length allele groups (G1, G2) and a third monomorphic group (G3) with a deletion encompassing the canonical Mamu-E exon 6. We found that G1 alleles composed ~90-97% of Mamu-E expression in animals with alleles from multiple allele groups, indicating that the spliceosome recovered was likely that of G1 alleles. Using whole blood mRNA-seq samples from these same animals we phased Mamu-E haplotypes that matched the G1 alleles and recovered additional variation in the 3' UTR not covered by the alleles. Finally, we interrogated the isoform expression of G1 alleles and show that isoform usage is stable throughout the RhCMV/SIV pre-challenge phase and is dominated by the canonical Mamu-E isoform. We expect these novel genetic and splicing annotation resources will facilitate future Mamu-E research, which in turn will better define the translatability of RhCMV-based vaccines.

10:25 AM: Dillon Llyod

Title: Exposome Factors Associated with Development of Type 2 Diabetes

Rationale: Type 2 Diabetes (T2DM) is a growing concern in the United States. While prevention research focused on a person's physical and/or genetic make-up has produced important insight into disease etiology, less is understood about the environmental components of disease risk.

Objectives: To comprehensively associate specific environmental factors with T2DM, we conducted an Environmental-Wide Association Study (EWAS), in which epidemiological data are comprehensively and systematically interpreted in a manner analogous to a Genome Wide Association Study (GWAS).

Methods: We performed an association analysis using 650 environmental factors with T2DM status using data from the Personalized Genes and Environment Study (PEGS). PEGS is a diverse North Carolina based cohort (71% white, 22% Black, 67% female). Exposome data was collected through three surveys, with assessments concerning endogenous and exogeneous exposure agents at home and work, and health and medical history (n=9,386). Logistic regression was used for single exposure association analyses. Single exposure analysis was adjusted for age, race, sex and BMI. Additionally, we built multi-exposure models for all outcomes applying the deletion-substitution-addition (DSA) algorithm. To examine nonlinear relationships, we used the machine learning methods LASSO regression and Knock Off Boosted Trees (KOBT).

Results: We found associations with asbestos [OR=1.5, 95% CI=(1.16-1.90)], gasoline exhaust [1.5, (1.15-2.00)], diesel [1.42, (1.08-1.87)], unusual irritability [1.29, (1.08-1.54)], and sleep trouble [1.09, (1.05-1.12)]. We found that sleep trouble, asbestos and gasoline were selected by the DSA algorithm as being in the most parsimonious model to explain T2DM risk. Sleep was also selected as being a variable of importance in our analysis using KOBT.

Conclusions: While hypothesis generating, our findings highlight the importance of considering multi-exposures for T2DM and demonstrate the potential of EWAS to better understand disease etiology. Starting to understand the environmental factors associated with T2DM, we can combine environmental risk with genetic risk of T2DM and understand important factors in predicting the development of T2DM.

10:35 AM: Madison Moore

Title: An in-silico approach for taxonomic assignment of a novel isolate of Lactocaseibacillus (Lactobacillus) rhamnosus NCB 441

Here, we describe the basis for the taxonomic classification of a novel strain of Lactocaseibacillus rhamnosus NCB 441, which was isolated from pickled white Domiati Egyptian cheese. Historically, misclassification of bacteria occurs when relying only on phenotypic characteristics and/or the variable 16S sequence regions. Today, microbial taxonomic assignment requires enhanced resolution beyond sequencing the 16S genes. Modern whole genome sequencing approaches ensure the correct classification and identification of unique strain features leading to more accurate phylogenetic and evolutionary relationships. Therefore, we conducted an extensive comparative genomic analysis utilizing average nucleotide identity (ANI) values to verify the assignment of the NCB 441 isolate to the species L. rhamnosus. The ANI of NCB 441 was over 97% comparable to each of the 192 L. rhamnosus strains with genome sequences publicly available in the NCBI GenBank database. Additional comparative genome analysis performed to the nine fully sequenced L. rhamnosus strains and two strains previously sequenced from our research group, confirmed over 97% nucleotide identity to three of the strains and over 99% identity to the other eight. Together, this data provides conclusive evidence for the novelty and the correct assignment of NCB 441.

