Petrov, Dmitri

**Background.**

Adaptation is the *central* evolutionary process and is at the core of some of the greatest challenges facing humanity. HIV would cause a mere fever without the ability of the virus to continuously adapt to the immune system. Cancer is a case of cellular adaptation within the body and would be much more straightforward to treat if not for its ability to adapt to anti-cancer drugs. Rather than killing hundreds of thousands of people every year, malaria could be treated with inexpensive drugs like quinine. We too often find ourselves in an evolutionary arms race without understanding the rules of engagement.

Until recently, adaptation was thought to be so idiosyncratic and infrequent that it seemed impossible to obtain enough empirical data about adaptive evolution to answer any of the above questions in a systematic way. However, we now know that adaptation is at times sufficiently rapid, recurrent, and involves mutations of large effect such that adaptive dynamics can be quantified on short time scales and with powerful statistical replication. Thanks to the development of genomic technology, we are finally in a position to study the process of adaptation in a comprehensive and statistically meaningful way.

Below I detail three interrelated projects in the lab. The first project focuses on the inference of adaptive dynamics and identification of targets and agents of adaptation from genomic data. In the second project we aim to map rapid seasonal adaptation from standing variation in Drosophila in order to understand the nature of fitness-related variation in populations. In the third, we use ultra-diverse molecular barcoding to track adaptation at unprecedented resolution in experimental evolution of yeast.

All three projects focus, to a large extent, on adaptation in large populations, where dynamics are not limited by the waiting time for adaptive mutations. Cancer and microbial infections fall within this adaptive regime (106- 1012 cells), making understanding of the dynamics of adaptation in large populations essential for conquering such diseases.

**I. Adaptation in the genome.**

**Recent progress.**

**1. Recurrent adaptation in Drosophila and humans is frequent and driven by strong selection.**

Adaptation leaves detectable signatures in genomes[1](#_ENREF_1). The rapid expansion of adaptive alleles in populations leads to (i) an excess of functional changes between species and (ii) distortions in the patterns of polymorphism known as selective sweeps. The signatures of selective sweeps, which include a reduction in levels of polymorphism and an excess of rare alleles, have been used by multiple groups to scan for adaptation in the genome[2](#_ENREF_2). My laboratory pioneered a distinct approach of quantifying the rate and the strength of adaptation by searching for a correspondence between genomic regions with reduced neutral polymorphism within species and increased functional divergence between species. This relationship is unexpected under the Neutral Theory but is naturally predicted if selective sweeps are common and driven by functional, adaptive substitutions[3-5](#_ENREF_3).

Our application of these methods in Drosophila confirmed that adaptation was frequent (an adaptive substitution in the genome every 50-100 generations) and led to the discovery that adaptation often involves mutations of large fitness effect (~1% or even larger)[3](#_ENREF_3). In humans, our initial claim of abundant positive selection[4](#_ENREF_4) proved controversial[6](#_ENREF_6). Our latest analysis[7](#_ENREF_7), however, confirmed that strong adaptation was indeed common in the recent human past. We estimate that on the order of 100-200 substantial adaptive changes took place in the past 100 thousand years of human evolution and further argue that many of the adaptive changes occurred in regulatory regions.

**2. Adaptation even by *de novo* mutation commonly leads to soft sweeps in Drosophila.**

Much of the current population genetic theory of adaptation by *de novo* mutation assumes the spread of a single adaptive mutation linked to a single haplotype leaving behind the signature of a hard selective sweep[8](#_ENREF_8). This assumption is based on the estimates of the effective population sizes (Ne) being much smaller than the reciprocal of the mutation rate per base pair. When this is the case, by the time the second adaptive mutation at the same site establishes in the population, the first one is already at such a high frequency that the second one has no chance to contribute to adaptation in a meaningful way[9](#_ENREF_9). Adaptation in this *weak mutation, strong selection* regime is slow (limited by mutation), with complex multi-step adaptations taking a particularly long time.

