ABSTRACT

REININGA, JENNIFER MARGARET. Evolutionary Inferences from Genomic Patterns of Polymorphism in *Arabidopsis* and Rice. (Under the direction of Michael D. Purugganan.)

Understanding the genetic basis of adaptation is a major goal of modern evolutionary biology and partitioning the effects of the evolutionary forces that shape levels and patterns of nucleotide variation is an essential component of both disease locus mapping and the identification of functionally important genetic variants in organismal genomes. Describing the effects of natural selection on patterns of genomic variability is an important aspect of each of these efforts. In line with these objectives, we used molecular population genetics to investigate the potential role of selection in shaping nucleotide diversity in high-diversity regions of the *Arabidopsis* genome and across a non-recombining region of a rice centromere.

We conducted a preliminary genome-wide scan for high-diversity genes in the *A. thaliana* genome to identify chromosomal regions with levels and patterns of nucleotide polymorphism that were consistent with the action of balancing selection. Detailed investigations of linked loci across multiple high-diversity regions revealed several candidate genes that might harbor balanced polymorphisms. These genes include *SOLUBLE STARCH SYNTHASE I* (*SSI*), *PLASTID TRANSCRIPTIONALLY ACTIVE 7* (*PTAC7*), and *BELL LIKE HOMEODOMAIN 10* (*BLH10*). To investigate possible functional differences that might be maintained by alternate alleles of these loci we tested for allele specific expression differences and examined the distribution of nonsynonymous polymorphism across each gene. We observe multiple replacement polymorphisms within and in the 32 amino acids just upstream of the protein-protein

interacting BELL domain at the *BLH10* locus and allele-specific expression differences at both *SSI* and *PTAC7*.

To investigate the relationship between recombination and diversity and to examine centromere evolutionary dynamics we sequenced fragments from ~18 expressed single copy genes distributed across a ~3.3 Mb region encompassing rice centromere 8 in ~ 20 individuals each of three domesticated *O. sativa* varieties (*indica*, *temperate japonica*, and *tropical japonica*) as well as from the wild progenitor, *O. rufipogon*. Theory predicts that diversity might be low in this non-recombining region due to hitchhiking with selected variants. We find that both *temperate japonica* and *tropical japonica* centromere 8 regions possess significantly elevated levels of nucleotide variation when compared to an empirical genome-wide distribution while *indica* and *O. rufipogon* show decreased diversity. HKA tests do not support the hypothesis that natural selection is acting at loci in this region. Demography and selective interference might best explain patterns of diversity and divergence at rice centromere 8.

Evolutionary Inferences from Genomic Patterns of Polymorphism in *Arabidopsis* and Rice

by
Jennifer Margaret Reininga

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

GENETICS

Raleigh, North Carolina

2007

APPROVED BY:

Dr. Michael D. Purugganan Chair of Advisory Committee Biology, NYU	Dr. Trudy F. C. Mackay Committee member Genetics
Dr. Jeff Thorne Committee member Genetics and Statistics	Dr. Jose Alonso Committee member Genetics

DEDICATION

In loving memory of my grandparents, Margaret and Earl Richert, who, through their support, taught me the value of an education.

BIOGRAPHY

Jennifer Reininga was born on May 25, 1978 and was raised in Evansville, Indiana by her parents, Gerald and Carol Reininga.. She attended Central High School where she enjoyed taking elective biology courses including human anatomy and advanced biology. In 1996 she decided to attend Western Kentucky University to pursue studies in genetics and graduated from the honors program with a degree in Recombinant Genetics in 2001. During her time as a Hilltopper she discovered her passion for evolutionary biology and population genetics and began her first research project in the field of wildlife conservation genetics. Her continued interest in population genetics and the adaptive process brought her to North Carolina State University where she began her graduate studies in the Department of Genetics.

ACKNOWLEDGEMENTS

I would like to thank the Department of Genetics for excellent training and support over the past six years. A special thanks to my advisor, Michael Purugganan, for taking a chance on me and for being there through both scientific and personal highs and lows. I am especially grateful to my committee members, Drs. Trudy Mackay, Jeff Thorne, and Jose Alonso for their encouragement and advice and to my teacher and friend Bill Atchley who seems to somehow always have my back. I would also like to thank John Willis at Duke University for providing wet lab space and a thriving plant evolution environment for me to enjoy after Michael's move. I have been fortunate to receive funding from the North Carolina State University Department of Genetics and from the National Institute of Health Pre-doctoral training grant.

I am indebted to many former Purugganan lab members who have all moved on to new and exciting things: Kristin Shepard, Marianne Barrier, Mark Ungerer, Ken Olsen, Amy Lawton-Rauh, Jane Suddith, Rich Moore, Kentaro Shimizu, Stephanie Rusza and Ana Caicedo and thanks to present members of the lab as well, especially Megan Hall, Kristie Mather and Ian Ehrenreich. I would like to thank my collaborators Philip Awadalla and Dahlia Nielsen for their help and contributions and my dear friends Alexis Edwards and Amanda Rice for their consistent friendship and for sharing common grad school woes. Where would I be without all of you?

Last, but certainly not least, I'd like to thank my parents, Gerald and Carol Reininga, who have sacrificed a great deal for my successes. They have always been, and continue to be my role models and are a pillar I know I can always rely on. I am

grateful to my whole family for their love and support. Finally, I'd like to express my affection and appreciation for Summer and Hallow, the wagging tails I come home to each night. Their goofy antics find me always astonished and entertained. Thanks for helping me burn the midnight oil!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

I. *Overview*

A major goal of modern evolutionary biology is to identify the specific genetic changes that are selected during adaptation, and providing evidence for the molecular basis of adaptation may help elucidate important genotype/phenotype relationships including those related to human disease or traits of agronomical importance. One powerful approach has been to use DNA sequence variability to identify cases of natural selection that have occurred at specific loci in the genome. However, patterns of genetic variation across organismal genomes are shaped by a variety of evolutionary processes including those with ubiquitous affects across the whole genome (e.g. demography) and those which occur more variably throughout (e.g. mutation, recombination, and natural selection). It is therefore important to disentangle the effects of these different evolutionary processes on genetic diversity so that specific footprints of natural selection can be detected.

Here, I discuss the evolutionary forces that act in conjunction to shape levels and patterns of intraspecific diversity and describe how signatures of selection can be detected in local genomic regions. I consider two empirical datasets from two model plant species to investigate the role of natural selection in shaping patterns of genetic variation: (1) in high-diversity regions of *A. thaliana* and (2) across the non-recombining region encompassing rice centromere 8.

II. The evolutionary forces shaping genetic diveristy

Mutation provides the raw material for heritable differences within a species and adaptation. Heritable genetic changes that have direct fitness consequences will be

visible to natural selection, which should act to remove deleterious mutations that arise in populations of a species (negative or purifying selection) and to preserve adaptive variants (positive selection) (BAMSHAD and WOODING 2003). According to the neutral theory of molecular evolution (KIMURA 1968) most mutations segregating in a population (or a species) should have effects equal to the best fit allele in that population, with their allele frequencies governed by the random process of genetic drift (LI 1997). Although the neutral model accounts for the presence of strongly deleterious mutations, it is assumed that they are immediately removed by negative selection and will therefore not usually be observed (KIMURA 1968). Positive selection, however, is considered rare, and is thought to have little impact on levels of genomic variability. Ever since the development of the neutral theory, the relative roles of drift and selection have been debated and, as a consequence, great efforts have gone towards describing instances where natural selection has shaped genomic diversity in order to assess the extent to which positive selection explains both intra- and inter-species variation (NIELSEN 2005).

Neutral theory predictions are woven throughout the body of theoretical population genetics and tests of neutrality (also called tests of selection) are often conducted to identify regions of the genome with patterns of polymorphism that are inconsistent with a null neutral hypothesis (NIELSEN 2001). Since neutral predictions can be violated by evolutionary forces outside of selection, characterizing patterns of genetic variation within and between populations requires an understanding of the species' natural history or demography. Events such as population bottlenecks, subdivision, and expansion all have predictable effects on the level and distribution of polymorphism within a species. For example, subdivided populations accumulate genetic differences in

isolation from one another and diverge over time at a rate dependent on the level of migration, thereby increasing species-wide diversity (BECKWITT and CHAKRABORTY 1980; TAJIMA 1993). Population (or domestication) bottlenecks, on the other hand, result in a reduction of genetic diversity since variation is lost due to an abrupt decrease in population size (TAJIMA 1989). Finally, population expansion is characterized by an excess of low frequency variants (SLATKIN and HUDSON 1991). These historical events can result in deviations from the standard neutral model in ways that are indistinguishable from natural selection and must be taken into account if selection is to be inferred.

A primary difference between demography and selection is that demographic events affect the entire genome while selection should target only specific sites. Because of this, it has become commonplace to test for selection at a candidate locus using multi-locus comparisons to assess the fit of the data as a whole to the standard neutral equilibrium model. If most loci fit neutral expectations, while a candidate adaptive trait locus shows strong statistical evidence against it, then this is usually considered reasonable evidence that the gene has indeed been selected (NIELSEN 2001). We have now entered an era of unprecedented genomic capability and empirical genome-wide distributions for various population summary statistics (see Table 1) are easily compiled to provide a basis for comparison against which candidate targets of selection (outliers) can be identified (NORDBORG *et al.* 2005; SCHMID *et al.* 2005). Several recent studies have exploited genome-wide data to generate lists of candidate adaptive trait loci (AKEY *et al.* 2002; HARR *et al.* 2002; NIELSEN *et al.* 2005; SWANSON *et al.* 2001; WRIGHT *et al.* 2005) and maps of selection in the genome (VOIGHT *et al.* 2006; WILLIAMSON *et al.* 2007).

In addition to mutation, selection and demography, the distribution of polymorphism within organismal genomes is also shaped by recombination, which promotes the mixing of genetic variants during sexual reproduction. Through this process, beneficial mutations that arise in different individuals can be placed together in novel combinations on the same chromosome (OTTO and BARTON 1997). Therefore, one consequence of recombination is that linked genetic variants share common evolutionary histories that are independent from those of other genomic regions (AWADALLA 2003). Because of their common histories, non-random associations of polymorphic sites, or linkage disequilibrium (LD), is often observed among variants in linked genomic regions. This LD will persist until recombination sufficiently acts to break associations among linked sites. Since the probability of a recombination event increases with distance, the extent of linkage disequilibrium can be used as a proxy for evaluating levels of recombination. Recent studies have revealed that rates of recombination vary greatly among different regions of the genome (CRAWFORD *et al.* 2004; MCVEAN *et al.* 2004) and that this variance is greater in plants than animals (GAUT *et al.* 2007). One consequence of this heterogeneous recombination landscape is that the extent of LD will also vary accordingly, due to chance events.

Empirical observations of positive correlations between rates of recombination and diversity, but not divergence, in *Drosophila melanogaster* (AGUADE *et al.* 1994; AGUADE *et al.* 1989; BEGUN and AQUADRO 1992; STEPHAN and LANGLEY 1989) gained attention by demonstrating that the responsible mechanisms must act at the population level (LANGLEY *et al.* 2000). These findings suggested that selection might be quite prevalent throughout the genome (BEGUN and AQUADRO 1992). As a result,

investigations of reduced variation in regions of low crossing over, such as centromeres and telomeres, have been subject to much theoretical scrutiny and models incorporating the action of both positive (directional hitchhiking) (KAPLAN *et al.* 1989) and negative selection (background selection) (CHARLESWORTH *et al.* 1993) have been proposed.

Depending on the recombination rate, strength of selection, and local mutation rate, selection acting on a specific functional variant can have profound affects on levels and patterns of neutral polymorphism segregating in the linked chromosomal region through the process of hitchhiking. During directional hitchhiking (also called a selective sweep), a beneficial mutation experiencing positive selection will rapidly increase in frequency in a population or a species, dragging along with it all of the neutral variation in the linked chromosomal region (KAPLAN *et al.* 1989; MAYNARD SMITH and HAIGH 1974). Directional hitchhiking should therefore initially increase linkage disequilibrium while decreasing nucleotide diversity. Often chromosomal regions that have recently experienced directional selection are further characterized by an excess of low frequency mutations, which gradually re-accumulate in the region (BAMSHAD and WOODING 2003). In regions of the genome with low rates of crossing over per physical distance, such as centromeres and telomeres, hitchhiking should have more pronounced effects, with this signal of selection extending across a broader portion of the chromosome (KAPLAN *et al.* 1989).

Background selection was proposed as an alternative to directional hitchhiking and is based on the theoretical prediction that selection against deleterious alleles introduced into a population by mutation should also cause a reduction in the level of genetic diversity at linked neutral sites (CHARLESWORTH *et al.* 1993) .Since neutral

mutations will be rapidly removed from the population if they are associated with deleterious alleles, only gametes that remain free from deleterious mutation will persist for long periods of time. As with directional selection, background selection is expected to produce its most extreme effects in expansive regions of low recombination. One important distinction, however, is that background selection can produce milder reductions in diversity than directional selection and tests of neutrality may not always indicate significant deviations from the null hypothesis (CHARLESWORTH 1996; LANGLEY *et al.* 2000).

Diversity can also be selectively *maintained* within a population or a species through the process of balancing selection through mechanisms such as frequency dependant selection (AWADALLA and CHARLESWORTH 1999), heterozygote advantange (overdominance) (AIDOO *et al.* 2002; TISHKOFF *et al.* 2001), and spatial-temporal selection (TIAN *et al.* 2003). Population genetic theory predicts a characteristic molecular footprint of elevated diversity for regions of the genome experiencing strong balancing selection, including increased peaks of nucleotide diversity that decrease somewhat symmetrically with distance to either side of the selected site(s) (HUDSON and KAPLAN 1988). As with other forms of selection, the strength and breadth of this signal depend strongly on local rate of recombination, time since the selective event, strength of selection, and the local mutation rate. Balanced polymorphisms commonly segregate at intermediate frequency in a the population or species and can result in trans-specific polymorphism if alleles are maintained over evolutionary times that predate the ages of species, a finding that is highly unlikely under neutrality (SCHIERUP *et al.* 2001). Thus,

one common effect of balancing selection is to deflate divergence while increasing diversity (BAMSHAD and WOODING 2003).

Statistical tests of selective neutrality have been developed and employed to identify genomic regions of adaptive and functional importance. Each test is sensitive to different patterns of variability in the data. For example, tests can be based on the mutational frequency spectrum, comparisons across multiple loci or species, or levels of variability among different categories of mutation. Table 1 includes a description of several common measures of genetic diversity and tests of neutrality.