10:40 AM: Avery Roberts

Title: Evaluating functionality of transposon-encoded CRISPR-Cas *with a cell-free expression platform*

CRISPR-Cas systems provide adaptive immunity in prokaryotes and serve to defend against foreign genetic elements such as bacteriophages. Currently, CRISPR-Cas systems are classified into two classes, six subtypes, over thirty subtypes, and variants thereof. Class 2 CRISPR-Cas systems possess single-protein effector modules, namely Cas9, Cas12, or Cas13. Class 1 systems, however, possess multi-protein effector modules, like the Cascade complex specific to type I CRISPR-Cas systems. Type I system functionality typically involves RNA-guided DNA targeting by the Cascade complex, followed by recruitment of the helicase/nuclease Cas3 for processive DNA cleavage. Recently, transposon-encoded type I-F3 variant CRISPR-Cas systems, which lack Cas3, were characterized and shown to be co-opted by Tn7-like transposons for RNA-guided transposition. As a genetic tool, these CRISPR-Cas systems and their associated transposon proteins can be employed for programmable, site-specific integration of large DNA payloads, circumventing the need for DNA cleavage and homology-directed repair that utilizes endogenous repair machinery. Here, we describe a testing approach for assessing previously characterized and orthogonal type I-F3 CRISPR-Cas systems through a cell-free transcription-translation expression platform.

10:45 AM: Tanchumin Xu

Title: Relationship between IGC rates and paralog divergence

Interlocus gene conversion (IGC) is a type of mutation that homogenizes DNA repeats by replacing a stretch of sequence in one repeat from a corresponding stretch in another. Although repeated DNA constitutes a large fraction of the genomes in humans and many other species, the dependence between repeats that is induced by IGC causes challenges for evolutionary inference. These challenges lead to IGC tending to be ignored when molecular evolution is studied. As a result, the evolutionary impact of IGC remains largely uncharacterized. This is especially unfortunate given that earlier work from our group demonstrated that IGC might be responsible for a large fraction of all sequence change in some duplicated protein-coding genes.

Building upon this earlier work, I am trying to determine how that rate of IGC depends on the divergence between duplicated sequences. There is strong evidence that IGC rates are negatively correlated with divergence, but this relationship between divergence and IGC rates remains to be quantified. In this talk, I will overview my data augmentation approach for studying the relationship between IGC rates and sequence divergence as well as present some preliminary simulation results.

10:55 AM: Kuncheng song

Title: Systematic Comparisons for Microbiome-Based Disease Predictions

Microbiome composition profiles generated from 16S rRNA sequencing have been extensively studied for their usefulness in phenotype trait prediction, including for complex diseases such as diabetes and obesity. These microbiome compositions have typically been quantified in the form of Operational Taxonomic Unit (OTU) count matrices. However, alternate approaches such as Amplicon Sequence Variants (ASV) have been used, as well as the direct use of k-mer sequence counts. The overall effect of these different types of predictors when used in concert with various machine learning methods has been difficult to assess, due to varied combinations described in the literature. Here we provide an in-depth investigation of more than 1,000 combinations of these three clustering/counting methods, in combination with varied choices for normalization and filtering, grouping at various taxonomic levels, and the use of more than ten commonly used machine learning methods for phenotype prediction. The use of short k-mers, which have computational advantages and conceptual simplicity, is shown to be effective as a source for microbiome-based prediction. Among machine-learning approaches, tree-based methods show consistent, though modest, advantages in prediction accuracy. We describe the various advantages and disadvantages of combinations in analysis approaches, and provide general observations to serve as a useful guide for future trait-prediction explorations using microbiome data.

11:20 AM : Jackson Parker

*Title: Impact of early-life TCDD exposure on molecular aging in the mouse
Liver.*

Consistent with the DOHaD hypothesis, environmental perturbation occurring early in life can have lasting impacts on liver health. In utero exposure to potent environmental toxicant dioxin (TCDD) has been linked with increased susceptibility to non-communicable diseases including metabolic disorder and liver cancer. Previous studies have shown that immediate response to TCDD in the liver alters the local epigenetic landscape around canonical metabolic enzymes resulting in increased expression of these genes. In this study, we assessed the impacts of early-life TCDD exposure on global profiles of chromatin accessibility and gene expression in adolescent and adult mice.

Early-life TCDD exposure disrupted chromatin accessibility and gene expression profiles across the life-course of mice. However, signatures of past exposure were primarily sex- and age-specific with very few differentially accessible regions or differentially expressed genes persisting from initial exposure into adult age points. Across both assays, five-week old mice exhibited little molecular evidence of toxicant exposure, yet considerable disruption to chromatin accessibility and gene expression was detected in three-week mice as well as in adults. Our data suggests that the initial response to TCDD sets forward a complex cascade of events impacting molecular aging and ultimately leading to widespread disruption of long-term chromatin accessibility and gene expression profiles.