The work in my lab challenged this notion. We first focused on careful dissection of rapid evolution of organophosphate resistance at the *Ace* (acetylcholine esterase) locus in *D. melanogaster*[10](#_ENREF_10). We demonstrated that this three-site, complex adaptation arose in fewer than 1000 generations in local populations by multiple concurrently sweeping *de novo* mutations in a pattern known as a **soft selective sweep**[9](#_ENREF_9). These results are difficult to reconcile with the predictions of the weak mutation, strong selection regime. Instead, we proposed that on time scales relevant to rapid adaptation, i.e. tens to thousands of generations, Ne in Drosophila is orders of magnitude larger than previously thought (>108). We argue that the much smaller estimates of Ne (~106), inferred from the levels of neutral polymorphism, are driven primarily by rare but strong bottlenecks and intermittent genetic draft[10](#_ENREF_10),[11](#_ENREF_11). These rare events drive down the long-term equilibrium values of neutral polymorphism and estimates of Ne. Because of the separation of time scales, however, these low values cannot be used to predict dynamics of short-term rapid adaptation[12](#_ENREF_12).

Our work firmly established that rapid adaptation in Drosophila can leave signatures of soft selective sweeps, with multiple haplotypes rising to high frequency simultaneously. Our review of literature and theoretical considerations further suggested that soft sweeps are common in many species, especially those with large census population sizes[13](#_ENREF_13). Given that most current methods focus on detecting hard sweeps, we reasoned that genome scans might be missing an entire class of important adaptive events[13](#_ENREF_13).

Soft sweeps leave subtle signatures that are more difficult to detect than hard sweeps. We have successfully developed a suite of new statistics (H1, H12, H2/H1, and H) that rely on distortions in haplotype spectra to detect both hard and soft sweeps with similar power and distinguish them from each other[14](#_ENREF_14). We utilized these statistics to discover that soft sweeps are a dominant form of adaptation in *D. melanogaster*[14](#_ENREF_14). In collaboration with Tim Anderson, we successfully detected selective sweeps (two hard and three soft) at all of the five tested drug-resistance loci in the malaria parasite *P. falciparum*, validating the utility and power of these approaches. We continue to develop additional statistics to detect these largely ignored, but possibly dominant, signatures of adaptation.

**Future studies:**

1. **Statistical development.**

Genome divergence and polymorphism data should in principle allow us to make detailed inferences about the dynamics of recurrent adaptation in the genome. Inference procedures must account for the facts that (i) the dynamics of adaptation are complex, generating a mixture of partial and complete, hard and soft sweeps of varying selective strengths, (ii) that background selection has a pervasive effect on patterns of polymorphism, and (iii) that rates and patterns of mutation vary across the genome. We will approach this problem by primarily focusing on haplotype statistics of the kind we have developed (H, H12, H2/H1) and the ones that already exist (such as iHS[15](#_ENREF_15" \o "Voight, 2006 #184), nSL[16](#_ENREF_16" \o "Ferrer-Admetlla, 2014 #183), XPHH[17](#_ENREF_17)). Our work showed that haplotype statistics in general are not sensitive to background selection[7](#_ENREF_7) and differentially sensitive to hard and soft sweeps. We will combine analytical theory, powerful forward simulations, and empirical work with genomic data to develop statistical inference machinery able to make valid inferences in the face of known complexity of adaptation. Our forward simulations are made possible by extremely efficient program SLiM[18](#_ENREF_18" \o "Messer, 2013 #186) (**S**election with **Li**nked **M**utation) developed in my lab by Philipp Messer.

1. **Population genetics of adaptation at known targets.**

Too often, statistical methods are evaluated with only idealized *in silico* simulations that may or may not reflect the full complexity of natural selection and demography in the field. We will therefore continue investigation of population signatures of adaptation at positive-control loci that underlie resistance to drugs and pesticides. Beyond the obvious practical importance of understanding evolution of drug resistance, I also believe that it is essential to test population genetic methods with empirical controls. Our focus right now is on evolution of resistance at *Ace* and *Cyp6g1* (locus of DDT resistance) in multiple Drosophila species, and a number of known drug resistance loci in *P. falciparum*. We are also beginning to study evolution of resistance in HIV, both to single drugs and drug cocktails.

