Tanja Pyhäjärvi

Population genomics of *Zea mays* spp. *parviglumis* – natural selection from the species to the genome

Site of research: Department of Plant Sciences, University of California, Davis, USA

Table of Contents

1 BACKGROUND

1.1 Genome-wide patterns of natural selection

Patterns of molecular genetic diversity among and within species are shaped by natural selection, but also by neutral evolutionary forces such as mutation, recombination and genetic drift. Since Darwin (1), one of the main aims in evolutionary biology has been to understand the relative importance of selection in patterning genetic and phenotypic diversity. While Kimura's (2) neutral theory, claiming most of the observed molecular polymorphism to be neutral, has been the null hypothesis in molecular evolutionary genetics for decades, in recent years there has been growing interest in and evidence for purifying and adaptive selection in the genome (3). As molecular techniques and genomic resources have advanced, scientists are now for the first time able to estimate – sometimes to the level of individual nucleotides – which parts of the genome have been affected by purifying and adaptive selection, and what fraction is neutral (4, 5).

Currently, there are only two model genera – *Drosophila* and *Arabidopsis* – for which both interspecific and intraspecific data on genetic variation is abundant enough to elucidate the role of selection at a genomic scale (6). Results from *Drosophila* are striking: more than 90% of new mutations at non-synonymous sites were targets of purifying and up to 50% of amino acid substitutions were estimated to be adaptive (4, 7). Moreover, a considerable percentage of noncoding regions that have previously been considered neutral seem to experience both purifying and adaptive selection (4). In *Arabidopsis,* however, the pattern is somewhat different: purifying selection appears weak and adaptive substitution rates are lower than in *Drosophila* (6, 8-10). Mating system, population structure and differences in effective population size have been suggested to explain these differences (8, 10), but data from more species are necessary to untangle the effects of these biological factors on genomic patterns of selection.

Most studies on the genome-wide effects of natural selection are based on species-wide sampling (11). However, demographic history as well as patterns of selection may vary among populations (e.g. 12). Also the gene flow and hierarchical structure among populations may result in processes that cannot be detected from species-wide sample assumed to represent panmictic population (13).

Geographical variation in selection pressures (i.e. differences in environment) can lead to populations specifically adapted to local conditions, i.e. that show the highest relative fitness at their original site (14). Understanding the genetic bases of local adaptation offers particular insight into evolutionary mechanisms, because such adaptations are relatively recent, and in many cases the environmental differences driving adaptation are still observable (15).

As data on nucleotide diversity continue to accumulate, it has become clear that natural populations are far from the panmictic, constant size population that is assumed in many models of molecular evolution. Due to biotic and abiotic environmental changes such as climate and inter-specific competition, species' distributions and effective sizes fluctuate, which in turn changes the relative contributions of drift and selection on molecular diversity. For example, population bottlenecks may increase the variance of polymorphism among loci, leading to patterns that can be mistaken for positive selection (16). On the other hand, hierarchical population structure may cause excess variation in measures of genetic differentiation, potentially mimicking the effects of local adaptation (17). It is thus imperative to understand the demographic history of a population and its connection to other populations before making conclusions about the effects of natural selection on genomic variation.

1.2 *Zea mays* spp. *parviglumis* as a model plant for population genomics

Zea mays spp. *parviglumis* (hereafter *parviglumis,* also called “teosinte”) is the closest wild relative of domesticated maize (*Z. m.* spp. *mays*). It is a highly outcrossing annual grass endemic to south-central Mexico. It is found in two main regions in the states of Jalisco and Guerrero, growing in often dense populations at moist middle-elevations (400-1000 m) (Fig. 1). Populations vary greatly in a number of environmental characteristics (exposure, slope, competition) (M.B. Hufford, R. Miranda-Medrano and P. Gepts, unpublished data) and show considerable morphological and phenological variation (18). Previous studies suggest strong genetic structure among populations and non-equilibrium demographic history at both the species and population level (19-21). Recent work has also found signs of local adaptation in the immune gene *wip1*, suggesting differences in pathogen pressure among populations (22).