III. Adaptation in model plant species

Plant species possess diverse morphological and functional phenotypes, many of which are thought to be adaptive. If the genes and molecular mechanisms responsible for naturally occurring variation among individuals of a species are to be understood, the availability of genomic resources, such as whole genome sequence data, is essential. These resources allow for the development of tools to aid in determining gene functions and the genetic mechanisms that underlie adaptive phenotypes and have helped to make model plant species like *Arabidopsis thaliana* and rice powerful systems for the study of evolutionary and population genetics.

a. Arabidopsis thaliana and balanced polymorphisms

Arabidopsis thaliana is a member of the mustard (Brassicaceae) family and has become the major plant model system, partly owing to its small genome (125Mb over 5 chromosomes), fast generation time (6 weeks), and inbreeding nature (~99% inbreeding)

(ABBOT AND GOMEZ 1989). Since the sequencing of the Col-0 accession genome, a variety of molecular and bioinformatic tools have been developed and applied to help elucidate gene function including knockout mutant lines (ALONSO *et al.* 2003) , expression databases/resources (MEYERS *et al.* 2004), and efforts to generate a species wide hapotype map (NORDBORG *et al.* 2005). These provide valuable resources for the study of evolutionary and ecologically relevant traits and have helped to reveal candidate adaptive trait loci on the basis of function and homology.

Demographic effects including inbreeding and a weedy life history are likely to affect levels and patterns of diversity in *A. thaliana*. Selfing organisms have low effective rates of recombination and more extensive LD (up to ~50 kb) relative to outcrossing organisms (NORDBORG *et al.* 2005). Moreover, the weedy nature of this species means extinction, recolonization and population expansion may be quite common (GAUT *et al.* 2007). Indeed, the mutational frequency spectrum in *A. thaliana* is skewed towards rare variants, a finding that has been interpreted as evidence of a recent population expansion (BERGELSON *et al.* 1998). One key point is that the neutral equilibrium model assumes constant population size and random mating, both of which are violated by *A. thaliana*. Furthermore, recent data have revealed significant population structure and isolation by distance (NORDBORG *et al.* 2005; SCHMID *et al.* 2005; SHARBEL *et al.* 2000), and geographic patterns of genetic differentiation among naturally occurring accessions have led to the suggestion that *A. thaliana* experienced a population expansion following a period of isolation in Pleistocene glacial refugia on the Iberian peninsula and in the Balkans (SHARBEL *et al.* 2000).

Many genes are reported to have experienced positive selection in *A. thaliana (*reviewed in Wright and Gaut 2004*).* However, until recently, population structure and demography have been largely ignored in *Arabidopsis* population genetics, primarily due to a lack of a complete understanding of how these processes affect patterns of diversity in this model plant system. Because both population expansion and directional selection are expected to produce common molecular signatures characterized by an excess of rare variants, claims for directional selection in this species should be interpreted conservatively. *A. thaliana*, however, is thought to be much better suited for identifying balanced polymorphisms (TIAN *et al.* 2002). The species' low effective recombination rate and extensive LD means that this signature should be maintained over larger stretches of the chromosome, easing the detection of selected genomic regions. Thus, one caveat is that the identification of the responsible functional polymorphisms might be a difficult challenge.

Strong cases for balancing selection have been made for several genes in *A. thaliana*. The best examples come from genes with roles in disease resistance (STAHL *et al.* 1999; TIAN *et al.* 2003) and protection against herbivory (KROYMANN *et al.* 2003); in both cases presence/absence polymorphisms are thought to be maintained due to fitness trade-offs. For example, a strong molecular signature of balancing selection has been identified in the ~3kb genomic region flanking the *RPS5* disease resistance locus and 250 bp sliding window estimates of levels of nucleotide diversity reach ~ 0.12, 17 fold above the genome average, at the insertion/deletion (indel) junction (STAHL *et al.* 1999). In the absence of the *Pseudomonas* pathogen individuals possessing the gene have been shown to have decreased fitness relative to those lacking a functional copy. These data have

been interpreted to reflect the maintenance of alleles due to fluctuating selection pressure across space and time (spatial-temporal selection), depending on the occurrence of the pathogen (TIAN *et al.* 2002; TIAN *et al.* 2003). Similarly, the chromosomal region flanking the *MAM2* herbivory locus also displays increased levels of diversity, consistent with the maintenance of a balanced polymorphism. In this case, however, fitness tradeoffs are thought to result from differences in allelic performance against specialist vs. generalist herbivores (KROYMANN *et al.* 2003).

Genome scanning approaches based on identifying patterns of molecular evolution that deviate from the neutral equilibrium model or from empirical distributions have implicated the action of natural selection at many loci across a variety of species (HARR *et al.* 2002; NIELSEN *et al.* 2005; SWANSON *et al.* 2001; VOIGHT *et al.* 2006; WILLIAMSON *et al.* 2007; WRIGHT *et al.* 2005). Interestingly, most genome scanning efforts have focused on identifying loci subject to directional selection, while balancing selection has received much less attention (but see Bubb *et al.* 2006). *Arabidopsis thaliana* is especially well suited for the identification of novel targets of balancing selection via genome scanning due to its inbreeding nature and low effective rate of recombination. The arrival of single pass shotgun sequence data from the *A. thaliana* accession Landsberg *erecta* (L*er*-0) (JANDER *et al.* 2002) has made possible large scale intraspecific comparisons of diversity by which genes with elevated levels of polymorphism consistent with the maintenance of balanced polymorphisms can be readily detected.

The identification of novel targets of balancing selection is of great importance since the relative contribution of this form of selection to the maintenance genomic

diversity is not well understood. Herein I present the results of a genome-wide scan for candidate balanced polymorphisms in *A. thaliana*. Several interesting chromosomal regions are investigated using molecular population genetics to assess the hypothesis that selection has acted at loci in these regions (see chapter 1). Novel candidate balanced polymorphisms are identified and naturally occurring variation among alleles at candidate loci is associated with potential functional differences, thereby providing a putative functional basis for the selective maintenance of alleles (see chapter 2).

b. Centromere evolution in rice

The Asian cultivated rice, *Oryza sativa* L., is grown world-wide under many different environmental conditions, highlighting the importance of this crop as the primary staple food for more than half of the human population (LONDO et al. 2006). *O. sativa* is comprised of multiple cultivars which possess a large amount of phenotypic diversity, reflecting the preferences of humans during artificial selection (JACKSON 1997; TAKAHASHI *et al.* 1997). Two main sub-species, *O. sativa indica* and *O. sativa japonica*, exist, although the japonica classification is further categorized into both tropical and temperate varieties. *Indica* rice is distributed across lowland tropical Asia while *tropical* and *temperate japonica* varieties grow in upland tropical and in temperate regions respectively. The agronomical importance of this plant species and its small genome size (relative to other domestic grains) has helped establish a role for rice a genetic model for the study of plant domestication. To this end, the 420 Mb *Oryza sativa* L. *ssp. japonica* (Nipponbare ecotype) genome was shotgun sequenced resulting in 6-fold coverage with

98% confidence on all base calls (GOFF *et al.* 2002). Whole genome sequencing of the *indica* cultivar has been undertaken too, albeit with lesser coverage (YU *et al.* 2002).

Both inbreeding and domestication bottlenecks should affect levels and patterns of diversity across loci in the *O. sativa* genome. Two independent domestication events from the wild progenitor, *O. rufipogon,* are thought to have occurred ~12,000 years ago during the Neolithic revolution (NORMILE 1997) to give rise to separate *indica* and *tropical japonica* varieties (VITTE *et al.* 2004), while t*emperate japonicas* are thought to be the result of a second domestication from the *tropical japonica* group. The demographic effects of both selfing and domestication should reduce levels of genetic diversity in *O. sativa* and empirical observations provide support for these theoretical expectations with average genomic diversity for *indica* and *tropical japonica* being approximately half that of the partially outcrossing ancestor *O. rufipogon* and with *temperate japonica* possessing approximately half the diversity of *tropical japonica* (CAICEDO *et al.* submitted).

Recently, rice has emerged as a model for centromere structure, function and evolution (YAN and JIANG 2007). This is largely a consequence of the completion of full length *O. sativa* (Nipponbare) centromere sequences for chromosomes 4, 5 and 8 (MATSUMOTO *et al.* 2005; NAGAKI *et al.* 2004; WU *et al.* 2004; ZHANG *et al.* 2004), a task accomplished due to the diminished occurrence of repetitive elements in these centromeres relative to other centromeres in rice and other taxa (MA *et al.* 2007). Although initially perceived as entirely heterochromatic, recent work has shown that the functional centromere, or kinetochore, may also possess euchromatic or other distinctive methylation features (CAM *et al.* 2005; SULLIVAN and KARPEN 2004; WONG *et al.* 2006;

YAN and JIANG 2007). Consistent with these observations, expressed genes have been identified in the kinetochore of rice centromere 8, a region thought to be transcriptionally inert until this demonstration (NAGAKI *et al.* 2004). This finding has led to suggestions that centromere 8 represents an early stage of centromere formation (NAGAKI *et al.* 2004; YAN and JIANG 2007).

Centromeres are known to have reduced levels of meiotic recombination, a characteristic that has typically been attributed to the repetitive nature of these regions and their associated heterochromatin, but may also be epigenetically regulated (YAN *et al.* 2005). In this context, centromeres represent the extreme and hence simplest example of regions in which the relationship between recombination, selection and diversity can be explored. Moreover, the existence of actively expressed genes in the centromere 8 region permits investigations into the evolutionary mechanisms shaping variation at functional genes as neocentromeres morph into fully mature centromeres.

In chapter three I present a molecular population genetic study of diversity at loci across rice centromere 8, and the first population data from genes located within a functioning kinetochore. Patterns of polymorphism and divergence are compared among three major domesticated rice varieties (*indica*, *temperate* and *tropical japonica*) and the wild progenitor *O. rufipogon* using *O. meridionalis* as an outgroup. We show contrasting patterns of variation between rice groups across the region and discuss possible explanations for empirical deviations from theoretical expectations of models based on hitchhiking and background selection.

Table 1. Population genetic parameters and tests of neutrality.

Parameter or Statistic	Basis	Description
Population mutation parameter		
π (Tajima 1983)	Estimator of $4N_e\mu$[a]	Estimates the probability of polymorphism per nucleotide site from the mean number of nucleotide differences between two sequences that are randomly chosen from a common sample.
θ_w (Watterson 1975)	Estimator of $4N_e\mu$[a]	Estimates the probability of polymorphism per nucleotide site from the proportion of segregating sites in a sample.
Linkage disequilibrium		
D	Measure of association	A measure of LD based on the difference between the observed frequency of a two-locus haplotype and the expected haplotype frequency assuming the random segregation of alleles.
$\|D'\|$ (Lewontin 1964)	Measure of association	A measure of LD obtained by standardizing *D* by its maximum possible value, given the allele frequencies at the two loci. Values range between 0 and 1 reaching 1 (complete LD) only when there has been no historical recombination between considered sites. Intermediate values of |D'| are highly affected by sample size.
r^2 (Hill and Robertson 1966)	Measure of association	The statistical correlation between two variable sites obtained by dividing D^2 by the product of the four allele frequencies at the two loci. Values range between 0 and 1, reaching a value of 1 (perfect LD) if no recombination has occurred between the considered sites and if they have a common allele frequency. Intermediate values of r^2 are easily interpreted.
Recombination		
ρ (Hudson and Kaplan 1985)	Estimate of $4N_er$[b]	A measure of how much recombination would be required to generate observed levels of LD.
Tests of neutrality		
Tajima's *D* (Tajima 1993)	Frequency distribution	A test based on the difference between θ_w and π that is sensitive to a skew in allele frequency.
Z_{nS} (Kelly 1997)	Linkage disequilibrium	The average standardized linkage disequilibrium across all pairwise comparisons among informative sites at a common locus. Significant elevations in the statistic are indicative of positive selection.
HKA (Hudson *et al.* 1987)	Multilocus comparison	A test based on comparisons among levels of polymorphism and divergence at two or more loci.
McDonald-Kreitman (McDonald and Kreitman 1991)	Site category comparison	A test based on comparisons among the ratios of synonymous and non-synonymous nucleotide substitution among loci within and between species.
d_N/d_S (or K_a/K_s) (Li 1981)	Site category comparison	A conservative test based on the variability and divergence of the rate of mutation among non-synonymous sites (d_N) and synonymous sites (d_S) in protein coding regions of the genome.

[a] Where N_e is the effective population size and μ is the neutral mutation rate per generation.
[b] Where N_e is the effective population size and *r* is the rate of recombination.

REFERENCES

ABBOT, R. J., and M. F. GOMES, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

AGUADE, M., W. MEYERS, A. D. LONG and C. H. LANGLEY, 1994 Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the su(s) and su(wa) regions of the Drosophila melanogaster X chromosome. Proc Natl Acad Sci U S A **91:** 4658-4662.

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced Variation in the Yellow-Achaete-Scute Region in Natural Populations of Drosophila Melanogaster. Genetics **122:** 607-615.

AIDOO, M., D. J. TERLOUW, M. S. KOLCZAK, P. D. MCELROY, F. O. TER KUILE *et al.*, 2002 Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet **359:** 1311-1312.

AKEY, J. M., G. ZHANG, K. ZHANG, L. JIN and M. D. SHRIVER, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res **12:** 1805-1814.

ALONSO, J. M., A. N. STEPANOVA, T. J. LEISSE, C. J. KIM, H. CHEN *et al.*, 2003 Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science **301:** 653-657.

AWADALLA, P., 2003 The evolutionary genomics of pathogen recombination. Nat Rev Genet **4:** 50-60.

AWADALLA, P., and D. CHARLESWORTH, 1999 Recombination and selection at Brassica self-incompatibility loci. Genetics **152:** 413-425.

BAMSHAD, M., and S. P. WOODING, 2003 Signatures of natural selection in the human genome. Nat Rev Genet **4:** 99-111.

BECKWITT, R., and R. CHAKRABORTY, 1980 Genetic structure of Pileolaria pseudomilitaris (Polychaeta: Spirorbidae). Genetics **96:** 711-726.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519-520.

BERGELSON, J., E. STAHL, S. DUDEK and M. KREITMAN, 1998 Genetic variation within and among populations of Arabidopsis thaliana. Genetics **148:** 1311-1323.

BUBB, K. L., D. BOVEE, D. BUCKLEY, E. HAUGEN, M. KIBUKAWA, M. PADDOCK, A. PALMIERI, S. SUBRAMANIAN, Y. ZHOU, R. KAUL, P. GREEN, and M. V. OLSON. 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics **173**: 2165-2177.