**3. Identification of agents of adaptation with a focus on viruses and pathogens.**

Our studies of adaptation in Drosophila and humans helped reveal that adaptation is pervasive. The identity of the key drivers of adaptive change remains mysterious, however. This mystery is further compounded by the wide variety of genes that show signatures of adaptive change. We are testing the hypothesis that adaptation is driven to a large extent by viruses and other pathogens. This hypothesis is attractive not only because it is well known that viruses exert powerful selective pressures, but also because they are known to interact with a wide range of cellular and physiological functions, and are thus capable of driving adaptation in a broad assortment of genes.

Preliminary results appear to validate this hypothesis in the mammalian lineage and in humans in particular. Using manual curation of biomedical literature, we identified ~1200 genes in the human genome that show defined interactions with viruses in low-throughput studies. We were able to show that these 1200 genes indeed adapt at a much higher rate than genes in the same functional (GO) categories that are not known to interact with viruses. The effects are not limited to immune genes and in fact are strongest in genes involved in central cellular functions. We will continue using these approaches to quantify the adaptive importance of various agents of selection in humans and other mammals, aiming to extend this approach to other organisms as well over the longer term.

**II. Adaptation of *Drosophila* to seasonal fluctuations.**

**Recent progress.**

**Discovery of polygenic adaptive oscillations over seasonal time scales in D. melanogaster.**

*D. melanogaster* goes through ~10 generations each growing season in temperate environments and contends with sharply and cyclically varying environments. In the winter, flies must survive prolonged periods of starvation, cold, and desiccation, while in the summer they are subject to strong selection for early and high fecundity. Previous studies revealed that, as a result, *D. melanogaster* populations undergo cyclic phenotypic evolution. For instance, *D. melanogaster* strains established in the spring are much more likely to undergo reproductive diapause when subjected to appropriate conditions in the lab (low temperature and short day) compared to strains established in the fall[19](#_ENREF_19).

In collaboration with Paul Schmidt from the University of Pennsylvania, we set out to determine the extent and nature of genomic response to seasonality in *D. melanogaster*. We sequenced pooled samples of ~100 *D. melanogaster* adults collected in June and late October/November of 2009, 2010, and 2011 at an orchard outside of Philadelphia. From ~5x105 common SNPs, we identified ~1500 SNPs that cycle reliably between spring and fall in this population (Fig. 1A)[20](#_ENREF_20). These SNPs are spread across the genome, tend to fall within genes, and are associated with key phenotypes such as chill coma recovery and starvation resistance. We further demonstrated that seasonal SNPs respond to a single incidence of frost by shifting frequency in the predicted, winter-like direction (Fig. 1A). Almost all seasonal SNPs are present in the ancestral, sub-Saharan African populations and are found at a higher than expected proportion to be segregating in the sister species *D. simulans*[20](#_ENREF_20" \o "Bergland, 2014 #189).

These results suggest that (i) adaptation of *D. melanogaster* to shifting environments can be rapid and highly polygenic and (ii) that consequential genetic variation can be maintained for long periods of time despite or maybe even because of rapidly fluctuating directional natural selection. The strength of selection experienced by individual seasonal alleles appears to be substantial, capable of shifting their frequencies by ~20% in the course of about 10 generations.

We also argue that even though most segregating polymorphisms are likely to be neutral or nearly so, a substantial number of polymorphisms with strong fitness effects are maintained by balancing selection. Moreover it is these balanced alleles that are responsible for the ability of populations to adapt to rapid environmental shifts. Understanding of the identity, function, and evolutionary history of these balanced polymorphisms should shed much light on the process of adaptation in general.

|  |
| --- |
| Macintosh HD:Users:dmitri:Library:Mail Downloads:three_year_post-2.png |
| Figure 1.Seasonal SNPs in *D. melanogaster*  A. The seasonal cycling of the identified SNPs and their response to the first frost of 2011. B. The temperature profile at the orchard in 2011. |

**Future studies:**

**1. Direct experimental studies of tradeoffs in the lab and in the field.**

We are following up the studies in the field with direct experimental analyses. In collaboration with Paul Schmidt, we have set up large population cages both in the lab and in the field to test our hypothesis that summer-favored alleles are beneficial during periods of abundance, while winter-favored alleles are favored during periods of privation. Both sets of cages were seeded with flies from a large outbred population started from ~150 isogenic DGRP[21](#_ENREF_21) strains. We are allowing populations in the indoor cages to expand to the size of ~800,000 individuals over four generations with abundant food (mimicking periods of abundance/summer) and then taking a sample of flies to seed cages for additional serial transfers, allowing the rest to die without food (mimicking periods of sharp privation/winter). By sequencing pools of flies through the growth and starvation, we seek to map variants that exhibit fitness tradeoffs between these environmental conditions. Our goal is to test whether seasonal SNPs exhibit hypothesized tradeoffs in manipulated conditions. Our mapping leverages knowledge of the haplotypes that initiated the cages, thereby allowing us to measure SNP frequency with high precision for a small fraction of the usual cost.