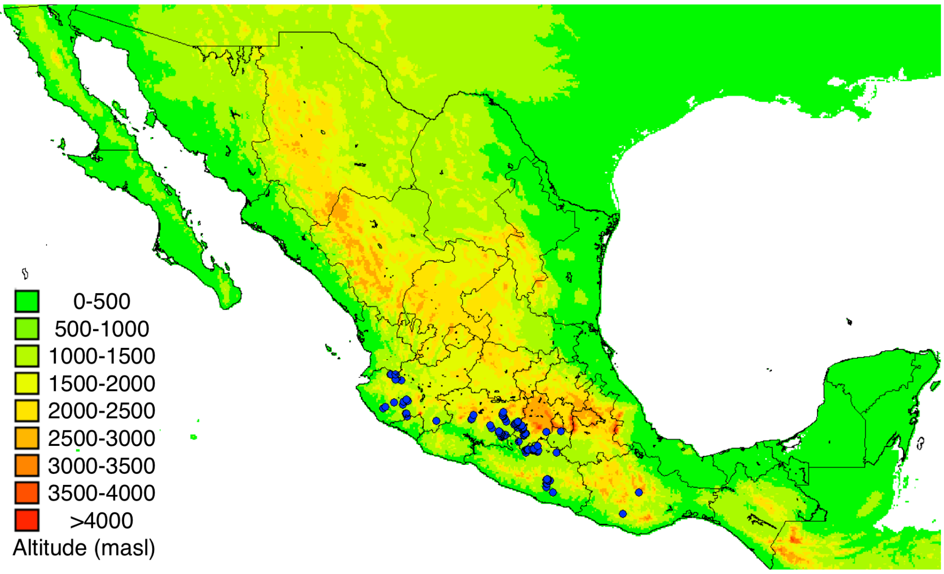

Figure 1 Map of sampled populations (blue dots) of *Zea mays* spp. *parviglumis*

Parviglumis is an ideal organism to study the effects of natural selection on genomic variation across multiple scales. Numerous wild populations enable the study both of differences in selection pressure among locations as well as the effects of distinct demographic histories. Because it is an endemic wild taxa, *parviglumis* is more representative of other wild plants than *A. thaliana*, which is a human commensal that has only recently spread worldwide. As an outcrossing species with high effective population size and strong population structure, it can be used to compare the relative roles of demography and selection. High effective population size is expected to facilitate species-wide selective sweeps even for weakly selected alleles, whereas strong population structure might lead to local adaptation specific to individual populations. In addition to these biological characteristics, the ability to make use of the new maize genome sequence and the array of state-of-the-art genomic tools that have been developed for maize is an enormous advantage for studying *parviglumis*. Moreover, a number of other grass genomes are available, facilitating comparative genomics approaches as well as evaluation of longer-term patterns of adaptive and purifying selection. In brief, *parviglumis* provides a unique opportunity to examine factors affecting molecular evolution on multiple scales, from the range-wide variation in diversity to nucleotide differences within the genome.

1.3 Previous research

For my PhD thesis, I studied the relative roles of demography and selection in several populations of an outcrossing, woody perennial, *Pinus sylvestris* (Scots pine). Levels of nucleotide diversity at 16 nuclear loci in Scots pine point to the existence of an ancient population bottleneck which has had a strong impact on diversity across the genome (23). Mitochondrial haplotype diversity further reveals an important role for post-glacial colonization history in effecting the current the distribution of molecular diversity (24). There were no obvious signs of selection, suggesting that the visible effect of selection is weak compared to the variance among loci induced by population demography. The main conclusion in my thesis was that patterns of neutral variation in Scots pine are dominated by population history and, due to long-generation time, Scots pine is unlikely to ever reach mutation-drift equilibrium. However, this does not prevent natural selection being very effective on adaptive mutations, because the current census size of the species is very large.

During my current postdoc period, I have studied nucleotide diversity of *Arabidopsis lyrata*, an outcrossing perennial relative of *A. thaliana*. I have inferred times of divergence and amount of gene flow among natural populations of *A. lyrata*. Together with my colleagues, we have also discovered strong evidence of the selective sweep caused by local adaptation in *phytochrome A* in a Northern Norwegian population of *A. lyrata*.