11:30 AM : Caizhi Huang

Title: A Meta-Analysis of the Vaginal Microbiome and Preterm Birth

Preterm birth (PTB) is the primary cause of neonatal morbidity and mortality. Many studies have suggested PTB is related to vaginal microbiome, however, the conflicting reports have often caused confusion. Here we perform a meta-analysis of 12 PTB studies (n= 6281 from 1926 individuals) using 16S rRNA gene sequencing to explore the factors leading the inconsistencies, such as technical difference, population, and definition of PTB. We collect a common set of metadata from each study, process the sequencing data based on DADA2, and then perform intra-, cross- and combined-study analysis to compare the prediction accuracy for different (1) taxonomic resolutions; (2) data transformation methods and classifiers; (3) definitions of PTB; (4) population characters. Our current results show PTB group defined as gestational age at delivery (GAAD) less than 32 weeks have better prediction accuracy than PTB group defined as GAAD in 34-37 weeks, which might indicate the vaginal microbiome profile difference between early PTB women and term birth (TB) women is larger than the difference between late PTB women and TB women. Moreover, taxonomic resolution at genus level and amplicon sequence variant level have equally well performance and are overall better than other higher taxonomy level. Random forest classifier and centered log-ratio transformation have overall best performance than other combination.

11:40 AM: William Kohlway

Title: Metagenomic analysis of Phytophthora-infected Trojan Fir roots

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. Additionally, microbial communities associated with the host's root system have been linked with a probiotic effect against root pathogens. A set of seedlings from an open-pollinated family of Trojan fir was inoculated with a single strain of Phytophthora cinnamomi. After a week of incubation with Phytophthora, the root tip from each seedling was harvested and the seedling was transplanted into sterile medium. Mortality of the transplanted seedlings was observed over 16 weeks, after which RNA was extracted from each of the harvested root tips. The extracted RNA were pooled and prepared using ribosomal depletion and sequenced in groups of ten based on similarity of survival time. Four pools contain samples from seedlings that died soon after inoculation, four contain samples from seedlings that survived the longest after inoculation, and two pools contain samples from non-inoculated control seedlings. The RNA sequences were assembled, and the different phenotypic groups were used to identify genes differentially expressed in resistant or susceptible individuals. About 75% of the differentially expressed transcripts were microbial derived. The transcripts were then analyzed using Mash and Kraken2 to ascertain members of the microbial community associated with resistant and susceptible fir roots. This study will explore the diversity of microbes within the rhizosphere which may be important for the interaction between Phytophthora and fir roots.

11:45 AM: Matt Nethery

Title: CRISPR classify: repeat-based classification of CRISPR loci

Detection and classification of CRISPR-Cas systems in metagenomic data has become increasingly prevalent in recent years due to their potential for diverse applications in genome editing. Traditionally, CRISPR-Cas systems are classified through reference-based identification of proximate cas genes. Here, we present a machine learning approach for detection and classification of CRISPR loci using repeat sequences in a cas-independent context, enabling identification of unclassified loci missed by traditional cas-based approaches. Using biological attributes of the CRISPR repeat, the core element in CRISPR arrays, and leveraging methods from natural language processing, we developed a machine learning model capable of accurate classification of CRISPR loci in an extensive set of metagenomes. Although the performance of cas-based identification will exceed that of a repeat-based approach in many cases, CRISPRclassify provides an efficient approach to classification of CRISPR loci for cases in which cas gene information is unavailable, such as metagenomes and fragmented genome assemblies.

11:50 AM: Yueyang Huang

Title: Chromosome Structure Guided Rare Variant Association Test

With the advancement of chromosome conformation capture technologies, we can now identify DNA-DNA contacts at high resolution. An important discovery of DNA 3D folding of the eukaryotic genome is the topologically associating domains (TADs). TADs are genomic regions that self-interact, but rarely contact regions outside the domain. TADs may drive gene expression by bringing enhancers, which boost the expression of distant genes. For example, the noncoding sequences within FTO gene act as long-range enhancers that directly interact with the promoter of IRX3 gene. Obesity-associated SNPs have been found both in the noncoding region of FTO and in the promoter of IRX3. Following the observation that the 3D structure of DNA influences gene expression, we propose a chromosome structure guided association test to provide variant-set association information using structure-guided aggregation of signal. Constructed under a kernel machine framework, the test performs association testing by borrowing information from neighboring variants in the DNA 3D space in a data-adaptive fashion.

