Research Interests

My main research interest is in understanding the evolutionary forces shaping genome structure and variability. To date, my research has focused on three different questions. First, what role does natural selection play in the evolution of duplicate genes? Second, are adaptive substitutions in nature generally dominant or recessive? Third, how can reliable inferences be made about the demographic histories of natural populations using genetic data? My approach to these questions has included both experimental and computational methods, often gathering new data that then require new tools to be developed for analysis [e.g. 7 (numbers refer to publications in attached CV)]. As a postdoc at Cornell, I have had access to large computer clusters that have allowed the development of new computationally-intensive methods that have been successfully applied to genetic data from natural populations of *Drosophila melanogaster* [1,3]. Future research will continue in this manner, integrating the collection of new data with the development of new methods and tools designed to leverage large amounts of computing power. In the sections below, I describe these areas of research, the directions they will take in the future, and current and future collaborators.

Evolution of Duplicate Genes

Background

Duplicated genes are a ubiquitous feature of eukaryotic genomes. The availability of complete genome sequences has reinvigorated evolutionary studies of gene families, in part by facilitating the discovery of species-specific duplication events and population-genetic studies of gene families. A major issue in the field is understanding why duplicate loci are maintained in genomes–i.e. do they provide new or specialized functions? Or are neutral processes (such as subfunctionalization) a sufficient explanation for the prevalence of duplicate loci?

Patterns of duplicate gene evolution differ between sex chromosomes and autosomes

As a graduate student I studied the evolution of duplicated genes in *Drosophila melanogaster*, taking advantage of the complete genome sequence of that species. This work revealed two novel, genome-wide, differences between the X and autosomes. First, X-linked duplications are significantly more divergent at the amino acid level when compared to autosomal duplicates, evidence that pervasive natural selection may drive the divergence of duplicates on the X [10]. Second, there is an excess of genes that have been duplicated from the X to the autosomes via transposon-mediated reverse-transcription [9]. The autosomal copies have recruited male-specific patterns of gene expression, which we interpreted as evidence that the duplicates were maintained because they provide functions in the male germline that would otherwise be lost when the X becomes transcriptionally inactive during gametogenesis, a hypothesis supported by the whole-genome expression studies of Parisi *et al.* (2003 *Science* 299: 697).

Natural selection on young X-linked duplications

In a follow-up study[4], I obtained polymorphism data from 17 duplicated loci in *D. melanogaster*, including the majority of X-linked duplicates with high amino acid divergence between copies. Analysis of this work required the development of new software designed to analyze polymorphism data from closely-related duplicate loci. The analyses revealed that these X-linked copies are most likely not accumulating mutations and degenerating into pseudogenes, but rather that the divergence was driven by positive selection. This study provides strong evidence that the evolution of new or specialized functions may be a dominant mode in the maintenance of young duplications on the *D. melanogaster* X chromosome.

Future directions

In the future, I will continue this work using both computational and experimental methods, and it will be a main focus of my research in the immediate future. On the computational end, methods will have to be developed to identify duplicated loci using the 12 *Drosophila* genomes that are currently being sequenced. Of particular importance here will be controlling for the variable coverage in each of the genome projects. This project will allow the rates and patterns of gene duplication to be quantified. A complementary project will involve three avenues of experimental work. First, tiling arrays will be used to directly probe these genomes for duplications that may be absent from the publicly-available assemblies. Second, polymorphism data will be gathered from duplicated and single-copy genes in these species, to ask about the relative degree of selective constraint on amino acid sites in young genes. Third, expression technology which can distinguish closely related sequences (such as pyrosequencing) will be applied to study the rate at which expression patterns evolve in duplicate genes.

Evolution of Sex Chromosomes

Background

The work on gene duplicates described above demonstrates that gene families on the X and the autosomes evolve at different rates. Theory predicts that the X chromosome will accumulate more substitutions if adaptive mutations are recessive on average (the "fast X" hypothesis), while dominant beneficial mutations result in faster evolution of autosomes. To date, evidence for a general fast-X effect is equivocal or weak, largely due to limited data availability.

No evidence for a "Faster-X" effect in *Drosophila*

In collaboration with Doris Bachtrog and Peter Andolfatto (UC San Diego), I have obtained sequence data from and analyzed 200 pairs of orthologous loci from 4 *Drosophila* species (*D. melanogaster, yakuba, pseudoobscura*, and *miranda*). All 200 loci are X-linked in *pseudoobscura* and *miranda*, but 100 of them are autosomal in *melanogaster* and *yakuba*. We can use these data to test the fast-X theory by contrasting rates of evolution between closely-related

species for *the same* genes that are X-linked in one comparison, but autosomal in another. Specifically, we can ask whether the rate of amino acid substitution, relative to the rate of silent substitutions, is accelerated in lineages where loci are X linked compared to lineages where the same loci are autosomal. Our analysis reveals that there is no difference in rates of evolution between X-linked and autosomal loci, and no evidence for accelerated rates of evolution of genes with male-biased expression patterns on the X, providing evidence that adaptive mutations are not recessive on average and that there is no general fast-X effect, at least in *Drosophila*. When considered in light of the results for gene duplications (discussed above), these results show that the nature of adaptive substitutions differs between single copy and duplicated genes.

Future directions

The genome sequences for twelve *Drosophila* species are currently becoming available. These allow a unique opportunity to study rates and patterns of molecular evolution in a model organism. These data facilitate detailed comparisons of rates of evolution of sex chromosomes and autosomes. One question of particular interest will be the comparison of (putative) gene regulatory regions and coding sequences, as the fitness effects of substitutions in these two regions may have different properties. Where necessary, the genome sequence data will be augmented by obtaining further sequence data from close relatives of the twelve sequenced species, such as *D. miranda*, a sister species of *D. pseudoobscura*.

Statistical Inference from Population Genetic Data

Background

The question of what forces shape genomic variability in natural populations is the central question of population genetics. In particular, the question of how much natural selection, particularly adaptive evolution, has shaped genetic variation in the recent past, is of primary interest. With the complete genome sequences of several model genetic systems now available, it is natural to ask about the forces governing genetic polymorphism in these species. The inference of the action of natural selection from population data has traditionally come from assessing how well the data are explained by an idealized model of a constant-size, random-mating, population undergoing no selection. However, demographic departures from this model, especially changes in population size, can lead to departures from the standard null model that "look" like the action of natural selection. At the center of this work is the need for consistently sampled data [e.g. 3] and for computational approaches that appropriately account for biological processes like recombination [1,2].

Inferring demographic parameters in *Drosophila*

One avenue towards uncoupling the effects of selection and demography is to analyze consistently sampled multilocus datasets. As a first approach, I have been investigating the ability

of simple models of population bottlenecks to predict patterns of variability in natural populations in *D. melanogaster* [1,3, in collaboration with Peter Andolfatto at UC San Diego]. Through analyses of data from non-protein-coding regions, we have found that recent, severe, reductions in population size are compatible with data sampled from non-African populations.

Future Directions

In collaboration with Peter Andolfatto, this work will be extended in two directions. The first task will be to compare coding and non-coding regions of the genome, as patterns of genetic variability differ markedly between the two, and the reasons why are currently unclear. The second task will be to develop models of natural selection in populations with non-equilibrium demography, and use these models to infer parameters from data. These projects will require collection of polymorphism data from a large number of regions of the genome, sampled from appropriate natural populations, and the development of new computational methods for inference.