My research both on *P. sylvestris* and *A. lyrata* have dealt with the roles of population structure, demography, and selection in natural plant populations, with special focus on local adaptation. Working on the same research questions in *parviglumis*, a species with more genomic resources available, will enable more detailed analysis, the experience on working with very different types of plants (gymnosperm, monocot, dicot) will give me a broad view on plant molecular evolution.

2 OBJECTIVES

Quantifying the relative roles of demography and natural selection at the population level is essential to understanding the mechanisms of evolution at the molecular level. I propose to investigate these questions by studying the demographic history, genetic structure and evidence of natural selection in natural populations of the wild ancestor of maize, *parviglumis*. This project has three main research questions: 1) How much does demographic history affect molecular diversity at the population scale and how much does it differ among natural populations of a single plant species? 2) How does geographic variation in selection pressure affect the distribution of genome-wide molecular variation among populations? 3) How common is positive and negative selection at a species-wide level and what genomic regions have been the targets of selection throughout the species distribution? The project will make use of three unique datasets, ranging from species-wide sampling with molecular markers to genome-wide resequencing of a sample of individuals, to pursue each of these objectives.

Detailed goals are to:

1. a) Investigate patterns of genetic structure among populations
   b) Infer demographic histories of natural populations of *parviglumis*

2. a) Assess evidence for local adaptation among six populations of *parviglumis*
   b) Identfiy the types of genes and traits affected by local adaptation
   c) Investigate associations between genotype, phenotype and environmental factors varying among *parviglumis* populations.

3. a) Investigate patterns of purifying and adaptive selection in the *parviglumis* genome
   b) Determine the strength and rate of species-wide selective sweeps caused by natural selection

3 MATERIAL AND METHODS

3.1 SNPs and genome-wide resequencing

This study will make use of both single nucleotide polymorphisms (SNPs) and next-generation (Solexa) genome-wide resequencing. A SNP is a single base pair position known to be polymorphic in a population or collection of samples. SNPs are efficient genetic markers for examining patterns of genetic diversity at genome level. With chip-based technology, tens of thousands of SNPs can be genotyped quickly and inexpensively. Modern SNP genotype data is high quality and requires little manual editing. Next-generation sequencing-by-synthesis methods (25) produce millions of short sequence reads that can be mapped to a reference genome, enabling the inexpensive re-sequencing of very large fractions of individual genomes.

Using both types of data allows me to utilize the advantages of both. Genomic variation in large samples of individuals and populations is most effectively captured via SNP genotyping. On the other hand, resequencing data allows in-depth coverage of individual plants genomes, allowing for analyses requiring information from contiguous sequence regions or differentiation among genomic regions or types of sites. Moreover, the amount of data generated from such studies provides considerable power to detect even subtle patterns of genetic variation.

3.2 Datasets

The data for this project consist of three partially overlapping datasets:

Dataset 1. Nearly 1000 individuals of parviglumis *genotyped for nearly 1000 SNPs*

In total, 934 individuals from natural populations covering the known range of *parviglumis* (Fig. 1) have been genotyped for 983 SNPs representing both candidate (18, 26) and random loci. Sampling includes 34 populations represented by 20-30 individuals as well as smaller samples of 1-8 individuals from >50 additional locales. Genotyping was performed by collaborator John Doebley using the Sequenom MassARRAY System (27). This dataset is ready for analysis.

Dataset 2. Six parviglumis populations genotyped for 100 000 SNPs

100 000 SNPs will be genotyped for ~10 individuals each from six of the above populations of *parviglumis*. The SNP genotyping will be conducted using the Infinium HD genotyping platform (Illumina Inc.), at the UC Davis Genome Center. The chip is designed for maize and contains SNPs predominantly from genic regions of the genome. The chip contains on average two polymorphic markers in each annotated gene in the maize genome. Because maize and *parviglumis* have abundant shared variation (21), the SNPs on the chip are informative for *parviglumis*. Work collecting these data has already begun and the data will be ready by the start of my postdoc period.