12 noon: Meichen Pan

Title: Abundant Occurrence of CRISPR-Cas systems in Bifidobacterium
Genomes

The clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated cas) systems constitute the adaptive immune system in prokaryotes, which provides resistance against bacteriophages and invasive genetic elements. The landscape of applications in bacteria and eukaryotes relies on a few Cas effector proteins that have been characterized in detail. However, there is a lack of comprehensive studies on naturally occurring CRISPR-Cas systems in beneficial bacteria, such as human gut commensal Bifidobacterium species. In this study, we mined 954 publicly available Bifidobacterium genomes and identified CRIPSR-Cas systems in 57% of these strains. A total of five CRISPR-Cas subtypes were identified as follows: Type IE, IC, IG, II-A, and II-C. Among the subtypes, Type IC was the most abundant (23%). We further characterized the CRISPR RNA (crRNA), tracrRNA, and PAM sequences to provide a molecular basis for the development of new genome editing tools for a variety of applications. Moreover, we investigated the evolutionary history of certain Bifidobacterium strains through visualization of acquired spacer sequences and demonstrated how these hypervariable CRISPR regions can be used as genotyping markers. This extensive characterization will enable the repurposing of endogenous CRISPR-Cas systems in Bifidobacteria for genome engineering, transcriptional regulation, genotyping, and screening of rare variants.

12:10 PM: Nathan Harry

Title: Genetic and regulatory architecture of a developmental dimorphism
in the marine polychaete Streblospio benedicti

Maternal genetic effects on development play a role in phenotypic variance that is relevant to understanding genetic diseases and epigenetic inheritance. We investigate maternal effects on development by exploring gene expression in the marine polychaete Streblospio benedicti which has a unique developmental dimorphism. This Spiralian worm is useful because it exhibits poecilogony: two different development modes within a species that converge to produce the same adult phenotype. These types can be experimentally crossed to produce viable intermediate offspring in both directions which allows us to interrogate parental background effects, such as maternal genetic effects. To identify key genes that differentiate the maternal expression background of the two modes we analyze patterns of expression between the eggs (or oocytes) of females of the two types. Using the F1 intermediate offspring, we can identify which expression differences are due to cis or trans acting factors in the oocyte, demonstrating the effect of the maternal background on expression.

1:30 PM: Ian Huntress

Title: Computational Prioritization of long noncoding RNA in respiratory
Infection

Long noncoding RNAs (lncRNAs) may play important roles in early host-response to respiratory infection. However, it remains difficult to identify key regulatory lncRNAs from lowly expressed and incompletely annotated lncRNAs that appear in high throughput sequencing experiments. We propose simple data-integration methods to address these challenges by taking advantage of several types of publicly available respiratory infection data.

To analyze an RNA-Seq experiment, our lncRNA analysis pipeline follows three major steps. First, we anchor our analysis to the RNA-Seq experiment by filtering all lncRNAs down to a tractable subset of expressed candidates. Next, we generate hypotheses for lncRNA-induced patterns of gene expression based on public genomic data. Finally, we prioritize candidates based on how well these hypothesized expression patterns generalize beyond our specific experiment.

The new influx of public data from single cell respiratory infection experiments is well-suited to complement our prioritization of lncRNAs. Treating each single cell RNA-isolation as a replicate allows us to externally test our hypothesized lncRNA-induced gene expression patterns. To mitigate the risk of confounding our analysis with biases from external data, we present our simplest possible implementation of these methods. And, we ensure that each new data integration step remains interpretable and open to experimental validation.

1:40 PM: Vaishnavi Venkat

Title: Obtaining Polygenic Risk Score using Random Forest

Random forests (RF) have been considered an attractive alternative to linear models for outcome prediction with whole-genome data due to its capability of accommodating high dimensional predictors and non-linear polygenic effects. In this project, we explore various strategies to construct polygenic risk scores (PRS) from RF predictive models using RF. Using different models of simulations, we will assess the performance of the RF-based PRS by comparing the PRS using the standard pruning and thresholding methods of marginal association as well as the PRS using LASSO regression of joint association.