CAICEDO, A.L., WILLIAMSON, S.H., HERNANDEZ, R.D., FLEDEL-ALON, A., YORK *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. Submitted to PLOS Genetics.

CAM, H. P., T. SUGIYAMA, E. S. CHEN, X. CHEN, P. C. FITZGERALD *et al.*, 2005 Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. Nat Genet **37:** 809-819.

CHARLESWORTH, B., 1996 Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res **68:** 131-149.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CRAWFORD, D. C., T. BHANGALE, N. LI, G. HELLENTHAL, M. J. RIEDER *et al.*, 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet **36:** 700-706.

GAUT, B. S., S. I. WRIGHT, C. RIZZON, J. DVORAK and L. K. ANDERSON, 2007 Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet **8:** 77-84.

GOFF, S. A., D. RICKE, T. H. LAN, G. PRESTING, R. WANG *et al.*, 2002 A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296:** 92-100.

HARR, B., M. KAUER and C. SCHLOTTERER, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. Proc Natl Acad Sci U S A **99:** 12949-12954.

HILL, W.G., and A. ROBERTSON, 1966 The effect of linkage on the limits to natural selection. Genet Res **8:** 269-294.

HUDSON, R.R., and N.L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147-164.

HUDSON, R.R., M. KREITMAN, and M. AGUADE. 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153-159.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831-840.

JACKSON, M. T., 1997 Conservation of rice genetic resources: the role of the International Rice Genebank at IRRI. Plant Mol Biol **35:** 61-67.

JANDER, G., S. R. NORRIS, S. D. ROUNSLEY, D. F. BUSH, I. M. LEVIN *et al.*, 2002 Arabidopsis map-based cloning in the post-genome era. Plant Physiol **129:** 440-450.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

KELLY, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146**: 1197-1206.

KIMURA, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res **11:** 247-269.

KROYMANN, J., S. DONNERHACKE, D. SCHNABELRAUCH and T. MITCHELL-OLDS, 2003 Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci U S A **100 Suppl 2:** 14587-14592.

LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics **156:** 1837-1852.

LEWONTIN, R.C., 1964 The interaction of selection and linkage. I. General Considerations; Heterotic Models. Genetics **49:** 49-67.

LI, W.-H., T. GOJOBORI, and M. NEI, 1981 Pseudogenes as a paradigm of neutral evolution. Nature **292**:237-239.

MA, J., R. A. WING, J. L. BENNETZEN and S. A. JACKSON, 2007 Plant centromere organization: a dynamic structure with conserved functions. Trends Genet **23:** 134-139.

MATSUMOTO, T., J. Z. WU, H. KANAMORI and E. AL., 2005 The map-based sequence of the rice genome. Nature **436:** 793-800.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genetical Research **23:** 23-35.

MCDONALD, J. and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351**:652-654.

MCVEAN, G. A., S. R. MYERS, S. HUNT, P. DELOUKAS, D. R. BENTLEY *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581-584.

MEYERS, B. C., D. K. LEE, T. H. VU, S. S. TEJ, S. B. EDBERG *et al.*, 2004 Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol **135:** 801-813.

NAGAKI, K., Z. CHENG, S. OUYANG, P. B. TALBERT, M. KIM *et al.*, 2004 Sequencing of a rice centromere uncovers active genes. Nat Genet **36:** 138-145.

NIELSEN, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641-647.

NIELSEN, R., 2005 Molecular signatures of natural selection. Annu Rev Genet **39:** 197-218.

NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res **15:** 1566-1575.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3:** e196.

NORMILE, D., 1997 Archaeology - Yangtze seen as earliest rice site. Science **275:** 309-309.

OTTO, S. P., and N. H. BARTON, 1997 The evolution of recombination: removing the limits to natural selection. Genetics **147:** 879-906.

SCHIERUP, M. H., A. M. MIKKELSEN and J. HEIN, 2001 Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics **159:** 1833-1844.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601-1615.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol **9:** 2109-2118.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-562.

STAHL, E. A., G. DWYER, R. MAURICIO, M. KREITMAN and J. BERGELSON, 1999 Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature **400:** 667-671.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. Genetics **121:** 89-99.

SULLIVAN, B. A., and G. H. KARPEN, 2004 Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat Struct Mol Biol **11:** 1076-1083.

SWANSON, W. J., A. G. CLARK, H. M. WALDRIP-DAIL, M. F. WOLFNER and C. F. AQUADRO, 2001 Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci U S A **98:** 7375-7379.

TAJIMA, F., 1989 The effect of change in population size on DNA polymorphism. Genetics **123:** 597-601.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437-460.

TAJIMA, F., 1993 Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol **10**: 677-688.

TAJIMA, F., 1993 Statistical analysis of DNA polymorphism. Jpn J Genet **68:** 567-595.

TAKAHASHI N, K. HAMAMURA, S. TSUNODA, S. SAKAMOTO, Y. SATO, 1997 Differentiation of ecotypes in cultivated rice. In: Matsuo T, Futsuhara Y, Kikuchi F, Yamaguchi H, editors. Science of the Rice Plant. Tokyo: Food and agriculture policy research center. pp. 112-160.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A **99:** 11525-11530.

TIAN, D., M. B. TRAW, J. Q. CHEN, M. KREITMAN and J. BERGELSON, 2003 Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature **423:** 74-77.

TISHKOFF, S. A., R. VARKONYI, N. CAHINHINAN, S. ABBES, G. ARGYROPOULOS *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293:** 455-462.

VITTE, C., T. ISHII, F. LAMY, D. BRAR and O. PANAUD, 2004 Genomic paleontology provides evidence for two distinct origins of Asian rice (Oryza sativa L.). Mol Genet Genomics **272:** 504-511.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. PLoS Biol **4:** e72.

WATTERSON, G. A. 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7:** 256-276.

WILLIAMSON, S. H., M. J. HUBISZ, A. G. CLARK, B. A. PAYSEUR, C. D. BUSTAMANTE *et al.*, 2007 Localizing Recent Adaptive Evolution in the Human Genome. PLoS Genet **3:** e90.

WONG, N. C., L. H. WONG, J. M. QUACH, P. CANHAM, J. M. CRAIG *et al.*, 2006 Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. PLoS Genet **2:** e17.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310-1314.

WU, J., H. YAMAGATA, M. HAYASHI-TSUGANE, S. HIJISHITA, M. FUJISAWA *et al.*, 2004 Composition and structure of the centromeric region of rice chromosome 8. Plant Cell **16:** 967-976.

YAN, H., and J. JIANG, 2007 Rice as a model for centromere and heterochromatin research. Chromosome Res **15:** 77-84.

YAN, H., W. JIN, K. NAGAKI, S. TIAN, S. OUYANG *et al.*, 2005 Transcription and histone modifications in the recombination-free region spanning a rice centromere. Plant Cell **17:** 3227-3238.

YU, J., S. HU, J. WANG, G. K. WONG, S. LI *et al.*, 2002 A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296:** 79-92.

ZHANG, Y., Y. HUANG, L. ZHANG, Y. LI, T. LU *et al.*, 2004 Structural features of the rice chromosome 4 centromere. Nucleic Acids Res **32:** 2023-2030.

CHAPTER TWO

High-Diversity Genes in the *Arabidopsis* genome

High-Diversity Genes in the *Arabidposis* Genome

Jennifer M. Cork[1]* and Michael D. Purugganan[1]

[1]Department of Genetics, Box 7614
North Carolina State University, Raleigh, NC 27695

*Name change; this author is now Jennifer M. Reininga

REFERENCE:

Cork, J. M., and M. D. Purugganan. 2005. High-diversity genes in the Arabidopsis genome. *Genetics* **170**:1897-1911.

DOI: 10.1534/genetics.104.036152

High-Diversity Genes in the Arabidopsis Genome

Jennifer M. Cork and Michael D. Purugganan[1]

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695

Manuscript received January 3, 2005
Accepted for publication April 5, 2005

ABSTRACT

High-diversity genes represent an important class of loci in organismal genomes. Since elevated levels of nucleotide variation are a key component of the molecular signature for balancing selection or local adaptation, high-diversity genes may represent loci whose alleles are selectively maintained as balanced polymorphisms. Comparison of 4300 random shotgun sequence fragments of the *Arabidopsis thaliana* L*er* ecotype genome with the whole genomic sequence of the Col-0 ecotype identified 60 genes with putatively high levels of intraspecific variability. Eleven of these genes were sequenced in multiple *A. thaliana* accessions, 3 of which were found to display elevated levels of nucleotide polymorphism. These genes encode the *myb*-like transcription factor *MYB103*, a putative soluble starch synthase I, and a homeodomain-leucine zipper transcription factor. Analysis of these genes and 4–7 flanking genes in 14–20 *A. thaliana* ecotypes revealed that two of these loci show other characteristics of balanced polymorphisms, including broad peaks of nucleotide diversity spanning multiple linked genes and an excess of intermediate-frequency polymorphisms. Scanning genomes for high-diversity genomic regions may be useful in approaches to adaptive trait locus mapping for uncovering candidate balanced polymorphisms.

UNCOVERING the genetic basis of adaptation has been a central goal of evolutionary genetics for nearly a century (Orr and Coyne 1992), and recent advances in genetic analysis have permitted the identification and isolation of loci responsible for speciation (Greenberg *et al.* 2003; Barbash *et al.* 2004), species differences (Doebley *et al.* 1997; Gompel and Carroll 2003), and adaptive intraspecific variation (Johanson *et al.* 2000; Kroymann *et al.* 2003). Several approaches based on patterns of molecular evolution have been proposed to scan genomes for genes associated with adaptation (Nielsen 2001; Swanson *et al.* 2001a,b; Schlotterer 2002; Bamshad and Wooding 2003; Barrier *et al.* 2003). These methods provide opportunities to analyze evolutionary diversification at both molecular genetic and phenotypic levels.

Approaches for mapping adaptive trait loci are based on detecting regions of the genome in which intraspecific sequence variation and/or interspecific divergence deviate either from predictions of a neutral-equilibrium model (Nielsen 2001) or from the norm of a genome-wide distribution (Otto 2000; Luikart *et al.* 2003). Evolutionary expressed sequence tag (EST) (Swanson *et al.* 2001a,b; Barrier *et al.* 2003) and comparative genomic approaches (Clark *et al.* 2003), for example, use interspecific patterns of nonsynonymous/synonymous substitution ratios (K_a/K_s) to identify candidate adaptive genes on the basis of accelerated rates of protein evolution ($K_a/K_s > 1$). Genes and genomic regions associated with directional selection have also been identified by scanning dense sets of genome-wide molecular markers for reduced levels of variation (Harr *et al.* 2002; Payseur *et al.* 2002; Schlotterer 2002; Vigouroux *et al.* 2002; Wootton *et al.* 2002; Storz *et al.* 2004). The latter approach, referred to as hitchhiking mapping, is based on the premise that a beneficial mutation that rapidly spreads in a population will also reduce nucleotide variation at linked neutral loci. Hitchhiking mapping has successfully identified several genomic regions containing putative adaptive trait loci that were thought to contribute to the worldwide colonization of *Drosophila melanogaster* out of Africa ~10,000 years ago (Harr *et al.* 2002). Although genome scanning for putative adaptive trait loci on the basis of levels of molecular diversity has focused largely on identifying genes associated with directional selection, this approach could also be employed in identifying genes and/or genomic regions that harbor balanced polymorphisms. This could complement other genome-scanning approaches, such as screens for elevated F_{ST} estimates in marker loci between two populations (Akey *et al.* 2002), in identifying genes under this selective regime.

Balanced polymorphisms are characterized by two or more alleles that are selectively maintained at intermediate frequencies within populations or species. Frequency-dependent selection (Charlesworth and Awadalla 1998; Bergelson *et al.* 2001), heterozygote advantage

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ132063–DQ132370.

[1]*Corresponding author:* Department of Genetics, Box 7614, 3513 Gardner Hall, North Carolina State University, Raleigh, NC 27695. E-mail: michael_purugganan@ncsu.edu

(Bamshad and Wooding 2003), and local adaptation (Kohn *et al.* 2000; Schulte *et al.* 2000; Gilad *et al.* 2002; Harr *et al.* 2002; Kohn *et al.* 2003; Storz *et al.* 2004) are some of the major selective mechanisms that can maintain balanced polymorphisms. These polymorphisms are also associated with specific levels and patterns of nucleotide variation at the selective target and in linked genomic regions (Strobeck 1983; Hudson and Kaplan 1988) and thus provide a molecular signature of adaptation that can aid in their identification. This signature can include increased levels of within-species diversity as well as intermediate-frequency polymorphisms (Bamshad and Wooding 2003), which, when maintained for long periods of time, are thought to result in *trans*-specific polymorphism (Schierup *et al.* 1998). Models of balancing or spatially heterogeneous selection also predict peaks of increased nucleotide diversity centered on a balanced polymorphism, which decrease symmetrically with distance (Hudson and Kaplan 1988; Nordborg 1997). Additional, less definitive features of a balanced polymorphism include high levels of linkage disequilibrium and a deficiency in the number of observed haplotypes (Charlesworth 2003), which result from selective hitchhiking. These characteristics of the signature of a balanced polymorphism have been observed in studies of several genes and/or gene regions in diverse species, including the human class I and II MHC (Garrigan and Hedrick 2003), Drosophila *Adh* (Kreitman and Aguade 1986; Kreitman and Hudson 1991), Fundulus *Ldh* (Schulte *et al.* 2000), *Arabidopsis thaliana MAM* (Kroymann *et al.* 2003), and disease resistance loci in plants (Stahl *et al.* 1999; Bergelson *et al.* 2001; Tian *et al.* 2002; Charlesworth *et al.* 2003).

Although balanced polymorphisms have been observed in various organisms, highly self-fertilizing species, like the model organism *A. thaliana*, are thought to be especially well suited for the identification and analysis of genes subject to balancing selection (Nordborg 1997; Tian *et al.* 2002; Shepard and Purugganan 2003). *A. thaliana* outcrosses at a rate of ~1%, resulting in a low effective rate of recombination (Abbot and Gomes 1989), and linkage disequilibrium that can extend over genomic regions spanning ~50–250 kb (Nordborg *et al.* 2002). This reduced effective recombination rate in *A. thaliana* should maintain correlations between nucleotide polymorphisms over larger distances and longer persistence times and facilitate the discovery of balanced polymorphisms.