Dataset 3. Haplotypes of 16 inbred parviglumis lines

16 inbred *parviglumis* lines, sampled from across the range of the taxa (12 from Balsas, three from Jalisco and one from Oaxaca), will be used to create a haplotype map of the genic (~25%) fraction of the *parviglumis* genome. Inbred individuals have reduced intra-individual variation, which eases the mapping of reads. Methylation sensitive restriction allows targeting sequencing to low-copy, genic regions of the highly repetitive genome (28). The raw data consists of millions of short reads that will be mapped to the maize reference genome (B73), generating gigabases of sequence data and >1.5 million SNPs (5). Ed Buckler's group (www.maizegenetics.net) has already finished sequencing, and the haplotype map of these data should be submitted for publication and the data available by the start of my postdoc period.

Information management plan

All the data will be uploaded to www.panzea.org or www.rilab.org by the time of publication. The Datasets 1 and 2 will be deposited in dbSNP archive and Dataset 3 to the short read archive, both at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). Plant material or DNA will be available either from the United States Department of Agriculture (USDA) or from Ross-Ibarra's lab upon request.

3.3 Statistical methods

Objective 1. Genetic structure and demography.

The overall genetic structure among *parviglumis* populations will be investigated by principle component analysis (PCA) (29). In PCA, covariance among genotypes is described by uncorrelated variables, called principal components (PCs). PCA is applicable for large samples sizes and can analyze structure on the level of individual genotypes rather than population frequencies. As an output, individuals are distributed along the axes of variation (Fig. 2). PCA is an objective method of describing population structure which does not require *a priori* information on population assignment, but is also well-rooted in population genetic theory (30).

Historical demographic processes, such as population size changes and gene flow between populations, leave a signature in genomic variation, seen as distortions from patterns expected under a standard neutral equilibrium (SNE) population. Deviations from expectations of the allele frequency distribution, variation among loci, or amount of linkage disequilibrium (LD, non-random association of alleles), contain information about past demographic changes (16, 31). I will use SNP data to infer population demography and the extent of gene flow among different *parviglumis* populations via Approximate Bayesian Computation (32). In brief, ABC consists of 1) summarizing genetic variation by statistics describing nucleotide diversity, the allele frequency spectrum, and population structure 2) conducting coalescent simulations under various demographic scenarios 3) comparing the observed data to that produced by simulations to infer posterior distributions of demographic parameters or choose among models. The ABC method has been succesfully applied to infer demographic history from polymorphism data in maize (21), *A. lyrata* (12) and *Populus tremula* (33).

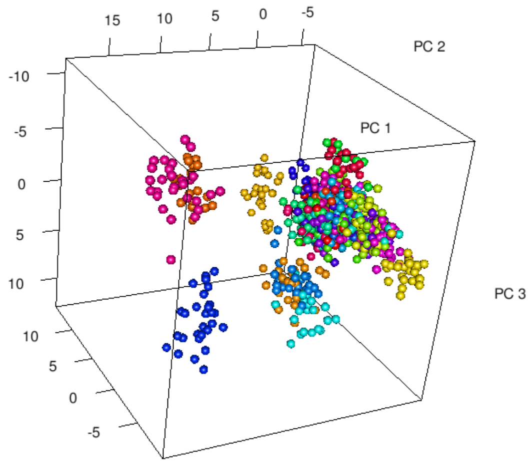

Figure 2 PCA of 34 *parviglumis* populations showing separation based on the first 3 PCs

Although bias in the choice of SNPs used on genotyping chips can impact demographic inference (34), I will match the simulated data according observed patterns to correct for this ascertainment bias.

Objective 2. Local adaptation

Local adaptation is expected to cause genetic differentiation in loci at or near the target of selection among populations under different selection pressures (35). First I will identify candidate loci for selection among the 100 000 SNPs by using F_{ST} outlier -method (36). The candidate loci are identified by comparing the observed F_{ST} values among 6 populations of Dataset 2 to simulated F_{ST} distributions. The neutral distribution of F_{ST} values is generated by coalescent simulations under the demographic scenarios inferred in Objective 1. This helps to tease apart deviations in diversity patterns caused by geographic variation in natural selection from those caused by demographic effects.