**2. Rapid evolution across time and space.**

We are continuing sampling flies (*D. melanogaster* and *D. simulans*) across seasons in multiple populations both in the US and in Europe. To do so, we have organized a consortium of Drosophila researchers to collect flies near their home institutions and send us samples. We now have two years of seasonal collections across eight different sites (Wisconsin, New York, Massachusetts, Virginia, Pennsylvania, Maine, Vienna, Barcelona) in addition to three years of collections at four sites in California. We are planning to sequence these samples to assess the predictability of seasonal adaptation across space and environmental conditions. Our plan is to continue seasonal collections across as many sites as possible for the foreseeable future. Such extensive sampling is essential in order to separate the true signal of adaptation from the noise driven by spurious linkage and to be able to associate signatures of selection at the molecular level with specific environmental variables.

**3. Phenotypic and molecular effects of seasonal SNPs.**

We are beginning to investigate phenotypic and molecular effects of the most robust seasonal SNPs. The current approach involves setting up cages that are fixed for the alternate versions of one specific seasonal SNP and measuring multiple phenotypes of the flies in the cages. Each cage starts with a large number of progenitor strains from DGRP[21](#_ENREF_21), ensuring that the effect of each SNP is measured in an outbred population on a diverse genetic background. Preliminary data are extremely promising with every one of the three tested SNPs showing predicted phenotypic effects. For instance, flies with the winter-preferred SNP in the TRP channel[22](#_ENREF_22), known to be involved in temperature perception, show strong preference for colder temperatures compared to the flies with the summer-adaptive SNP, but do not show any other detectable phenotypic effects. We will continue investigating molecular and physiological mechanisms by which seasonal SNP mediate their phenotypic effects. Our next step is to begin incorporating seasonal SNPs into the constant genetic background using Crispr-Cas9 technology, which has the potential to increase experimental precision and allow us to study epistasis among multiple seasonal SNPs.

1. **Theoretical studies of strong and cyclical polygenic adaptation.**

Our data suggest that highly polygenic fitness-related variation can cycle in response to selection by as much as 20% while simultaneously being maintained in the population for millions of generations. We are carrying out theoretical studies aimed at understanding the plausibility of these patterns and the conditions under which they may be expected. Preliminary data suggest that even moderately large populations can indeed maintain variation of this type without incurring problems with “costs of selection” or “genetic load”, provided a specific form of dominance-by-epistasis interactions in fitness among seasonal SNPs. The future work will specifically focus on making predictions about the dominance-by-epistasis interactions among seasonal SNPs that we can test in the lab and in the field.

**III. Studies of experimental evolution in yeast using ultra-high resolution barcodes.**

**Recent progress.**

1. **Development of high-throughput barcoding in experimental evolution.**

Without the ability to systematically document a large number of adaptive events, we are doomed to draw our conclusions from a few anecdotal cases. Experimental evolution allows us to replay the “tape of life” with extreme replication, and should in principle provide statistically powerful samples. Unfortunately, even experimental evolution was by and large limited by relatively small total numbers of mutational events that can be studied in parallel and by the focus on the long-term winners.