The pattern of nucleotide variation associated with a balanced polymorphism in *A. thaliana* is illustrated by the *RPS5* disease resistance locus (Tian *et al.* 2002). A genomic area centered on *RPS5* ~5.8 kb in length shows increased sequence variability, with silent-site levels of nucleotide diversity (π) at 0.025. In this instance, the balanced polymorphism appears to result in persistent haplotype dimorphism across this genomic region, although the dimorphic allele classes of closely linked genes are no longer associated with particular disease resistance alleles. Several of the features of the balanced polymorphism at *RPS5* are also shared with the genomic region of *CLAVATA2*, a gene encoding a leucine-rich repeat involved in *A. thaliana* shoot meristem development (Shepard and Purugganan 2003), which also may be subject to balancing selection.

The predominantly selfing nature of *A. thaliana*, the availability of large-scale genome sequences from two Arabidopsis ecotypes, Columbia (Col-0) (Arabidopsis Genome Initiative 2000) and Landsberg *erecta* (L*er*) (Jander *et al.* 2002), and current efforts to develop a species-wide haplotype map (http://walnut.usc.edu/2010/) provide a unique opportunity to scan the whole genome of this model plant for high-diversity genes that may arise from balanced polymorphisms. It is unclear, however, whether such an approach can identify these adaptive polymorphisms, since other evolutionary forces such as mutation, gene duplication, and population structuring could also result in high-diversity genes. Disentangling these alternative possibilities and determining the utility of a genome-scanning approach for identifying balanced polymorphisms requires a better understanding of the levels and patterns of nucleotide variation for high-diversity genes and their associated genomic regions.

In this article we report on a molecular population genetic analysis of high-diversity genes and genomic regions in the model genetic organism *A. thaliana*. From a comparison of genome sequence data between the Col-0 and L*er* *A. thaliana* ecotypes, we have identified three gene fragments that show divergence between these two ecotypes of >5%. The levels and patterns of nucleotide polymorphism in the genomic regions spanning these high-diversity gene fragments were also ascertained, thereby providing the foundation for determining the utility of large-scale intraspecific genome scans in identifying candidate genes that may harbor balanced polymorphisms.

MATERIALS AND METHODS

Identification of genes for analysis: Approximately 4300 genome-wide shotgun sequence fragments from the *A. thaliana* L*er* ecotype (Jander *et al.* 2002) were compared to the Col-0 ecotype full-genome sequence (Arabidopsis Genome Initiative 2000) through a large-scale BLAST analysis (Altschul *et al.* 1990). Sequences that were 2–10% divergent between these two ecotypes were identified. Transposable elements, genes producing proteins <150 amino acids in length, pseudogenes, and duplicate gene copies were excluded from this sequence list. Eleven of these high-diversity gene fragments were sequenced in five to six additional *A. thaliana* ecotypes to confirm the high levels of within-species polymorphism at these loci. Genes were included in further analysis if elevated levels of nucleotide diversity were validated and if the gene fragments displayed significantly positive Tajima's *D* or Fu and Li's *D** value.

Isolation and sequencing of alleles: Genomic DNA was isolated from young leaves of 21 *A. thaliana* accessions (supplementary Table S1 at http://www.genetics.org/supplemental/) and from

one to three *A. lyrata* plants using the plant DNeasy mini kit (QIAGEN, Valencia, CA). The *A. thaliana* accessions primarily span the geographic range of this species in Europe, although some Asian and North African accessions are included (see supplementary Table S1 at http://www.genetics.org/supplemental/). *A. lyrata* seed from a Karhumaki, Russia, population was provided by O. Savolainen (University of Oulu) and Helmi Kuittinen (University of Barcelona). PCR primers were designed from Col-0 genomic BAC sequences using Primer3 (Rozen and Skaletsky 2000). Primers were designed to amplify ~1-kb regions of the three confirmed high-diversity genes identified by our BLAST analysis (AT1G63910, AT5G24300, AT1G19700) (Table 1 and supplementary Tables S2 and S3 at http://www.genetics.org/supplemental/). Primers were also designed to amplify 0.5- to 1-kb regions of genes flanking each of our identified genes to assess the extent to which elevated levels of polymorphism reach into the flanking chromosomal region. Flanking genes were sampled from each side of our identified gene until levels of nucleotide diversity dropped near the *A. thaliana* mean.

PCR of *A. thaliana* and *A. lyrata* samples was performed using either *Taq* DNA polymerase (Roche, Indianapolis) or *ExTaq* DNA polymerase (Takara, Madison, WI). Amplified DNA fragments were purified using QIAquick PCR purification or gel extraction kits (QIAGEN). *A. thaliana* PCR products were cycle sequenced directly with Big Dye terminators and run on Prism 3700 96 capillary automated sequencers (Applied Biosystems, Foster City, CA) at the North Carolina State University Genome Research Laboratory. Amplified *A. lyrata* products were cloned using the TOPO TA PCR cloning kit (Invitrogen, San Diego), and plasmid DNA was isolated using the QIAaprep spin miniprep (QIAGEN). The presence of inserts in plasmid clones was confirmed by restriction digests using *Eco*RI and five to six independent clones were identified for sequencing. The PHRED and PHRAP functions (Ewing and Green 1998; Ewing *et al.* 1998) of Biolign (Tom Hall, North Carolina State University) were used in base calling and creating sequence contigs. All polymorphisms were visually confirmed, and questionable polymorphisms were rechecked through PCR reamplification and sequencing. Nucleotide sequence alignments and tables of polymorphic sites are available upon request.

Molecular population genetic analysis: Sequences were visually aligned using the *A. thaliana* Col-0 sequence as a reference. DnaSP 3.99 (Rozas *et al.* 2003) was used for intraspecific analysis of polymorphism data. Nucleotide diversity was estimated for silent sites as both π (Tajima 1983) and θ_w (Watterson 1975). MEGA2.0 (Kumar *et al.* 2001) was used to calculate interspecific silent-site nucleotide divergence (K) between Col-0 and one *A. lyrata* individual for each gene, using the Kimura two-parameter model. Tajima's D (Tajima 1993) and both Fu and Li's D and D^* (with and without outgroup, respectively) (Fu and Li 1993), haplotype number, and the intragenic linkage disequilibrium statistic Z_{nS} (Kelly 1997) were also estimated for each gene. Statistical significance of these estimates was determined by coalescent simulations with 10,000 runs, conditioning on the number of segregating sites and under the conservative assumption of no recombination. Levels of linkage disequilibrium both within and between genes in a region were estimated using the r^2 statistic based on informative sites (Hill and Robertson 1968), and significant associations were determined using Fisher's exact test.

The HKA (Hudson *et al.* 1987) test of selection was applied using the multilocus HKA program available from Jody Hey (http://lifesci.rutgers.edu/~heylab/HeylabSoftware.html). Only exon sequences were used in HKA tests for the expressed protein-encoding gene AT5G24310 since, for this gene, intron sequences between species were difficult to align with confidence. Complete sequences were analyzed in all other cases. Tests were based on silent sites and comparisons were made between each of our sequenced genes and a set of three neutral reference genes. These three reference loci, *PI*, *API*, and *FAH*, all have θ *vs.* K values that fall within the 95% confidence limit of the regression of genome-wide nucleotide diversity on interspecific divergence (Schmid *et al.* 2005; R. C. Moore and P. Awadalla, personal communication). Bonferroni corrections for multiple testing were applied for all tests of selection that were conducted across multiple linked genes in a given region. Allelic relationships were inferred using the neighbor-joining algorithm in MEGA 2.0 under the Kimura two-parameter substitution model and handling gaps and missing data as pairwise deletions.

Estimation of local recombination rates: Genetic markers were obtained from the Lister and Dean Col × L*er* recombinant inbred map (Lister and Dean 1993), and marker physical distances were obtained from the The Arabidopsis Information Resource database (ftp://tairpub:tairpub@ftp.arabidosis.org/home/tair/Maps/mapviewer_data). Markers with known genetic and physical map positions were ordered according to their physical positions and noncollinear markers were removed. Recombination rates were calculated between each pair of adjacent markers. Local recombination rate estimates were taken as the estimated rate between the two closest markers flanking the region of interest.

RESULTS

Genome scanning for high-diversity genes in *A. thaliana*: A large-scale comparison of ~2.5 Mb of L*er* genomic shotgun sequence fragments against the Col-0 whole-genome sequence provides a preliminary scan for high-diversity genes in the *A. thaliana* genome. In this species, silent-site nucleotide diversity has been estimated to be ~0.7% (Yoshida *et al.* 2003). On the basis of this consideration, we chose gene fragments with an interecotype divergence range of 2–10% as representing putative high-diversity loci. The low end of this range is comparable to the nucleotide diversity estimate for the *RPS5* disease resistance gene ($\pi = 2.5\%$), which has been shown to be under balancing selection (Tian *et al.* 2002). The high end of the range is slightly less than the interspecific divergence estimate ($K = 12\%$) between *A. thaliana* and its sister species, *A. lyrata* (Barrier *et al.* 2003). We also removed repetitive sequences (*e.g.*, transposable elements, duplicate genes), pseudogenes, and short, hypothetical genes from the list of putative high-diversity loci.

We identified a list of 60 functionally annotated genes ranging from 2–10% divergence between Col and L*er*. From this list we chose 11 genes that spanned the specified range of divergence; these were arbitrarily chosen on the basis of functional annotation (*e.g.*, transcription factor genes). Since the divergence estimates are based on raw shotgun genome sequence data from L*er*, it is possible that several of these putative high-diversity estimates arose from sequencing errors or the presence of a rare divergent allele. We conducted another round of screening to confirm which genes truly represent high-diversity loci that could be candidates for further study. Fragments of ~0.5–1.0 kb in length were se-

quenced in each of the 11 putative high-diversity genes in five to six additional ecotypes to confirm the observed elevated levels of polymorphism for these loci. Silent-site nucleotide diversity and Tajima's D and Fu and Li's D^* were estimated for the genes in this second screen. Genes were included for further analysis if they had (i) silent-site nucleotide diversity (π) >3% and (ii) positive Tajima's D or Fu and Li's D^* that were significantly higher than neutral-equilibrium expectations.

Of the 11 genes examined in the secondary screen, 3 were confirmed to have high levels of silent-site nucleotide diversity and significantly positive Tajima's D and/or Fu and Li's D^* test statistics. These three genes are: (i) AT1G63910, which encodes the *myb*-like transcription factor *MYB103*; (ii) AT5G24300, which encodes a putative soluble starch synthase I enzyme; and (iii) AT1G19700, which encodes a member of the homeobox-leucine zipper transcription factor gene family.

High-diversity genomic regions: One known signature of a balanced polymorphism is a peak of elevated nucleotide diversity centered on the target of selection (Tian *et al.* 2002). The high-diversity gene fragments identified in this study may represent this peak of elevated nucleotide diversity or may represent the effect of genetic hitchhiking with the target of selection at a linked locus. Alternatively, the observed high diversity may not be due to a balanced polymorphism, but may arise from alternative genetic/genomic or demographic factors. Discriminating among these alternatives requires a detailed examination of the levels and patterns of nucleotide variation not only in the three high-diversity genes identified in the genome scan, but also in an extended genomic region surrounding these loci. If these high-diversity genes are associated with balanced polymorphisms, one might expect a broad region of elevated nucleotide polymorphism spanning several genes, given the reduced effective recombination rate in the predominantly selfing *A. thaliana*. We thus isolated and sequenced ~0.5- to 1.2-kb fragments from these high-diversity genes and from five to eight flanking loci in 14–20 *A. thaliana* accessions across each of these genomic regions (Table 1). The orthologous gene fragments in the sister species *A. lyrata* were also isolated and sequenced to serve as an outgroup for comparison.

Our genome scan initially identified the gene *MYB103* (Li *et al.* 1999), located in the middle of the bottom arm of chromosome I, as a high-diversity locus. We have designated a 44.2-kb genomic region associated with *MYB103* (AT1G63910) as high-diversity region 1. This region contains 12 annotated loci, and, aside from *MYB103*, we sequenced and analyzed genes encoding a putative monodehydroascorbate reductase (AT1G63940), two C3HC4-type zinc-finger proteins (AT1G68900 and AT1G63840), and a *PRLI*-interacting factor-related protein (AT1G63850) that together span this region. One striking feature of this genomic region is the presence of two putative TIR-NBS-LRR-type disease-resistance genes that encode proteins with Toll and interleukin-1 receptor (TIR), nucleotide-binding site (NBS), and leucine-rich repeat (LRR) domains and one disease-resistance pseudogene tandemly located together in a gene block in the Col-0 accession. PCR primers designed from the Col-0 sequence to amplify the second TIR-NBS-LRR putative disease resistance gene were successful in only 9 of the 19 ecotypes attempted. Previous work has suggested that balancing selection can act on the presence/absence of alleles of disease resistance genes; our results suggest that this TIR-NBS-LRR cluster may be a target of selection, and the elevated nucleotide variation in this genomic region may result from genetic hitchhiking.

Elevated nucleotide diversity at a putative soluble starch synthase I gene on chromosome V was used to identify high-diversity region 2. The region we analyzed encompasses 59.5 kb and includes 12 annotated loci. We examined an additional seven genes in this region, including five annotated genes that have no known function, but which all show evidence of being transcriptionally expressed. EST analyses indicate that two of these genes are associated with full-length cDNAs (AT5G24280 and AT5G24214); one gene is supported by a near full-length cDNA, which lacks only the first 20 bases of sequence (MOP9.15), and the remaining two genes are associated with an EST hit at least 500 bp in length (AT5G24210 and AT5G24250). Other genes in high-diversity region 2 that were sequenced include one encoding an integral membrane family protein (AT5G 24290) and a protein containing a 3′–5′ exonuclease domain (AT5G24340).

High-diversity region 3, in the middle of the top arm of chromosome I, was identified in the genome scan by elevated nucleotide polymorphism in a gene encoding a homeobox-leucine zipper family protein. We analyzed this region, which spans 21.3 kb and includes five annotated loci. Other sequenced genes in this region include a jacalin lectin family protein (AT1G19715), a glycosyl transferase family 1 protein (AT1G19710), and two other expressed proteins of unknown function (AT1G19690 and AT1G19680).

Silent-site nucleotide variation across high-diversity genomic regions: Increased nucleotide variation in high-diversity genomic regions 1 and 2 was not confined to the initially identified high-diversity gene. In both of these genomic regions, other genes also displayed elevated levels of nucleotide polymorphism. In high-diversity region 1, which spans 44.2 kb, we sequenced a total of 3441 nucleotide sites, including 1559 silent sites. A total of 195 single nucleotide polymorphisms (SNPs) were identified by the analysis, including 164 silent-site polymorphisms. Four of the five genes surveyed in high-diversity region 1 have silent-site nucleotide diversity levels (π) ranging from 0.022 to 0.089 (Table 2), which is 3- to 12-fold higher than the mean level of 0.007 observed from previously studied *A. thaliana* nuclear genes (Yoshida *et al.* 2003).