When a new, beneficial allele emerges in the population, its frequency rises as a result of natural selection. Due to LD, the beneficial allele drags nearby genomic regions to fixation with it. As a result of this ”hitchhiking”, genetic variation is lost in the region (37). This selective sweep is expected to reduce the level of polymorphism, change the pattern of LD and alter the allele frequency spectrum near the target of selection (38, 39). Thus, the second way to detect loci behind local adaptation is to find evidence of selective sweeps *within* population. To accomplish this I will use composite likelihood tests of the site frequency spectrum of SNPs (40), which have been shown to have high power to detect hitchhiking events and yet are relatively robust to demography and allow correction for ascertainment bias.

After identifying candidate SNPs for local adaptation, I will analyze what kind of genes have been the putative targets of selection. Because the position of each of these SNPs in the maize genome is already known, they can be identified as synonymous, nonsynonymous, or noncoding variants and information can be gleaned about the function of the gene in which each SNP lies. I will use this information (e.g. http://www.geneontology.org/) to investigate whether candidate genes are enriched for loci involved in particular plant functions such as immunity, drought tolerance, or flowering time. Finally, I will also investigate whether there are correlations between candidate SNP frequencies and environmental variables among populations in these loci.

For most *parviglumis* populations in this dataset, seeds are available for further experiments. Depending on the results of the Objective 2, we will investigate traits and phenotype-genotype associations that have putatively been under geographically divergent selection. For example, if genes affecting flowering time seem to be present among outlier loci, it is feasible to investigate whether population phenologies differ in a controlled common garden setup.

Objective 3. Patterns of selection across the genome

The effect of selection on genomic variation can be examined by comparing levels of polymorphism within species and divergence between species in neutral versus putatively non-neutral sites (41). The McDonald-Kreitman (MK) test is a versatile approach that accounts for variation in mutation rate among loci and is robust to demography. Under a neutral model of evolution, the ratio of polymorphism and divergence should be approximately equal at different sites. In theory, positive selection results in relatively higher divergence between species at selected sites than at neutral sites relative to polymorphism within species. In contrast negative or purifying selection would appear as deficiency of divergence relative to polymorphism at selected sites.

Resequencing Dataset 3 will be used to study patterns of selection across the genome. The MK approach is used to estimate the fraction of deleterious and adaptive alleles among new mutations at a species-wide level. Polymorphism among the 16 inbred *parviglumis* lines will be compared to divergence between *parviglumis* and the sorghum genome. Diversity and divergence in non-synonymous and different classes of non-coding sites (UTRs, introns, intergenic regions) will be compared to synonymous sites to identify targets of positive and negative selection (4).

In theory, positive selection results in a correlation between recombination, level of polymorphism and amino acid divergence along the genome. Selective sweeps are expected to reduce polymorphism more in regions of low recombination, and in regions where positive selection has been common, amino acid divergence is expected to be high. These relationship will be used to estimate frequency and strength of the selective sweeps (42-44).

3.4 Ethical issues

This project involves only *in silico* analysis. No special permits are required to conduct these analyses, and no specific ethical issues are involved.

4 IMPLEMENTATION

4.1 Schedule

The workflow of the project will follow the division of objectives (Fig. 3). I will start by analyzing genetic structure and demographic history (Objective 1) based on Dataset 1. Results on population structure and inferred demographic models from Objective 1 will be utilized in finding loci for local adaptation (Objective 2). I have reserved some time for common garden experiments, but the exact experiments depend on the results from Objective 2. Genome-wide resequencing data requires editing, cleaning and critical examination before in-depth selection analysis. I have reserved a relatively long time for preparing the data for Objective 3, because the dataset is large, and I need to learn to use many new bioinformatic tools and programming skills.

