Jason G. Mezey

RESEARCH INTERESTS

GENOMIC APPROACHES FOR STUDYING PHENOTYPIC VARIATION

How does genomic variation generate phenotypic variation? My research addresses this question by integrating analysis of genomic data with developmental genetic models of phenotypic traits. My approach is to focus on traits of *Drosophila* that are important models for quantitative genetics and developmental genetics. The goal is to understand why individuals and species have different phenotypes, how gene expression is regulated and evolves, and the connection between sequence, gene expression, and phenotypic variation.

My research program requires a diverse set of skills: experimental work, collection of genomic data, development of statistical tools for analyzing genomic data, and derivation of analytical theory. My grounding in these disciplines comes from my dissertation work in evolutionary biology and quantitative genetic theory and from my postdoctoral work in experimental quantitative genetics and genomics. I envision building a lab that both collects genomic data and has the statistical and data-mining expertise to analyze these data. I want a flexible lab that reaches out to people with other areas of expertise to find new approaches and address questions that complement and add to this work.

ONGOING RESEARCH

My current research is in three areas: expression analysis of phenotypic variation, genetics of multi-genic traits, and evolution of the *Drosophila* transcriptome. For all three, I use theoretical and statistical modeling approaches to analyze genomic data. Below, I briefly describe my current research in these areas and my plans for future work.

Expression Analysis of Phenotypic Variation

How does variation in gene expression generate variation in the distance between veins L3-L4 in the *D. melanogaster* wing? I study this trait because: (1) the genetic signaling pathways responsible for the placement of these veins are known, providing not only a list of candidate genes but also a network of gene interactions that can be analyzed, and (2) there is variation in this trait both within and among closely related species, making this an excellent model for understanding how evolution at the population level can produce the variation we observe among species. My interest in this system has developed from my postdoctoral work where I estimated an additive genetic covariance matrix (**G**) for twenty wing vein traits for *D. melanogaster*. A dimensionality analysis demonstrated that there is additive genetic variation for all possible combinations of traits (Mezey and Houle, submitted; Houle et al. 2004). This means that the position of veins can evolve in any direction in trait space.

How do gene networks generate this high dimensional variation in vein position? I have used whole-genome microarrays to assay gene expression levels in dissected wing discs of *D. melanogaster* using lines that were selected for large and small distances between veins L3-L4. Developmental geneticists have discovered that interruption or misexpression of genes in the Hedgehog (Hh) signaling pathway causes extreme effects on L3-L4 development such as truncated veins or partial L3-L4 fusion but it is unknown whether these pathways can account for quantitative variation. The wing disc expression data that I have collected are being used to determine whether variation in the expression of Hh genes is responsible for the quantitative variation in these lines (Mezey *in prep*). I am also using these data to develop spatial statistical models based on the Hh pathway to understand how variation in expression of *hedgehog* and a second locus, *engrailed*, which is necessary for *hedgehog* expression, works to produce variation by activating or repressing downstream genes such as *knot*, *vein* and *decapentaplegic*.

In the future, this system will be used to address a major question in evolutionary biology: are the factors responsible for variation within species also responsible for divergence among species? Expression variation among species in the *melanogaster* subgroup will be analyzed to determine if the genes responsible for variation in L3-L4 distance in *D. melanogaster* are also responsible for variation among species. I am currently writing an NIH grant in collaboration with Dr. Sergey Nuzhdin (UC Davis) to continue work on this system.

Genetics of Multi-Genic Traits

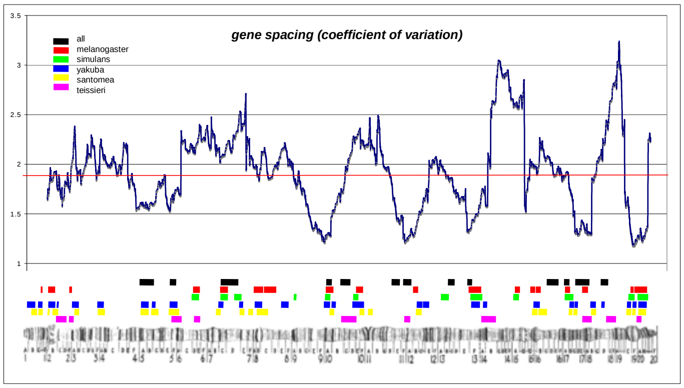
This research focuses on identifying sequence polymorphisms that are responsible for variation. The goal is to identify polymorphisms that cause variation in L3-L4 distance and variation in expression of genes in the Hh pathway. This approach complements my research on gene expression (above) by quantifying how polymorphisms generate phenotypic variation by producing variation in genetic pathways. I have completed the first step in this project: a quantitative trait locus (QTL) analysis of wing vein position in *D. melanogaster* (Mezey et al. *in press*). This study incorporated statistical techniques used in genomics to test for QTL interactions with the entire Hh pathway. This study identified a large number of QTLs with effects on vein position and many of these QTLs were found to have dominant interactions with genes in the Hh pathway indicating close association with Hh signaling.