TABLE 1

Genes contained in each of the three identified high-diversity regions listed in 5′–3′ order from top to bottom

Locus	Gene product description
	High-diversity region 1
AT1G63940[a]	Monodehydroascorbate reductase (putative)
AT1G63930	Expressed protein
AT1G63920	Pseudogene, similar to putative AP endonuclease/reverse transcriptase
AT1G63910[a]	MYB family transcription factor (*MYB103*)
AT1G63900[a]	Zinc-finger (C3HC4-type RING finger) family protein
AT1G63880	Disease resistance protein (TIR-NBS-LRR class) (putative)
AT1G63870	Disease resistance protein (TIR-NBS-LRR class) (putative)
AT1G63860	Pseudogene, disease resistance protein
AT1G63857	Expressed protein
AT1G63855	Expressed protein
AT1G63850[a]	PRLI-interacting factor related
AT1G63840[a]	Zinc-finger (C3HC4-type RING finger) family protein
	High diversity region 2
AT5G24280[a]	Expressed protein
AT5G24290[a]	Integral membrane family protein
AT5G24300[a]	Starch synthase (putative)
AT5G24310[a]	Expressed protein
AT5G 24313	Expressed protein
AT5G24314[a]	Expressed protein
AT5G24316	Proline-rich family protein
MOP9.15[a,b]	Expressed protein
AT5G24320	WD-40 repeat family protein
AT5G24330	PHD finger family protein/SET domain-containing protein
AT5G24340[a]	3′–5′ exonuclease domain-containing protein
AT5G24350[a]	Expressed protein
	High diversity region 3
AT1G19715[a]	Jacalin lectin family protein
AT1G19710[a]	Glycosyl transferase family 1 protein
AT1G19700[a]	Homeobox-leucine zipper family protein
AT1G19690[a]	Expressed protein
AT1G19680[a]	Expressed protein

[a] Genes included in analysis.
[b] BAC locus annotation for a predicted gene with EST support.

This increased intraspecific nucleotide diversity could reflect increased neutral mutation rates for these loci. Variation in neutral mutation rates should be mirrored by differences in interspecific nucleotide substitution rates, and we therefore examined the silent-site nucleotide divergence (K) between these *A. thaliana* genes and their *A. lyrata* orthologs. Silent-site interspecific nucleotide divergence (K) estimates for these genes range from 0.09 to 0.24; the mean K between *A. thaliana* and *A. lyrata* is 0.12 (Barrier *et al.* 2003). Figure 1 depicts the ratio of θ/K across this region; the mean value for previously studied *A. thaliana* genes is ~0.06 (Barrier *et al.* 2003). The ratio θ/K is three- to sevenfold higher than the mean for *A. thaliana* nuclear genes, and there appears to be a peak of diversity surrounding the putative disease resistance genes.

High-diversity region 2 contains the putative soluble starch synthase I gene (AT5G24300), and nucleotide variation in this region is also elevated across multiple linked loci. We sequenced 6528 nucleotide sites from eight genes, including 3861 silent sites across the 59.5-kb region. A total of 265 SNPs were identified by the analysis, including 240 silent-site polymorphisms. Values of silent site π for the putative soluble starch synthase I gene and the three loci immediately downstream (which all encode expressed proteins of unknown function) are all high. Estimates of π for these genes range from 0.024 to 0.056, values three- to eightfold higher than the mean for *A. thaliana* (Table 2). Interspecific nucleotide divergence estimates for the genes in this region range from 0.07 and 0.18, and the ratio of θ/K is shown in Figure 2. Like high-diversity region 1, a peak of elevated nucleotide polymorphism is also observed in this genomic region. An expressed protein gene

TABLE 2

Measures of diversity for the sampled genes in each high-diversity region

Gene	n^a	Length (bp)[b]	No. of silent sites	S[c]	S (silent)[d]	π[e]	θ_w[e]	K[f]
			High-diversity region 1					
AT1G63940	16	919	642.17	34	34	0.022	0.016	0.094 ± 0.012
AT1G63910	18	411	82.18	54	24	0.089	0.085	0.236 ± 0.049
AT1G63900	14	846	483.69	80	80	0.068	0.052	0.130 ± 0.017
AT1G63850	18	863	258.46	24	23	0.024	0.026	0.122 ± 0.021
AT1G63840	19	402	92.50	3	3	0.005	0.009	0.149 ± 0.036
			High-diversity region 2					
AT5G24280	17	711	340.2	6	2	0.002	0.001	0.173 ± 0.024
AT5G24290	18	704	341.7	5	4	0.002	0.003	0.122 ± 0.019
AT5G24300	18	1240	793.6	63	62	0.035	0.023	0.071 ± 0.008
AT5G24310	18	681	395.61	23	22	0.024	0.016	0.179 ± 0.027
AT5G24314	17	827	728.6	107	104	0.056	0.042	0.151 ± 0.017
MOP9.15[g]	15	761	250.89	45	29	0.044	0.036	0.143 ± 0.024
AT5G24340	15	797	416.38	3	1	0.001	0.001	0.086 ± 0.015
AT5G24350	16	807	594.24	18	16	0.006	0.008	0.124 ± 0.015
			High-diversity region 3					
AT1G19715	18	646	237.4	7	5	0.002	0.006	0.132 ± 0.023
AT1G19710	18	896	339.9	5	4	0.003	0.003	0.134 ± 0.019
AT1G19700	20	990	411.3	50	43	0.051	0.032	0.124 ± 0.017
AT1G19690	17	576	425.6	13	12	0.006	0.008	0.112 ± 0.017
AT1G19680	17	542	127.3	5	2	0.002	0.005	0.114 ± 0.027

[a] Number of samples.
[b] Length of sequenced region.
[c] Number of segregating sites in the sample.
[d] Number of segregating silent sites in the sample.
[e] Estimates are based on silent sites.
[f] Divergence between the *A. thaliana* Colombia-0 accession and *A. lyrata* based on silent sites including standard errors.
[g] BAC locus annotation for a predicted gene with EST support.

(AT5G24310) within this putative diversity peak also has elevated levels of intraspecific nucleotide variation (silent site $\pi = 0.024$, $\theta = 0.016$), but an elevated interspecific divergence estimate for this locus (silent site $K = 0.179$) results in a low level of θ/K. The four genes immediately upstream and downstream of this diversity peak, however, have θ/K ratios similar to or lower than the mean for *A. thaliana*. Both high-diversity genomic regions 1 and 2 share the pattern of multiple linked genes of elevated nucleotide polymorphism.

The pattern of nucleotide variation in high-diversity region 3, which includes the homeobox-leucine zipper transcription factor gene (AT1G19700) identified in the genome scan, differs from that observed in the other two regions. We sequenced 3650 nucleotide sites from five genes in this 21.3-kb region, including 1542 silent sites. A total of 80 SNPs were identified by the analysis, including 66 silent-site nucleotide polymorphisms. In high-diversity region 3, only the homeobox-leucine zipper gene has elevated levels of nucleotide diversity (silent site $\pi = 0.05$); this increased diversity is still evident in the θ/K ratio for this locus (Figure 3 and Table 2). The four other genes in this region have levels of nucleotide variation ranging from 0.002 to 0.006, which are all lower than the mean for *A. thaliana* nuclear genes (Figure 3 and Table 2).

Nonsynonymous nucleotide variation across high-diversity gene regions: Increased levels of nonsynonymous variation can be associated with balanced polymorphisms, and this pattern is exemplified by plant disease resistance and self-incompatibility loci (Bergelson *et al.* 2001; Charlesworth *et al.* 2003). Nonsynonymous polymorphism, however, is not generally high across the three high-diversity regions (Table 2). In high-diversity region 1, only the *MYB103* gene (AT1G63910) shows elevated levels of nonsynonymous polymorphism; 30 nonsynonymous polymorphisms exist in our sequenced portion of this gene, resulting in nonsynonymous nucleotide diversity of ~2.7%. Similarly, only one gene in high-diversity region 2, the expressed protein encoding gene MOP9.15, shows elevated nonsynonymous polymorphism with 16 replacement polymorphisms

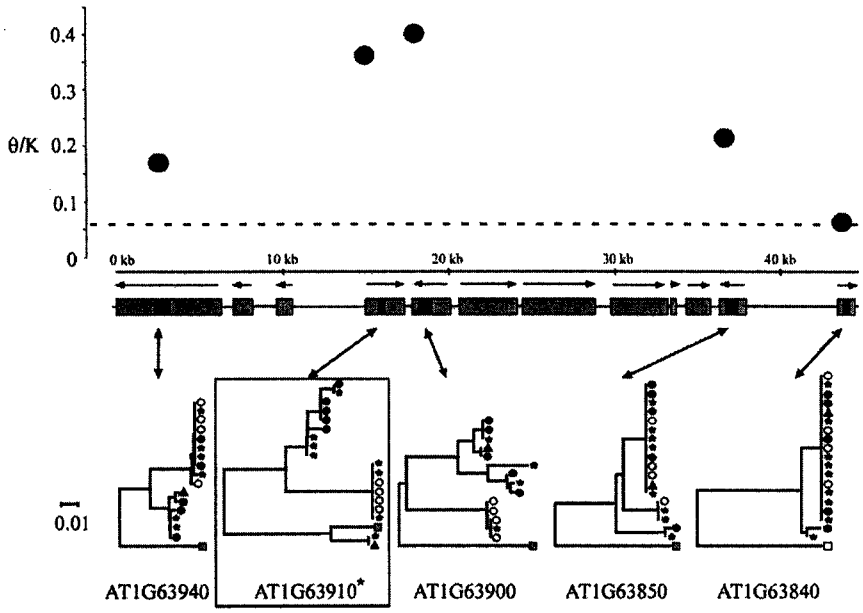


FIGURE 1.—Nucleotide diversity and neighbor-joining trees of high-diversity region 1. The dashed line indicates the average level of θ/K for *A. thaliana*. Sequenced sites from each gene are solid. The *MYB103* gene originally identified by our BLAST analysis is indicated by an asterisk. Symbols in neighbor-joining trees were applied to ecotypes on the basis of membership in a specific haplogroup in the most structured gene in the region, indicated by a boxed tree. Solid circles represent accessions possessing the Col-0-type allele, open circles represent *Ler*-type alleles, triangles are used if a third *A. thaliana* allele class is present, shaded squares represent the *A. lyrata* outgroup sequence, and stars indicate accessions that are not common across all genes in each region.

in the sequenced portion of the gene and estimated nonsynonymous nucleotide diversity of 1%. Although increased nonsynonymous variation has been previously identified in genes thought to harbor balanced polymorphisms, the presence of polymorphism at this class of sites is dependent on the mechanism of selection. Patterns of variation at nonsynonymous sites can also be influenced by other evolutionary mechanisms; for example, differing levels of constraint can allow for different patterns of variation at nonsynonymous sites to emerge between genes or regions of genes (BERGELSON *et al.* 2001). Therefore, silent-site polymorphism is likely

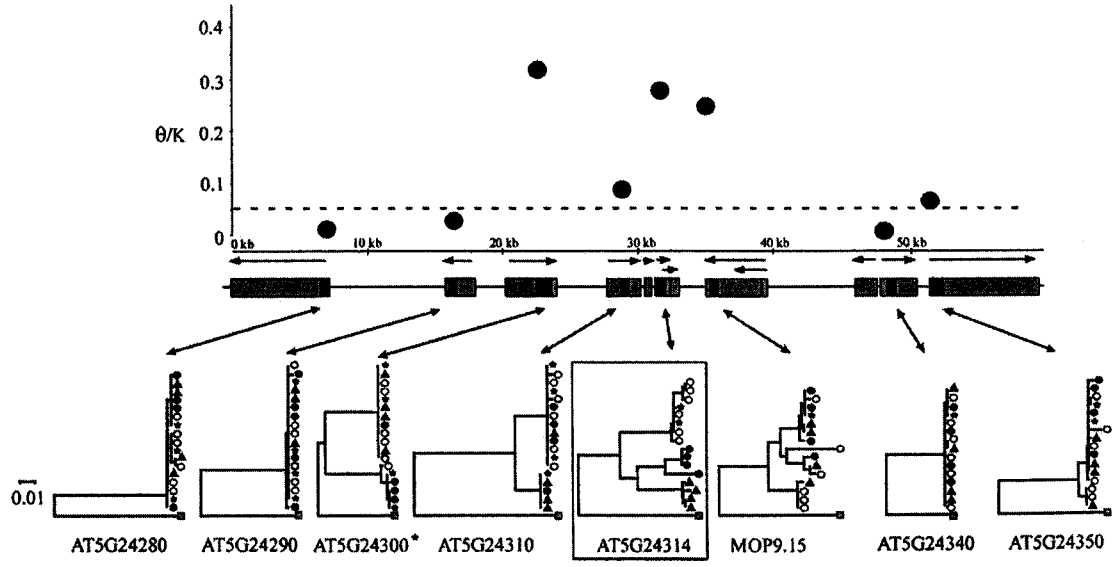


FIGURE 2.—Nucleotide diversity and neighbor-joining trees of high-diversity region 2. The dashed line indicates the average level of θ/K for *A. thaliana*. Sequenced sites from each gene are solid. The putative soluble starch synthase I gene originally identified by our BLAST analysis is indicated by an asterisk. Symbols are as described in Figure 1.