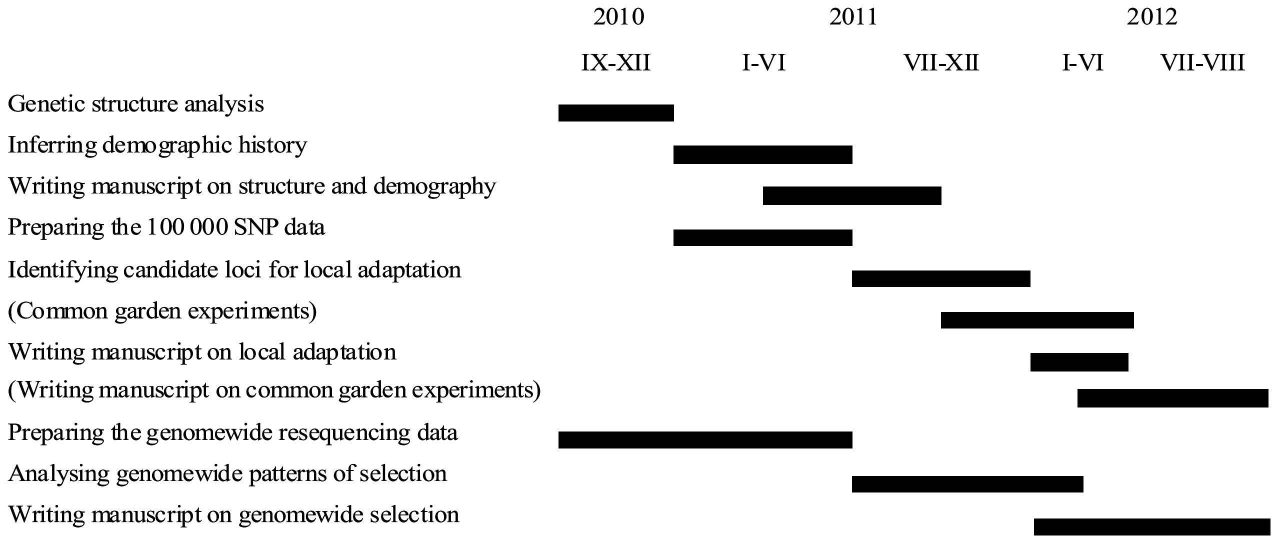

Figure 3 Timeline of workflow

4.2 Financial plan

I plan to attend annual maize meetings, which costs about 500€. My family (husband and 1 year-old daughter) would follow me to the location of postdoc period, which will increase the cost of travelling, accomodation (rent about 1000$ per month) and health-insurance. Health-insurance for the whole family is about 900$ per month. The part of insurance that is for postdoc alone is about 300$ per month and will be mostly covered by Dr. Jeffrey Ross-Ibarra.

Table 1 Financial Plan (numbers in €)

	Year		
	2010	2011	2012
Grant (T. Pyhäjärvi)[1]	14260	42780	28520
Travel[2]	5000		5000
Other expenses[3]	500	1000	500
Total	19760	43780	34020
Both years			97560

[1] Includes 15% doctoral raise, 20% family raise and 20% high-cost raise

[2] Estimated Oulu-Davis / Davis-Oulu one-way trip for 2 adults and 1 child

[3] Other upcoming expenses (e.g. literature, conference fees)

5 RESEARCHERS AND RESEARCH ENVIRONMENT

The prospective research will be conducted in University of California, Davis (UC Davis). My postdoc advisor, Dr. Jeffrey Ross-Ibarra, is a new assistant professor in the Department of Plant Sciences. The department is a great research community comprising ~90 faculty and ~150 students. Dr. Ross-Ibarra has quickly launched his own research group, studying the evolutionary genetics of adaptation in plants, focusing on the study of plant domestication and the evolution of crop plants. Dr. Ross-Ibarra's lab combines an understanding of plant population genetics with computation and bioinformatic analyses, exactly matching my research interests. Among other projects in plant evolutionary genetics, he has collaborations with top scientists working on maize genetics (Jim Birchler, Ed Buckler, R. Kelly Dawe, John Doebley, Brandon Gaut, Jiming Jiang, Gernot Presting). At the moment he has funding from USDA and University of California Institute for Mexico and the United States. He already published widely on plant population genetics, including in high-level journals such as Science and PNAS. He received the Dean's Award for Postdoctoral Excellence in UC Irvine in 2008 for his work in Brandon Gaut's lab.