To address these current limitations, we are harnessing the power of experimental evolution, next-generation sequencing, and an ultra high-resolution molecular barcoding system (up to 1010 unique barcodes) that we recently developed in collaboration with the labs of Gavin Sherlock and Daniel Fisher. In our first proof-of-principle experiment, we placed 5x105 molecular barcodes at a single site in the yeast genome and evolved the culture by serial transfer in glucose-limited media[23](#_ENREF_23). At every transfer, we sequenced the barcodes at very high coverage and followed the trajectories of individual lineages at an unprecedented level of resolution.

|  |
| --- |
|  |
| Figure 2.Trajectories of ~500,000 molecularly tagged yeast genomes in an evolution experiment. Red lineages increase in frequency beyond what one would expect by chance. |

Figure 2 shows the resulting trajectories. The red lines mark lineages that show evidence of adaptation because their frequencies increases beyond what is expected by random drift. A second replicate experiment showed very similar evolutionary trajectories and overall fitness increase, suggesting that adaptation in this system is indeed predictable in a gross statistical sense.

Using this experimental design, we were able to show that the rate of adaptive mutagenesis is very high and that large-effect adaptive mutations are common. **On the order of 25,000 lineages obtain adaptive mutations and increase in frequency, but only 12 increase in frequency beyond 1%.** These “winning” lineages, which are the general subject of experimental evolution studies, must have obtained their adaptive mutations early on in the experiment. In addition, our sequencing study revealed that these lineages also fortuitously managed to obtain a series of adaptive mutations in quick succession.

1. **Isolation of a large number of individual adaptive mutations and remeasurement of their fitness.** Our experimental design allowed us to identify beneficial clones early in the evolution (~80-100 generations) when beneficial mutations are collectively common but individually rare. At this stage, only ~30% of isolated clones should have acquired any neutral mutations and even fewer should harbor deleterious mutations. We sequenced several hundred clones isolated at generation 88, finding support for these predictions. Most clones contain only single mutations, with the vast majority hitting genes in the *Ras* pathway.

**Future studies.**

**1. Evolution in diploids.**

Most work in experimental evolution has traditionally focused on haploids. This is unfortunate, as it is likely that adaptation in diploids proceeds in a qualitatively different fashion and is driven by a drastically different set of mutations. Adaptive mutations in diploids must be at least partially dominant in fitness (to pass through the so-called Haldane sieve), and we have predicted that adaptive mutations in diploids should often be overdominant in fitness generating widespread heterozygous advantage[24](#_ENREF_24).

We will use our molecular barcoding system to isolate hundreds of individual adaptive mutations in diploids, allowing us to conduct precise measurements of fitness effects in heterozygotes and homozygotes, generating the first comprehensive joint distribution of fitness effect, dominance, and molecular identity of diploid adaptive mutations.

**2. Pleiotropy and epistasis among adaptive mutations.**

We are carrying out experimental evolution in 12 distinct environments and remeasuring the fitness of all adaptive clones across all of the environments. Our experimental design will use distinct barcode sets for each experimental evolution allowing us to carry out all-by-all (12x12=144) fitness measurements in just 12 experiments. This project will produce the first comprehensive map of fitness by environment interactions among specifically adaptive mutations. We expect that most adaptive mutations will be adaptive in some but not all conditions. This map will help us make predictions about adaptation in variable environments and help us investigate the detailed molecular and cellular determinants of fitness benefits and costs of a large number of adaptive mutations.

**3. Epistasis among adaptive mutations.**

In collaboration with Mark Siegal at NYU we will investigate prevalence of epistasis among adaptive mutations. The first experiment will focus on the protein-folding chaperone, Hsp90, because it interacts with a large portion of genes, buffering the effects of most mutations[25](#_ENREF_25) but enhancing the effects of adaptive mutations[26](#_ENREF_26). Theoretical models predict inhibiting Hsp90 will impede adaptive evolution, and suggest this may be a strategy to prevent the evolution of drug-resistance in cancers[27](#_ENREF_27). We are carrying out experimental evolution in yeast with inhibited Hsp90 in order to test how this changes the rate of adaptation and the magnitude and identity of adaptive mutations.

**IV. Conclusions.**

Rapid advances of genomic technologies combined with the recent realization that adaptation is often rapid, pervasive, and driven by large-effect mutations have the potential to revolutionize our understanding of evolution. My lab has been at the forefront of this revolution. From quantifying adaptation in genomes using computational and theoretical approaches, to understanding evolutionary history and phenotypic impact of polymorphism underlying rapid adaptation in contemporary populations, to carrying out evolutionary experiments at the highest level of resolution and replication ever done, all the projects in our lab aim to build a theoretically and empirically rich theory of the evolutionary adaptation.