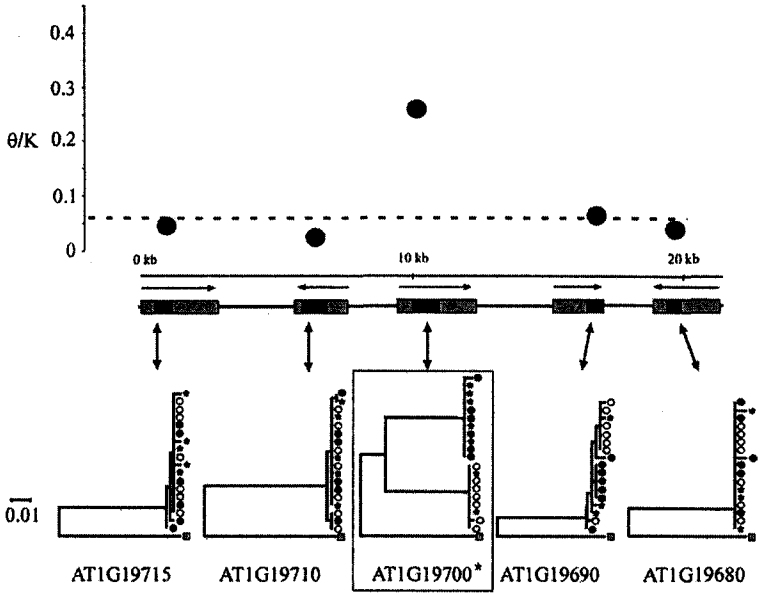


FIGURE 3.—Nucleotide diversity and neighbor-joining trees of high-diversity region 3. The dashed line indicates the average level of θ/K for *A. thaliana*. Sequenced sites from each gene are solid. The homeobox-leucine zipper gene originally identified by our BLAST analysis is indicated by an asterisk. Symbols are as described in Figure 1.

a more informative tool in delimiting a selected region of the genome and we focus further analyses on this category of sites.

Significant departures from the neutral-equilibrium model and genome-wide variation levels for high-diversity genes: The HKA test of selection is based on the assumption that under the neutral-equilibrium model, levels of intraspecific diversity and interspecific divergence should be governed by the neutral mutation rate and thus be correlated (HUDSON *et al.* 1987). The multilocus HKA test can be used to compare genes of interest to a neutral set of loci, thereby taking into account levels of neutral variation and divergence from multiple loci in the genome.

Multilocus HKA tests were applied to each of the genes in the three high-diversity regions, with a set of three previously studied genes as neutral reference loci (see Table 3). Significant deviations from neutrality were assessed following Bonferroni correction for multiple testing. The multilocus HKA tests indicate that two genes in high-diversity region 1, the AT1G63910 and AT1G63900 loci, have significantly elevated levels of intraspecific diversity ($P < 0.00583$ and 0.00525, respectively). For high-diversity region 2, the starch synthase gene ($P < 0.001$) and two expressed protein genes ($P <$ 0.00025 and 0.00558) are significant. In high-diversity region 3, the homeodomain gene shows significant departures from the expectations of the neutral-equilibrium model ($P < 0.00382$) (Table 3). In these cases, the significant results arise from both observed high intraspecific diversity and low interspecific divergence compared to expected values (data not shown).

The multilocus HKA tests indicated departures from a neutral-equilibrium model. We also determined where these genes fell within the distribution of θ/K studied in a genome-wide survey of variation in the *A. thaliana* genome (SCHMID *et al.* 2005). This genome-wide survey examined nucleotide variation in 195 unlinked, randomly selected gene fragments of ~400 bp in length from 12 ecotypes in SCHMID *et al.* (2005). Of these 195 gene fragments, 68 fragments had both intraspecific nucleotide diversity estimates in *A. thaliana* and interspecific nucleotide divergence between *A. thaliana* and *A. lyrata*, and this subset formed the basis of our θ/K distribution (see Figure 4). All genes identified as departing from neutral-equilibrium expectations in the multilocus HKA test were also found to be in the extreme tail of the genome-wide distribution (top 5%) of θ/K ratios for the genome-wide distribution (see Figure 4), indicating that these genes have exceptionally high levels of nucleotide variation compared to other loci in the *A. thaliana* genome.

Haplotype di- and trimorphism of high-diversity genes: The number of haplotypes in several genes in each of these high-diversity regions is significantly lower than expected under a conservative neutral-equilibrium model of no recombination (see Table 3); many of these are significant even with Bonferroni correction for multiple tests. This arises, in part, because all high-diversity genes identified in this study are organized into two or three distinct allele groups; these are referred to as di- and trimorphic haplotype structuring, respectively (Figures 1–3). Moreover, at least one gene, the *myb*-like gene (AT1G63910) in high-diversity region 1, also exhibits *trans*-specific polymorphism, with one of three allele classes (represented by two sampled *A. thaliana* accessions)

TABLE 3

Tests of selection for each sampled gene in all three high-diversity regions

Gene	n^a	Tajima's D	Fu and Li's D^*	Fu and Li's D	No. of haplotypes	Z_{ns}	HKA test (multilocus)[c]
			High-diversity region 1				
AT1G63940	16	1.6613*	0.8888	1.3556	7	0.6388*	0.39277
AT1G63910	18	0.3111	1.5013**	1.7841**	6***	0.5209*	0.00583**[d]
AT1G63900	14	1.4549	0.8756*	1.0912	10	0.5480*	0.00525**[d]
AT1G63850	18	−0.1764	1.1704	1.3269	5	0.4921	0.45463
AT1G63840	19	−1.1260	−0.1052	−0.6448	3	1.0000***	0.81329
			High-diversity region 2				
AT5G24280	17	0.5368	0.5941	0.5823	6	0.1723	0.01431*
AT5G24290	18	−1.1871	−1.1379	−0.8275	6***	0.4375	0.05413
AT5G24300	18	2.2290**	1.4535**	1.9336**	7***	0.9304***	0.001***[d]
AT5G24310	18	1.4491	1.1463	0.9705	4**	1.0000***	0.15324
AT5G24314	17	1.5564*	0.9236	1.2390	11	0.4126	0.00025***[d]
MOP9.15[b]	15	0.6575	0.2994	0.9110	10	0.3536	0.00558**[d]
AT5G24340	15	−0.0269	1.0566***	0.9025	4***	0.0452	0.01435*
AT5G24350	16	−1.0776	−0.8577	−0.8408	7	0.7851**	0.1492
			High-diversity region 3				
AT1G19715	18	−1.9331*	−2.3336	−1.3439*	8***	NA	0.63494
AT1G19710	18	−0.9452	−0.3590	0.1611	5	0.3500	0.14928
AT1G19700	20	2.9594***	1.3742*	1.5507*	5***	0.9613***	0.00382**[d]
AT1G19690	17	−1.0650	−1.7787*	−2.0333*	10*	0.2862	0.80176
AT1G19680	17	−1.9433*	−2.6319***	−2.7601*	6***	NA	0.57568

NA, no results were obtainable for this test with this sample. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$.

[a] Number of samples.

[b] BAC locus annotation for a predicted gene with EST support.

[c] Probability estimates based on silent sites for multilocus HKA test.

[d] Significant after Bonferroni correction.

more similar to the sequenced *A. lyrata* allele than to either of the other two *A. thaliana* allele classes.

Given the close linkage among the genes in these high-diversity regions and the reduced effective recombination rate in *A. thaliana* as a result of selfing, we would expect gene genealogies across these regions to be strongly correlated. Neighbor-joining trees show that phylogenetic relationships among adjacent genes are not perfectly correlated (Figures 1–3), indicating that intergenic recombination has occurred in these high-diversity regions to limit associations among di- or trimorphic haplogroups across loci. Patterns of linkage disequilibrium (LD; supplementary Figure S1 at http://www.genetics.org/supplemental/) observed across the investigated high-diversity regions, however, indicate that correlations between nucleotide polymorphisms exist among linked genes and are quite strong in high-diversity regions 1 and 2.

Excess of intermediate-frequency alleles in the high-diversity regions: Tajima's and Fu and Li's tests of selection were applied for all sequenced genes in each of the three high-diversity regions; these tests examine the frequency distribution of nucleotide polymorphisms along branches of a gene tree. Although these tests are often used to infer selection, they are also sensitive to population structure and demographic changes. Given the possible recent population size expansion, ancestral population structure, and selfing nature of *A. thaliana*, the results of these tests should be interpreted with caution. Despite these concerns, these tests prove useful in comparing patterns of polymorphism among *A. thaliana* genes and may also be indicative of the types of nonneutral forces acting at specific loci.

All three high-diversity-region genes with elevated levels of nucleotide polymorphism are accompanied by positive Tajima's D and/or Fu and Li's D/D^*, and multiple genes in high-diversity regions 1 and 2 show this pattern (Table 3). This is not surprising, given that the three focal genes identified in this study were initially chosen to have highly positive Tajima's D values. In high-diversity region 1, three of five genes have positive Tajima's D, while four have positive Fu and Li's D or D^*. In high-diversity region 2, five of eight genes have positive Tajima's D, and six of eight have positive Fu and Li's D/D^*. In high-diversity region 3, only the homeodomain-encoding gene has a positive value for these test statistics. Several of the genes with positive Tajima's D in these genomic regions are also in the top 5% tail

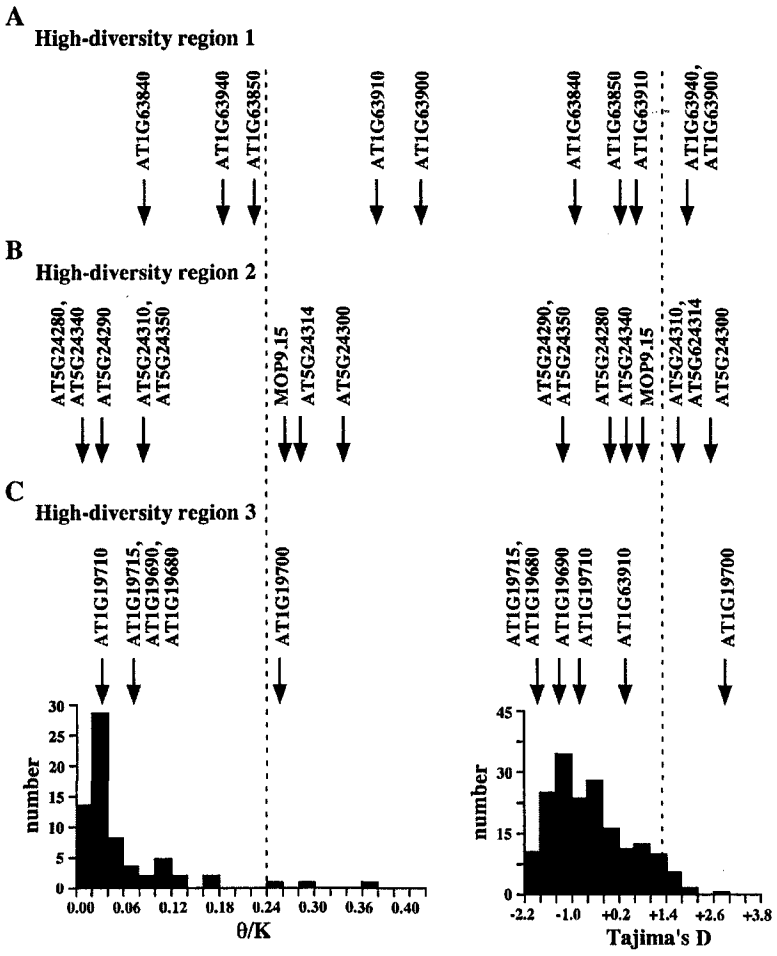


FIGURE 4.—Levels and patterns of nucleotide variation in high-diversity regions compared to a genome-wide distribution. Data for the distributions are from SCHMID *et al.* (2005). Data from 12 *A. thaliana* accessions were chosen to generate these distributions due to their use as mapping populations (Col-0, Cvi-0, L*er*, Nd-0, and Ws-0) and to obtain a maximum average genetic distance between surveyed accessions (Ei-2, CS22491, Gu-0, Lz-0, Wei-0, Ws-0, and Yo-0). The top 5% limits are indicated by dashed lines. A distribution of θ/K ratios for 68 randomly chosen unlinked gene fragments for which estimates of both intraspecific diversity within *A. thaliana* and interspecific divergence between *A. thaliana* and *A. lyrata* are available is shown on the bottom left. A distribution of Tajima's *D* for 195 randomly chosen unlinked gene fragments for which estimates of intraspecific diversity in *A. thaliana* only are available is shown on the bottom right. The location of the estimates within these distributions for genes in (A) high-diversity genomic region 1, (B) region 2, and (C) region 3 is shown by arrows.

of the distribution of this test statistic in a recent genome-wide survey of 195 unlinked gene fragments for which intraspecific data in *A. thaliana* are available (see Figure 4) (SCHMID *et al.* 2005). The latter result indicates that the numbers of intermediate-frequency polymorphisms in several of these high-diversity regions are exceptionally high compared to genes in the rest of the genome.

Linkage disequilibrium within and between genes: LD is a measure of genetic association at sites both within and between genes and is affected by a wide range of genetic, demographic, and selective factors (NORDBORG and TAVARE 2002; GAUT and LONG 2003). Z_{nS}, the standardized intragenic linkage disequilibrium averaged over all pairwise comparisons, is a test of selection that is expected to be significantly elevated if alleles at a locus are under balancing selection (KELLY 1997). We calculated Z_{nS} for each gene in all three high-diversity genomic regions, and significantly high Z_{nS} values were detected for genes in all three regions. In high-diversity region 1, all loci except for the *PRLI*-interacting-factor-related gene have significantly high Z_{nS} values (Table 3). Three genes in high-diversity region 2, the soluble starch synthase (AT5G24300) and two expressed protein genes (AT5G24310 and AT5G24350), also have significantly high Z_{nS} values. In contrast, only the homeobox-leucine zipper gene (AT1G19700) reveals a significantly elevated value of Z_{nS} in high-diversity region 3. For regions 2 and 3, the genes initially identified as having elevated levels of nucleotide variation [the soluble starch synthase gene (AT5G24300) in region 2 and the homeodomain gene (AT1G19700) in region 3] also have significantly high Z_{nS} estimates even after Bonferroni correction.

Linkage disequilibrium among informative polymorphic sites was also examined across each of the three high-diversity genomic regions using r^2, with significantly high levels of LD determined using Fisher's exact test. Statistically significant pairwise LD across each of the three regions is depicted in supplementary Figure

S1 at http://www.genetics.org/supplemental/. Significantly high levels of intragenic LD are observed in each of the three high-diversity regions, which conform to the high Z_{nS} estimates. Significantly high linkage disequilibrium between genes is also observed among genes with related neighbor-joining trees (see above) in high-diversity regions 1 and 2, which confirms previous findings of extensive LD in *A. thaliana* (NORDBORG *et al.* 2002; SHEPARD and PURUGGANAN 2003).

Local rates of recombination and patterns of variation in high-diversity regions: The extent of LD and the widths of the peaks of diversity observed in our three high-diversity regions should be affected by local recombination rates. We estimated local recombination rates for each of our three regions using information on genetic and physical map positions of markers flanking our genomic regions. The recombination rate in *A. thaliana* appears to range from 1 to 14 cM/Mb (ZHANG and GAUT 2003) and the genome-wide average was previously shown to be 4.8 cM/Mb (COPENHAVER *et al.* 1999; ZHANG and GAUT 2003).