As of January 2010, there will be two postdocs working on maize and teosinte population genetics and evolution in the lab, which will allow for considerable collaboration and discussion on research methods and development of bioinformatics tools. In addition, there are two undergraduate students, one graduate student and one visiting scientist in the group. People in the lab are enthusiastic, hard-working and work together in relaxed atmosphere. The lab has regular twice-a-week journal clubs and joint meetings with other research groups. Dr. Ross-Ibarra has an open door policy, encouraging interaction on a daily basis. The communal office space is shared with Dr. David Neale’s group, including several bioinformaticians and programmers, which adds additional possibilities for communication, interaction and learning of bioinformatic methods. Among standard wetlab resources, office space, etc., his lab offers computing facilities necessary for extensive bioinformatic analyses: a 128 cpu cluster with access to additional 384 cpu. Extensive field and greenhouse facilities are available at the department and college. Dr. Ross-Ibarra and his current postdoc Joost van Heerwaarden already have, in their first six months of collaboration, one paper in press (45) and two papers in review, reflecting the productivity of the work environment and the emphasis put on postdoctoral training.

UC Davis is the leading plant research university, with plant biology departments in both the College of Agricultural and Environmental Science and the College of Biological Sciences. The university possesses ample resources for molecular population genetics: the Genome Center provides state-of-the-art core facilties for bioinformatics, genomics (next-generation sequencing), metabolomics etc. The Center for Population Biology combines scientists and approaches from various fields (molecular, experimental, computational) to gather understanding on fundamental ecological and evolutionary processes. The Center also has one of the world's top evolutionary biology graduate programs, the Population Biology Graduate group, which includes a number of renowned evolutionary geneticists (Charles Langley, David Begun, Graham Coop, David Neale, Bruce Rannala, etc). In addition, there are several active seminar series and journal clubs on topics covering plant biology, genetics and genomics.

6 RESULTS

This study will broaden the insight on the effect of natural selection on genome-wide patterns of molecular variation that has so far been studied in detail in only a handful of species. The study provides knowledge on how different evolutionary forces affect outcrossing species with large population sizes and strong population structure. We will obtain estimates of how common deleterious and adaptive mutations are and how selective sweeps have affected patterns of diversity and divergence in the *parviglumis* genome. This study will be among the first to incorporate genome-wide data to investigate natural selection at both a local and species-wide scale. With these data we will be able to evaluate the importance of population-specific processes and evaluate whether local selective pressures outweigh species-wide selection in structured populations. Genome-wide analysis of local adaptation will reveal what types of genes have been the targets of selection and potentially provide abundant material for further, detailed analyses of how adaptations emerge as a response to environmental differences.

The processes of evolution and natural selection at a molecular level are of interest to a wide audience, and the manuscripts will be submitted to widely-read journals such as Genetics and Molecular Biology and Evolution. In addition, knowledge on population history of *parviglumis* will provide a more detailed background for studies on maize domestication, and loci that result in genetically locally adapted populations of *parviglumis* may provide new material for maize breeding.

7 REFERENCES

1. Darwin C (1859) *The Origin of Species,* (Bantam Books, New York).
2. Kimura M (1983) *The neutral theory of molecular evolution,* (Cambridge University Press, Cambridge).
3. Nei M (2005) Selectionism and neutralism in molecular evolution. *Mol Biol Evol* 22:2318-2342.
4. Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila. Nature* 437:1149-1152.
5. Gore MA*, et al* A first generation haplotype of maize. *Science*. In press
6. Wright SI & Andolfatto P (2008) The impact of natural selection on the genome: Emerging patterns in *Drosophila* and *Arabidopsis. Annu Rev Evol Ecol Syst* 39:193-213.
7. Sella G, Petrov DA, Przeworski M & Andolfatto P (2009) Pervasive natural selection in the *Drosophila* genome? *PLoS Genet* 5:e1000495.
8. Bustamante CD*, et al* (2002) The cost of inbreeding in *Arabidopsis. Nature* 416:531-534.
9. Wright SI, Lauga B & Charlesworth D (2002) Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis. Mol Biol Evol* 19:1407-1420.
10. Foxe JP*, et al* (2008) Selection on amino acid substitutions in *Arabidopsis. Mol Biol Evol* 25:1375-1383.
11. Wright SI & Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* 22:506-519.
12. Ross-Ibarra J*, et al* (2008) Patterns of polymorphism and demographic history in natural