This work will be extended by using two statistical approaches to analyze variation in L3-L4 distance and gene expression variation in a new set of recombinant inbred lines. First, association tests will be used to analyze sequence polymorphisms in the genomic regions identified in the initial QTL study. This will address questions such as: Which genes alter L3-L4 distance through the Hh pathway? Do these polymorphisms occur in coding or non-coding regions? Second, I will analyze the epistatic relationships among polymorphisms using techniques developed from new theoretical work. The theoretical work is based on my previous derivation of the set of genetic models that cause **G** matrices to have common principal components (Mezey and Houle, 2004). Such components occur only when there is a modular distribution of gene effects, implying that these special distributions can be identified by comparing **G** matrices. My new theory can be used to estimate the topology of the landscape that describes the epistatic relationships among sequence polymorphisms. This will be essential for addressing a fundamental question of multi-genic trait genetics: what is the distribution of epistatic effects?

Evolution of the **Drosophila** *Transcriptome*

The transcriptome is the collection of all gene transcripts in an organism. The *D. melanogaster* transcriptome is known to include sets of sequential genes that have correlated expression: co-expression domains. This research is focused on determining if co-expression domains have played a role in the evolution of the transcriptome among *Drosophila* species. I have assayed gene expression across the entire genome in five *Drosophila* species using Affymetrix microarrays. These microarrays were designed to assay expression in *D. melanogaster*. Although there is widespread interest in using these microarrays to examine expression in other species, there is a significant issue that must be resolved: sequence divergence among species introduces inaccuracies in gene expression measures. To deal with this problem, my collaborator Dr. Corbin Jones (UNC, Chapel Hill) and I have compared all the Affymetrix *D. melanogaster* probes to the genome of *D. yakuba*. Remarkably, we have found that the *melanogaster-yakuba* comparison allowed us to remove the confounding effects of sequence divergence on expression measures for all five species. Using this method we found variation among species that was missed by microarray studies that failed to account for sequence divergence (Jones and Mezey, *in prep*). Analysis of sequential co-expressed domains using the corrected data resulted in the identification a significant number of domains that are evolving among these species (Mezey and Jones, *manuscript under embargo*; see Figure). In addition, we have found that the average physical spacing of genes in these evolving domains is significantly less than the average spacing across the entire genome. This is expected if co-regulation of genes is the mechanism responsible for correlated evolution of these domains. As the *D. simulans* and *D. yakuba* genomes are annotated, we plan to examine the relative location of control elements relative to these regions to confirm that co-regulation is responsible.

Evolving Co-expression Domains:
X Chromosome



FUTURE WORK

I believe that genomic data present a tremendous opportunity to answer questions that were unapproachable just a few years ago. My future work will play a dual role in this context. First, I plan to develop theoretical, statistical and computational tools for applying genomic data to questions in a variety of biological fields. I see the application of genomic data as a chance for me to collaborate in areas beyond those in which I work directly. Second, I will address fundamental questions with experimental approaches that are streamlined to make use of large genomic data sets. While I will continue to work on *Drosophila*, I am excited by the prospect of expanding my approaches to include other organisms. Overall, my goal is to pursue projects that will make unifying contributions to genomics, genetics, and evolutionary biology.

Teaching Interests

My teaching interests reflect my research program, in which I synthesize genomic data and a broad range of statistical techniques to address fundamental questions in genetics, genomics and evolutionary biology. As an educator, I hope to provide a unique perspective on how borrowing from different fields can provide an opportunity to address difficult problems. I have been a teaching assistant for a diversity of classes as a graduate student and I have also developed and taught an undergraduate evolution course where I was the instructor. Reviews from this course were extremely positive and are available upon request. Below I provide brief descriptions of graduate and undergraduate courses I would like to teach.

Introduction to Genome Sciences

This course will provide an overview of genomic data and the computational tools necessary to utilize these data to address questions in a variety of biological fields. Topics will include genome sequencing, assembly and annotation, assays of transcript abundance, and functional proteomics. Students will be assigned small projects which will familiarize them with relevant databases and data mining/analysis tools. This class is intended for advanced undergraduate biology majors and could be adapted for graduate students by adding instruction in Perl and SAS.

GENES AND GENOMES

An undergraduate course intended for non-majors, aimed at students interested in obtaining a working understanding of genetics and genomics. A major focus of the course will be how advances in genomics are affecting our everyday lives. This course will include readings from popular science, as well as introductory biology texts. Three main topics will be covered: introduction to genetics, introduction to genomes and the genomic era, and evolution of genomes. Subtopics will include: history of genetics, gene structure and expression, genomic organization, medical and agricultural applications, nurture vs. nature, and popular misunderstandings of genetics.

Biostatistics

An undergraduate course for biology majors and a graduate level course. Topics will include: introduction to probability theory, parameter estimation, hypothesis testing, ANOVA, regression, non-parametric tests, construction of statistics, and appropriate tests. The course will focus on providing an intuitive grasp of statistics and application of statistics in biology. Examples and assignments will use data from studies that range from ecology and evolution, genetics and genomics, to medicine and psychology. The goal of this course is to make students proficient enough to use statistics as a rigorous modeling tool for addressing questions in any field of biology. The graduate level version of this course will additionally include the following topics: maximum likelihood estimation, advanced linear models, multivariate techniques emphasizing applications in genomics and programming in R and SAS. Assignments in the graduate course will include projects where students will be encouraged to analyze their own data.

Additional course topics

In addition to the courses described above, I am also interested in teaching introductory biology, evolution, molecular evolution, quantitative genetics, *Drosophila* development, and advanced population and quantitative genetic theory.