The local recombination rate for high-diversity region 1 was estimated to be 1.75 cM/Mb, which is low in comparison to average chromosome and genome-wide estimates. This low recombination rate is consistent with the observation of a broad peak of nucleotide diversity and significant intergenic LD in high-diversity region 1 (see Figure 1). In contrast, high-diversity region 2 has an estimated recombination rate of 11.50 cM/Mb, which is high compared to average genome-wide estimates. Although elevated levels of nucleotide diversity in this region are observed among several linked genes, the peak is narrower than observed in genomic region 1 (see Figure 2).

The pattern observed in high-diversity region 3, however, appears anomalous. The local recombination rate for high-diversity region 3 is estimated at 2.17 cM/Mb, which is slightly lower than average estimates in this species. As such, we might expect to observe a broad peak of diversity across this region; what we observe, however, is a narrow peak that is centered on only one gene. This departure from expectation may result in higher recombination at smaller scales in this region. Alternatively, differences in peak breadths may result from other factors, including the age of alleles.

DISCUSSION

High-diversity genes represent an important class of genes in organismal genomes. Population genetic theory predicts that high-diversity genes may contain balanced polymorphisms (HUDSON and KAPLAN 1988), which underlie adaptive variation within species. These balanced polymorphisms are maintained over long evolutionary periods and are characterized by elevated levels of nucleotide diversity in silent sites linked to the selective target (HUDSON and KAPLAN 1988). This prediction has been substantiated by studies of several genes known to be subject to balancing selection or local adaptation, such as plant disease resistance (TIAN *et al.* 2002) and self-incompatibility loci (TAKEBAYASHI *et al.* 2003). Hunting for high-diversity genes could thus form the basis of an adaptive-trait-locus-mapping approach for scanning genomes for selectively maintained alleles.

Comparison of the *A. thaliana* Col-0 whole-genome sequence with 4300 short sequence fragments from a genomic shotgun sequence of the L*er* ecotype initially identified 60 functionally annotated sequences with 2–10% divergence between the two ecotypes. Further analysis in a secondary screen with five to six other *A. thaliana* ecotypes confirmed that three of these gene fragments—the *myb*-like transcription factor gene *MYB103* (AT1G63910), a putative soluble starch synthase I gene (AT5G24300), and a locus encoding a homeodomain-leucine zipper protein (AT1G19700)—have elevated nucleotide diversity levels and are high-diversity genes that may represent loci that have or are linked to balanced polymorphisms. The presence of high levels of nucleotide diversity, however, is only one characteristic of the signature of a balanced polymorphism. Additional characteristics can include: (i) a symmetrical peak of nucleotide diversity surrounding the selective target, (ii) maintenance of intermediate-frequency alleles, (iii) a reduction in the number of haplotypes, (iv) high levels of linkage disequilibrium, and (v) the presence of *trans*-specific polymorphism. The case for balanced polymorphisms is strengthened if the genes that harbor elevated levels of nucleotide diversity also display these other characteristics of selection, although it is worthwhile to note that it is unlikely for every expectation to be fulfilled by every empirical data set. Moreover, although many of these features are not totally independent of each other, they do represent different facets of an underlying pattern of sequence variation associated with selection.

Analysis of the high-diversity genes identified in the genome scan, as well as the loci flanking these genes, reveals that high-diversity region 1 displays all of the characteristic signatures of balanced polymorphisms. This region is characterized by elevated nucleotide variation spanning a local region of the genome, significant levels of intermediate-frequency polymorphisms, intergenic linkage disequilibrium, a significant deficiency in the number of haplotypes among highly variable genes, and *trans*-specific polymorphism at the *MYB103* locus (AT1G63910) in the region. High-diversity region 2 displays all of the characteristics of high-diversity region 1, with the exception of *trans*-specific polymorphism. This region also contains several genes with high levels of nucleotide diversity, among which the expressed protein gene AT5G24314 has significantly elevated nucleotide polymorphism levels as well as significantly positive Tajima's *D*. These features are consistent

with the hypothesis that both high-diversity regions 1 and 2 harbor balanced polymorphisms.

One gene in high-diversity region 3, the homeodomain-leucine zipper gene (AT1G19700), shows elevated nucleotide diversity and positive Tajima's *D* and Fu and Li's *D*/*D**. In addition, this gene also displays significant LD and a significant deficiency in haplotype number. Interpretation of these results is complicated, however, by the fact that high nucleotide polymorphism in this region is confined to one gene and does not show the gradual symmetric decline with distance; the four loci flanking the homeobox-leucine zipper gene have levels of nucleotide diversity at or below the mean for *A. thaliana* nuclear genes (Table 2 and Figure 3). Alternative explanations for the mechanism and/or origin of the two divergent haplogroups observed in this gene may help explain inconsistencies in these observations. Interestingly, a transposable element was identified in the intergenic region 3′ of the homeobox-leucine zipper gene in complete association with the L*er* haplogroup (J. M. Cork and M. D. Purugganan, unpublished observations). Potential functional consequences of this insertion and its effect on the molecular evolution of this region are currently being explored.

The levels and patterns of nucleotide polymorphisms, particularly those in high-diversity genomic regions 1 and 2, do not conform to the neutral-equilibrium model and are at the extremes for the genome-wide distributions, consistent with the selective maintenance of differentiated alleles. Other alternative possibilities, however, need to be considered. The first possibility is that these differentiated alleles represent ancestral and/or contemporary structure in *A. thaliana*. Recent studies suggest some isolation by distance as well as evidence for genetic differentiation associated with possible Pleistocene refugia in *A. thaliana* (Sharbel *et al.* 2000; Schmuths *et al.* 2004). Differentiation among dimorphic alleles attributed to population structure, however, is generally modest in other surveys of variation (Kawabe *et al.* 1997; Kuittinen and Aguade 2000) and does not explain the strong divergence of allelic classes in these high-diversity genomic regions. As a comparison, only 1 of 10 previously reported dimorphic genes in *A. thaliana* (*CLV2*) has nucleotide diversity levels in the top 5% of the genome-wide distribution. In contrast, the dimorphic genes in each high-diversity genomic region studied here all show exceptionally high variation levels (see Figure 5).

The second possibility is that the elevated diversity observed at these genomic regions could arise from gene duplications that result in artifactual comparisons of paralogous rather than allelic sequences. This duplication scenario also requires the loss of alternate duplicates such that *A. thaliana* ecotypes possess only one or the other duplicate copy. Although this may not be common, such a pattern is indeed possible; at the *MAM* locus of *A. thaliana*, alternate duplicate copies in a tandem array are lost in different ecotypes (Kroymann *et al.* 2003). However, this pattern, like the geographic structuring scenario, is improbable, given the elevated nucleotide diversity observed across multiple linked, unrelated loci in high-diversity regions 1 and 2. In contrast, this scenario cannot be ruled out for high-diversity region 3, where only one gene has elevated nucleotide polymorphism levels.

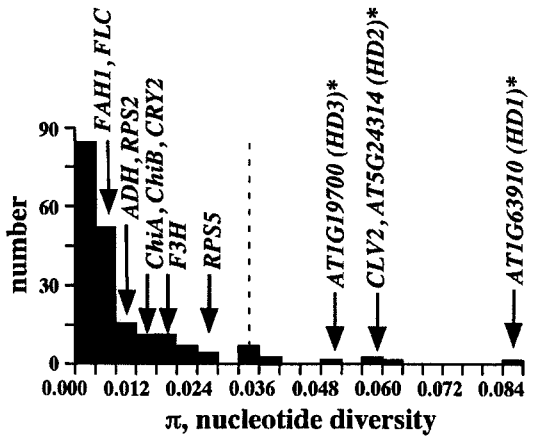


Figure 5.—Comparison of nucleotide diversity levels for dimorphic genes compared to a genome-wide distribution. Data for the distribution are from Schmid *et al.* (2005). The top 5% limit is indicated by a dashed line. Nucleotide diversity estimates (π) for 10 genes previously described as dimorphic are indicated by their names (Miyashita *et al.* 1996; Kawabe and Miyashita 1999; Aguade 2001; Tian *et al.* 2002; Mauricio *et al.* 2003; Caicedo *et al.* 2004; Olsen *et al.* 2004). One dimorphic gene from each of the three high-diversity regions is also shown.

It should be noted that both the geographic structure and the duplication scenarios are not mutually exclusive from a selection hypothesis. The former two scenarios relate to the *origins* of allelic differentiation, but selection can still be invoked to explain the intraspecific *maintenance* of these differentiated alleles. For example, geographic structure could explain the divergence in *CRY2* (Olsen *et al.* 2004) and *FLC* dimorphic haplotypes (Caicedo *et al.* 2004) and duplication accounts for differences in gene content and apparent nucleotide polymorphisms among genes at the *MAM* locus (Kroymann *et al.* 2003). In these cases, however, these different alleles are associated with trait variation in flowering time in the case of *CRY2* (Olsen *et al.* 2004) and *FLC* (A. L. Caicedo, J. R. Stinchmobe, K. M. Olsen, J. Schmitt and M. D. Purugganan, unpublished results) and with glucosinilate levels for the *MAM* locus (Kroymann *et al.* 2003), which suggests that maintenance of differentiated allele classes could result from selection of these ecologically relevant phenotypes. In each case, the precise mechanistic origins of alleles do not preclude the resultant phenotypic consequences that may lead to selective maintenance of alternate alleles.

While the genome-scan approach appears able to

identify high-diversity gene regions that may contain candidate balanced polymorphisms, identifying the specific polymorphism(s) and associated phenotype(s) that are possible targets of selection requires further work. In high-diversity region 1, the region of high diversity flanks tandemly repeated genes that belong to the TIR-NBS-LRR family of disease resistance loci. Disease resistance loci are known to be subject to various selective forces, including diversifying selection (BERGELSON *et al.* 2001). The presence/absence of deletion alleles, for example, is the basis for balancing selection at the disease resistance gene *RPS5* (TIAN *et al.* 2002). This previous knowledge and the pattern of variation observed in this region make the TIR-NBS-LRR genes the most likely putative target of selection in high-diversity region 1. Interestingly, PCR amplifications consistently fail to amplify the second TIR-NBS-LRR duplicate copy in this region in 9 of 19 *A. thaliana* ecotypes. Although this, as well as problems in designing copy-specific primers, makes it difficult to obtain completed sequence data sets for these putative disease resistance loci, these results suggest that this locus may segregate for the presence or absence of the second duplicate copy in Arabidopsis ecotypes. This finding also demonstrates the possible utility of the adaptive-trait-locus-mapping approach in exploiting the relationship between linkage disequilibrium and selection in identifying genes of potential adaptive significance when their direct sampling cannot be easily achieved.

The potential target of selection in high-diversity region 2 remains uncertain. One candidate is an expressed protein gene of unknown function (AT5G24314) that segregates for three distinct haplotype groups, yields a significant multilocus HKA test, and has the highest level of nucleotide polymorphism in the region. The precise function of this gene is unknown, but a T-DNA insertion mutant at this locus displays aberrant seed pigmentation associated with a defective embryo (BUDZISZEWSKI *et al.* 2001). Another candidate, however, is the putative starch synthase I locus, which is characterized by two highly divergent allele classes that show almost no polymorphism. This gene also has elevated levels of nucleotide diversity and highly positive Tajima's *D* estimates (see Figures 2 and 4).

Finally, only the homeodomain-leucine zipper transcription factor gene in high-diversity region 3 has high-diversity and intermediate-frequency alleles, although whether this is due to selection remains ambiguous, given that this gene does not show other key characteristics of a balanced polymorphism. At present, the precise function of this gene is unknown. Detailed functional reverse genetic studies are currently underway to determine the functions and the precise phenotypic consequences of the alternatively maintained alleles for all these candidate adaptive trait genes.

It is unclear how common these high-diversity genomic regions are in the genome. Moreover, not all balanced polymorphisms may have the extreme levels of diversity observed in this study, and thus this approach is inherently conservative. It remains important, however, to continue to identify and study high-diversity genes and genomic regions, and their possible contribution to adaptive variation. *A. thaliana* is particularly suited for these studies, given that the genomic resources (JANDER *et al.* 2002) and predominantly selfing nature of this species make it easier to identify high-diversity regions associated with balanced polymorphisms (NORDBORG *et al.* 1996; TIAN *et al.* 2002; SHEPARD and PURUGGANAN 2003). Detailed analysis of these genes may shed light on the extent of selection, as well as other evolutionary and genetic forces that act on these loci, and on their contribution to genome evolution.

We thank members of the Purugganan laboratory for a critical reading of this manuscript. We also thank M. Barrier for programming assistance and K. Shepard, R. C. Moore, P. Awadalla, M. Uyenonyama, and T. Mitchell-Olds for helpful discussions. This work was supported in part by National Science Foundation Integrated Research Challenges in Environmental Biology and Frontiers in Integrative Biological Research grants to M.D.P. and a National Institutes of Health Training Grant graduate fellowship to J.M.C.

LITERATURE CITED

ABBOT, R. J., and M. F. GOMES, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

AGUADE, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in Arabidopsis thaliana. Mol. Biol. Evol. **18:** 19.

AKEY, J. M., G. ZHANG, K. ZHANG, L. JIN and M. D. SHRIVER, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:**1805–1814.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796–815.

BAMSHAD, M., and S. P. WOODING, 2003 Signatures of natural selection in the human genome. Nat. Rev. Genet. **4:** 99–111.

BARBASH, D. A., P. AWADALLA and A. M. TARONE, 2004 Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus. PLoS Biol. **2:** 839–848.

BARRIER, M., C. D. BUSTAMANTE, J. YU and M. D. PURUGGANAN, 2003 Selection on rapidly evolving proteins in the Arabidopsis genome. Genetics **163:** 723–733.

BERGELSON, J., M. KREITMAN, E. A. STAHL and D. TIAN, 2001 Evolutionary dynamics of plant R-genes. Science **292:** 2281–2285.

BUDZISZEWSKI, G. J., S. P. LEWIS, L. W. GLOVER, J. REINEKE, G. JONES *et al.*, 2001 Arabidopsis genes essential for seedling viability: isolation of insertional mutants and molecular cloning. Genetics **159:** 1765–1778.

CAICEDO, A. L., J. R. STINCHCOMBE, K. M. OLSEN, J. SCHMITT and M. D. PURUGGANAN, 2004 Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc. Natl. Acad. Sci. USA **101:** 15670–15675.

CHARLESWORTH, D., 2003 Effects of inbreeding on the genetic diversity of populations. Philos. Trans. R. Soc. Lond. B Biol. Sci. **358:** 1051–1070.