1:50 PM: Lenora Kepler

Title: Decomposing the sources of SARS-CoV-2 fitness variation in the
United States

The ability of a pathogen to transmit between hosts, or it’s fitness, is determined by the complex interplay between viral genomic features, intrinsic host features, and spatiotemporal factors. In the case of rapidly adapting pathogens such as SARS-CoV-2, it is imperative that we are able to predict which lineages are going to rapidly spread. While the impact of viral mutations can often be assayed at a cellular (or within-host) level, it is more difficult to determine their impact at the population level. Thus, we have developed methods that combine phylodynamics and machine learning in order to infer the fitness a given SARS-CoV-2 lineage based on its genotype, sampling time, and region. We can further use these methods to determine which of these factors are the main drivers of transmission potential across time.

2:00 PM: Ashley Schoonmaker

Title: A Whole-Genome Assembly of St. Augustinegrass and Detecting
Resistance to Gray Leaf Spot

In the past couple of decades, St. Augustinegrass has become extremely popular in southern states due to some cultivars' ability to thrive in sandy soils and across many warm climate regions. Despite its popularity and the vast knowledge of physical and breeding aspects of the species, there is little known about the genetic makeup of the grass. Like all grasses, St. Augustinegrass is extremely heterozygous making it difficult to create genome assemblies, and to date no genome reference sequence exists. Raleigh, a prominent diploid cultivar developed by the North Carolina State University breeding program, is highly resistant to a common virus, St. Augustine Decline, and has been noted to be more cold-tolerant than other St. Augustinegrass accessions. A reference genome was developed using PacBio Circular Consensus Sequencing (CCS) technology and scaffolded using information from previously developed linkage maps. The 851.8MB assembly was separated into two haplotypes sized 422MB and 429MB. Gray Leaf Spot (GLS), a fungal disease, plagues many grass species including St. Augustinegrass, perennial ryegrass, and tall fescue. This disease is caused by many species of fungus, one of the most important being Magnaporthe grisea, and is generally found in warm, humid regions. An experiment was developed to identify and characterize possible regions containing or closely linked to genes for gray leaf spot resistance by comparing RNA-Seq data from the susceptible cultivar Raleigh to a resistant cultivar PI 410353. Currently, annotation is being generated using RNA-Seq data produced here and from other projects as evidence.

2:05 PM: Margot Ruffieux

Title: Understanding introgression and mito-nuclear incompatibilities in
Saccharomyces cerevisiae

A critical and long-standing biological question is how do diversification and speciation occur? Part of this question might be answered with a better understanding of the molecular genetic mechanisms that maintain reproductive isolation between different species. There are many sources of conflict during hybridization, and this project focuses specifically on mitochondrial-nuclear interactions. Nuclear and mitochondrial genomes are inherited independently and accumulate mutations at different rates. Despite separate inheritance, many genes necessary for mitochondrial function are encoded in the nuclear genome, and therefore these two genomes must function together for cell survival. During hybridization, problems can arise when the two genomes are incompatible. This project thus aims to explore selective pressures on nuclear-encoded mitochondrial genes to gain insight to the molecular mechanisms of mito-nuclear incompatibilities using the system Saccharomyces cerevisiae. Several instances of mito-nuclear incompatibilities have been documented in Saccharomyces, however, there are no existing studies that have examined the link between mito-nuclear interactions and selection against introgressed genomic regions. Introgression describes when a segment of a species' genome is integrated into another species' genome. Recent studies supplied vast new resources for population genomic work in Saccharomyces cerevisiae, as well as provided robust computational methods for detecting introgression in genomic data. Utilizing these resources, this analysis seeks to illuminate how this mechanism is maintaining species boundaries. Alleles that reduce hybrid fitness during interspecies hybridization may be selected against; consequently, observing an underrepresentation of nuclear-encoded mitochondrial genes in introgressed regions of the genome could be indicative of a molecular incompatibility. This study aims to identify introgressed regions in over 1000 strains of Saccharomyces cerevisiae. Bioinformatic approaches are employed to determine if nuclear-encoded mitochondrial genes are in fact underrepresented in areas of introgression of Saccharomyces cerevisiae isolate genomes. To propel our understanding of evolutionary biology, it is important to identify selective pressures on nuclear-encoded mitochondrial genes, leading to better comprehension of mito-nuclear incompatibilities that contribute to reproductive isolation and speciation in hybrids across many taxa.