populations of *Arabidopsis lyrata. PLoS One* 3:e2411.

13. Staedler T, Haubold B, Merino C, Stephan W & Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in non-equilibrium subdivided populations. *Genetics* 182:205-216.
14. Kawecki TJ & Ebert D (2004) Conceptual issues in local adaptation. *Ecol Lett* 7:1225-1241.
15. Savolainen O, Pyhäjärvi T & Knürr T (2007) Gene flow and local adaptation in trees. *Annu Rev Evol Ecol Syst* 38:595-619.
16. Depaulis F, Mousset S & Veuille M (2003) Power of neutrality test to detect bottlenecks and hitchhiking. *J Mol Evol* 57:S190-S200.
17. Excoffier L, Hofer T & Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* 103:285-298.
18. Weber A*, et al* (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics* 177:2349-2359.
19. Moeller DA, Tenaillon MI & Tiffin P (2007) Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics* 176:1799-1809.
20. Fukunaga K*, et al* (2005) Genetic diversity and population structure of teosinte. *Genetics* 169:2241-2254.
21. Ross-Ibarra J, Tenaillon M & Gaut BS (2009) Historical divergence and gene flow in the genus *Zea. Genetics*
22. Moeller DA & Tiffin P (2008) Geographic variation in adaptation at the molecular level: A case study of plant immunity genes. *Evolution* 62:3069-3081.
23. Pyhäjärvi T*, et al* (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* 177:1713-1724.
24. Pyhäjärvi T, Salmela MJ & Savolainen O (2007) Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. *Tree Genet Gen* 4:247-254.
25. Margulies M*, et al* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
26. Weber AL*, et al* (2008) The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): New evidence from association mapping. *Genetics* 180:1221-1232.
27. Jurinke C, van den Boom D, Cantor CR & Köster H (2002) The use of MassARRAY technology for high throughput genotyping. *Adv Biochem Eng /Biotech* 77:57-74.
28. Gore MA*, et al* (2009) Large-scale discovery of gene-enriched SNPs. *Plant Genome* 2:121-133.
29. Patterson N, Price AL & Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
30. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genet* 5:e1000686.
31. Voight BF*, et al* (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* 102:18508-18513.
32. Beaumont MA, Zhang W & Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025-2035.
33. Ingvarsson PK (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula. Genetics* 180:329-340.
34. Brumfield RT, Beerli P, Nickerson DA & Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* 18:249-256.
35. Lewontin RC & Krakauer J (1973) Distribution of gene frequency as as test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175-195.
36. Beaumont MA & Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 13:969-980.
37. Maynard Smith J & Haigh J (1974) The hitch-hiking effect of a favorable gene. *Genet Res* 23:
38. Fay JC & Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155:1405-

1413.
39. Kim Y & Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513-1524.
40. Nielsen R*, et al* (2005) Genomic scans for selective sweeps using SNP data*. Genome Res* 15:1566-1575.
41. McDonald JH & Kreitman M (1991) Adaptive protein evolution at the *adh* locus in *Drosophila. Nature* 351:652-654.
42. Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome*. Genome Res* 17:1755-1762.
43. Wiehe T & Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster. Mol Biol Evol* 10:842-854.
44. Macpherson JM, Sella G, Davis JC & Petrov DA (2007) Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* 177:2083-2099.
45. van Heerwaarden J, van Eeuwijk FA & Ross-Ibarra J (2009) Genetic diversity in a crop metapopulation*. Heredity*. In press