2:10 PM: Evan Walsh

Title: Development of single-cell based immune repertoire analysis in
rhesus macaques

Immunoglobulin (Ig) and T-cell receptor (TCR) repertoire analysis plays a key role in understanding the development of antigen-specific immunity during infection and vaccination. Due to the close phylogenetic relationship and highly similar physiology to humans, rhesus macaques (Macaca mulatta) have been one of the most popular and well-studied nonhuman primates (NHPs) for modeling immune responses in humans. However, current assays for rhesus Ig/TCR repertoire analysis have incomplete coverage and are not available at the single cell level.

Recently we computationally designed the first single-cell based Ig and TCR repertoire sequencing assays for rhesus macaques. Here we report the experimental assessment and optimization of these assays using rhesus PBMC and splenocyte samples. Using the generated single cell sequencing data we properly assemble and annotate full-length V(D)J sequences for every rhesus Ig and TCR isotype and chain type and obtain the pairing of Ig/TCR receptor transcripts in same cells. Additionally, we show that these rhesus Ig/TCR assays have capture specificities comparable to commercially available reagents. These results demonstrate that our optimized assays and custom computational workflows provide an unique opportunity for performing single-cell based Ig and TCR repertoire analysis in rhesus macaques.

2:20 PM: Preethi Thunga

Title: High-throughput chemical hazard identification using behavioral
assessments in zebrafish

The continual introduction of new chemicals into the market necessitates fast, efficient testing strategies for evaluating their toxicity. Ideally, these high-throughput screening (HTS) methods should capture the entirety of biological complexity while minimizing reliance on expensive resources that are required to assess diverse phenotypic endpoints. In recent years, the zebrafish (Danio rerio) has become a preferred vertebrate model to conduct rapid in vivo toxicity tests. Previously, using HTS data on 1060 chemicals tested as part of the ToxCast program, we showed that early, 24 hours post-fertilization (hpf), behavioral responses of zebrafish embryos are predictive of later, 120 hours post-fertilization, adverse developmental endpoints—indicating that embryonic behavior is a useful endpoint related to observable morphological effects. Here, our goal was to assess the contributions (i.e., information gain) from multiple phenotypic data streams and propose a framework for efficient identification of chemical hazards. We systematically swept through analysis parameters for data on 24 hpf behavior, 120 hpf behavior, and 120 hpf morphology to optimize settings for each of these assays. We evaluated the concordance of data from behavioral assays with that from morphology. We found that combining information from behavioral and mortality assessments captures early signals of potential chemical hazards, obviating the need to evaluate a comprehensive suite of morphological endpoints in initial screens for toxicity. We have demonstrated that such a screening strategy is useful for detecting compounds that elicit adverse morphological responses, in addition to identifying hazardous compounds that do not disrupt the underlying morphology. The application of this design for rapid preliminary toxicity screening will accelerate chemical testing and aid in prioritizing chemicals for risk assessment.

2:30 PM: Montana Knight

Title: Assessing the Nucleotide-Level Impact of Spaceflight Stress using
RNA-Sequencing Data

Space is an exciting frontier, but it presents unique environmental stressors like microgravity and space radiation. It is difficult to study the impact of spaceflight on terrestrial life due to the limited resources of the International Space Station (ISS). NASA created Genelab, a public Omics database for spaceflight relevant data, to give scientists access to data without needing a new experiment on the ISS. Transcription profiles are among the most common data type available on Genelab, and in the age of Next Generation Sequencing this commonly means RNA-Sequencing data. A developed pipeline that could find new information out of the public RNA-Seq data would be useful, especially in a case like this where the available data is extremely limited. This study aims to develop and use a pipeline analyzing Genelab's RNA-Sequencing data from Arabidopsis thaliana for sequence variants. The hypothesis is that space's environment will cause more variants to be called in the spaceflight A. thaliana samples than those in the ground control. RNA-Sequencing data is not the preferred method to call variants due an associated high false discovery rate, however recent studies show it can be done with appropriate precautions. The pipeline incorporates steps to combat factors leading to RNA Seq's high false variant discovery rate including 2-pass mapping methods and stringent filters. Preliminary results show A. thaliana samples from space tend to have higher variant counts than those from the ground. The development of this pipeline would demonstrate RNA-Seq can be analyzed beyond gene expression to see the impact of abiotic stressors like microgravity and space radiation. Further, if the results continue trending towards spaceflight A. thaliana having a higher number of variants then this will show the damage spaceflight can have at the nucleotide level. This may lead to a better understanding of what future precautions may need to be taken for the better safety of those in space.