CHARLESWORTH, D., and P. AWADALLA, 1998 Flowering plant self-incompatibility: the molecular population genetics of Brassica S-loci. Heredity **81** (Pt 1): 1–9.

CHARLESWORTH, D., C. BARTOLOME, M. SCHIERUP and B. K. MABLE, 2003 Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Mol. Biol. Evol. **20:** 1741–1753.

CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL

et al., 2003 Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302:** 1960–1963.

COPENHAVER, G. N., K. KUROMORI, T. BENITO, M. I. KAUL, S. LIN *et al.*, 1999 Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science **286:** 2468–2474.

DOEBLEY, J., A. STEC and L. HUBBARD, 1997 The evolution of apical dominance in maize. Nature **386:** 485–488.

EWING, B., and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:** 186–194.

EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

GARRIGAN, D., and P. W. HEDRICK, 2003 Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution **57:** 1707–1722.

GAUT, B. S., and A. D. LONG, 2003 The lowdown on linkage disequilibrium. Plant Cell **15:** 1502–1506.

GILAD, Y., S. ROSENBERG, M. PRZEWORSKI, D. LANCET and K. SKORECKI, 2002 Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl. Acad. Sci. USA **99:** 862–867.

GOMPEL, N., and S. B. CARROLL, 2003 Genetic mechanisms and constraints governing the evolution of correlated traits in drosophilid flies. Nature **424:** 931–935.

GREENBERG, A. J., J. R. MORAN, J. A. COYNE and C.-I WU, 2003 Ecological adaptation during incipient speciation revealed by precise gene replacement. Science **302:** 1754–1757.

HARR, B., M. KAUER and C. SCHLOTTERER, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

HILL, W. G., and A. ROBERTSON, 1968 The effects of inbreeding at loci with heterozygote advantage. Genetics **60:** 615–628.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

JANDER, G., S. R. NORRIS, S. D. ROUNSLEY, D. F. BUSH, I. M. LEVIN *et al.*, 2002 Arabidopsis map-based cloning in the post-genome era. Plant Physiol. **129:** 440–450.

JOHANSON, U., J. WEST, C. LISTER, S. MICHAELS, R. AMASINO *et al.*, 2000 Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. Science **290:** 344–347.

KAWABE, A., and N. T. MIYASHITA, 1999 DNA variation in the basic chitinase locus (ChiB) region of the wild plant *Arabidopsis thaliana.* Genetics **153:** 1445–1453.

KAWABE, A., H. INNAN, R. TERAUCHI and N. T. MIYASHITA, 1997 Nucleotide polymorphism in the acidic chitinase locus (ChiA) region of the wild plant Arabidopsis thaliana. Mol. Biol. Evol. **14:** 1303–1315.

KELLY, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

KOHN, M. H., H. J. PELZ and R. K. WAYNE, 2000 Natural selection mapping of the warfarin-resistance gene. Proc. Natl. Acad. Sci. USA **97:** 7911–7915.

KOHN, M. H., H. J. PELZ and R. K. WAYNE, 2003 Locus-specific genetic differentiation at Rw among warfarin-resistant rat (*Rattus norvegicus*) populations. Genetics **164:** 1055–1070.

KREITMAN, M. E., and M. AGUADE, 1986 Excess polymorphism at the Adh locus in *Drosophila melanogaster.* Genetics **114:** 93–110.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

KROYMANN, J., S. DONNERHACKE, D. SCHNABELRAUCH and T. MITCHELL-OLDS, 2003 Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc. Natl. Acad. Sci. USA **100** (Suppl. 2): 14587–14592.

KUITTINEN, H., and M. AGUADE, 2000 Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana.* Genetics **155:** 863–872.

KUMAR, S., K. TAMURA, I. B. JAKOBSEN and M. NEI, 2001 MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:** 1244–1245.

LI, S. F., T. HIGGINSON and R. W. PARISH, 1999 A novel MYB-related gene from Arabidopsis thaliana expressed in developing anthers. Plant Cell Physiol. **40:** 343–347.

LISTER, C., and C. DEAN, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *A. thaliana.* Plant J. **4:** 745–750.

LUIKART, G., P. R. ENGLAND, D. TALLMON, S. JORDAN and P. TABERLET, 2003 The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. **4:** 981–994.

MAURICIO, R., E. A. STAHL, T. KORVES, D. TIAN, M. KREITMAN *et al.*, 2003 Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana.* Genetics **163:** 735–746.

MIYASHITA, N. T., H. INNAN and R. TERAUCHI, 1996 Intra- and interspecific variation of the alcohol dehydrogenase locus region in wild plants Arabis gemmifera and Arabidopsis thaliana. Mol. Biol. Evol. **13:** 433–436.

NIELSEN, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641–647.

NORDBORG, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

NORDBORG, M., and S. TAVARE, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

NORDBORG, M., B. CHARLESWORTH and D. CHARLESWORTH, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond., Ser. B, Biol. Sci. **263:** 1033–1039.

NORDBORG, M., J. O. BOREVITZ, J. BERGELSON, C. C. BERRY, J. CHORY *et al.*, 2002 The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. **30:** 190–193.

OLSEN, K. M., S. S. HALLDORSDOTTIR, J. R. STINCHCOMBE, C. WEINIG, J. SCHMITT *et al.*, 2004 Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics **167:** 1361–1369.

ORR, H., and J. COYNE, 1992 The genetics of adaptation: a reassessment. Am. Nat. **140:** 725–742.

OTTO, S. P., 2000 Detecting the form of selection from DNA sequence data. Trends Genet. **16:** 526–529.

PAYSEUR, B. A., A. D. CUTTER and M. W. NACHMAN, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. **19:** 1143–1153.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. **132:** 365–386.

SCHIERUP, M. H., X. WEKEMANS and F. B. CHRISTIANSEN, 1998 Allelic genealogies in sporophytic self-incompatibility systems in plants. Genetics **150:** 1187–1198.

SCHLOTTERER, C., 2002 Towards a molecular characterization of adaptation in local populations. Curr. Opin. Genet. Dev. **12:** 683–687.

SCHMID, K. J., S. RASMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from the standard neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

SCHMUTHS, H., M. H. HOFFMANN and K. BACHMANN, 2004 Geographic distribution and recombination of genomic fragments on the short arm of chromosome 2 of Arabidopsis thaliana. Plant Biol. **6:** 128–139.

SCHULTE, P. M., H. C. GLEMET, A. A. FIEBIG and D. A. POWERS, 2000 Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. Proc. Natl. Acad. Sci. USA **97:** 6597–6602.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol. Ecol. **9:** 2109–2118.

SHEPARD, K. A., and M. D. PURUGGANAN, 2003 Molecular population genetics of the Arabidopsis CLAVATA2 region: the genomic scale of variation and selection in a selfing species. Genetics **163:** 1083–1095.

STAHL, E. A., G. DWYER, R. MAURICIO, M. KREITMAN and J. BERGEL-

SON, 1999 Dynamics of disease resistance polymorphism at the Rpm1 disease resistance locus of Arabidopsis. Nature **400:** 667–671.

STORZ, J. F., B. A. PAYSEUR and M. W. NACHMAN, 2004 Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol. Biol. Evol. **21:** 1800–1811.

STROBECK, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

SWANSON, W. J., A. G. CLARK, H. M. WALDRIP-DAIL, M. F. WOLFNER and C. F. AQUADRO, 2001a Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98:** 7375–7379.

SWANSON, W. J., Z. YANG, M. F. WOLFNER and C. F. AQUADRO, 2001b Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98:** 2509–2514.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1993 Statistical analysis of DNA polymorphism. Jpn. J. Genet. **68:** 567–595.

TAKEBAYASHI, N., P. B. BREWER, E. NEWBIGIN and M. K. UYENOYAMA, 2003 Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. **20:** 1778–1794.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A **99:** 11525–11530.

VIGOUROUX, Y., M. MCMULLEN, C. T. HITTINGER, K. HOUCHINS, L. SCHULZ *et al.*, 2002 Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc. Natl. Acad. Sci. USA **99:** 9650–9655.

WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

WOOTTON, J. C., X. FENG, M. T. FERDIG, R. A. COOPER, J. MU *et al.*, 2002 Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature **418:** 320–323.

YOSHIDA, K., T. KAMIYA, A. KAWABE and N. T. MIYASHITA, 2003 DNA polymorphism at the ACAULIS5 locus of the wild plant Arabidopsis thaliana. Genes Genet. Syst. **78:** 11–21.

ZHANG, L., and B. GAUT, 2003 Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res. **13:** 2533–2540.

Communicating editor: J. BERGELSON

CHAPTER THREE

Evidence for functional and geographic differentiation of three candidate balanced polymorphisms in *Arabidopsis thaliana*

Evidence for functional and geographic differentiation of three candidate balanced polymorphisms in *Arabidopsis thaliana*

Jennifer M. Reininga[1], Dahlia Nielsen[1] and Michael D. Purugganan[2*]

[1]Department of Genetics, Box 7614
North Carolina State University, Raleigh, NC 27695

[2]Department of Biology and Center for Comparative Functional Genomics, New York University, 1009 Silver, 100 Washington Square East, New York, NY 10003

*CORRESPONDING AUTHOR: Michael Purugganan. Telephone (212) 992 9628 and email: mp132@nyu.edu

Prepared for submission to *Molecular Biology and Evolution*

ABSTRACT

Molecular population genetic analysis of several broad chromosomal regions identified by an intraspecific genome scan in *Arabidopsis thaliana* suggested that balancing selection might operate to maintain variation at three novel candidate adaptive trait loci. These genes encode *SOLUBLE STARCH SYNTHASE I (SSI)*, *PLASTID TRANSCRIPTIONALLY ACTIVE 7(PTAC7)*, and *BELL-LIKE HOMEODOMAIN 10 (BLH10)*. If balanced polymorphisms are indeed maintained at these loci then we would expect to observe functional variation underlying the previously detected signatures of selection. We observe multiple replacement polymorphisms within and in the 32 amino acids just upstream of the protein-protein interacting BELL domain at the *BLH10* locus. In contrast, allele-specific expression differences were observed at both *SSI* and *PTAC7*. Geographic patterns of allelic differentiation are, for the most part, consistent with population structure findings for this species and a significant longitudinal cline was observed at all three candidate loci. These data support a hypothesis of balancing selection at all three candidate loci and provide a basis for more detailed functional work in the future by identifying possible functional differences that might be selectively maintained.

INTRODUCTION

Balancing selection maintains genetic variation within populations or species. Mechanisms of balancing selection have traditionally included overdominance (Tishkoff et al. 2001; Aidoo et al. 2002) and frequency-dependent selection (Charlesworth and Awadalla 1998), but now also encompass other selective mechanisms such as spatial-temporal selection (Tian et al. 2002; Charbonnel and Pemberton 2005), local adaptation to different environments (Schulte et al. 2000; Harr, Kauer, and Schlotterer 2002; Storz, Payseur, and Nachman 2004; Storz et al. 2007) and epistatic selection (Kroymann and Mitchell-Olds 2005) . This expansive definition of balancing selection results from the recognition that these various modes of selection result in similar molecular signatures at the genomic level, including elevated levels of nucleotide diversity that decrease symmetrically with distance to either side of the selected site(s). The intensity and breadth of this signal are highly dependent on the local rate of recombination, time since the selective event, and strength of selection (Charlesworth 2006).

Examples of balanced polymorphisms are best supported if one can establish the selective mechanism, provide molecular evidence for a signal of selection, and/or demonstrate functional or phenotypic effects of alternate putatively selected alleles. Expression phenotypes are often investigated for functional consequences of naturally-occurring genetic variation in gene regulation that may contribute to adaptive differences both within and between species (Crawford, Segal, and Barnett 1999; Schulte et al. 2000; Michalak et al. 2001; Bamshad et al. 2002; Rockman and Wray 2002; Lerman et al. 2003; Tian et al. 2003). Recent studies have identified molecular signatures of balancing

selection associated with differential expression phenotypes among allelic variants of the baboon homolog of human class II MHC locus *DQA1* (Loisel et al. 2006) and among *A. lyrata* self-incompatibility (SI) *S*-alleles (Prigoda, Nassuth, and Mable 2005). Alternatively, amino acid replacement polymorphisms that result in protein sequence changes could be maintained by balancing selection if alternate protein forms are favored at different points in space or time. In this regard, the example of local adaptation of mice to clines in oxygen availability with elevation though differentiated hemoglobin alleles presents a strong case. In addition to a clear molecular signature of balancing selection in the genomic region spanning the *Hb* locus, multiple replacement polymorphisms are observed in high linkage disequilibrium and with varying frequency from low to high elevations in deer mice, *Peromyscus maniculatus* (Storz et al. 2007).

The geographic distribution of alleles within and between populations can also provide further support for hypotheses of balancing selection and in identifying mechanisms that might be responsible for the maintenance of diversity. For example, temperature-dependent fitness differences among northern and southern populations of *Fundulus heteroclitus* have been associated with mutations in the *Ldh-B* locus promoter and suggests a role for this enzyme in local adaptation to the cold (Schulte et al. 2000). Alternatively, when frequency dependant selection is the driving evolutionary force, multiple alleles are often observed co-segregating within local populations across a broad species range. Plant self-incompatibility loci are one of the best documented examples of frequency dependent selection. SI alleles can persist for long evolutionary timescales and occur at intermediate frequencies within populations of a species due to the increased fitness effects of an allele when it becomes rare (Charlesworth and Awadalla 1998).

Genome scanning approaches to identify new targets of balancing selection have been largely unsuccessful in humans (Asthana, Schmidt, and Sunyaev 2005; Bubb et al. 2006), although specific signatures of local adaptation are clear (Kayser, Brauer, and Stoneking 2003; Storz, Payseur, and Nachman 2004; Voight et al. 2006; Williamson et al. 2007). The model plant species *A. thaliana*, however, may provide a better system for identifying a broader set of balanced polymorphisms (Tian et al. 2002) given its inbreeding nature, and resulting low effective rate of recombination (Abbot and Gomez 1989) and extensive linkage disequilibrium (Nordborg et al. 2005). Genome scanning has indeed identified the molecular signature of balancing selection at several candidate loci in this species (Cork and Purugganan 2005), including (i) increased levels of nucleotide diversity and (ii) intermediate frequencies of divergent alleles in three genomic regions. A detailed analysis of several identified regions revealed one putative disease resistance locus and three novel candidate genes as potential targets of balancing selection. These candidate genes include *SOLUBLE STARCH SYNTAHSE I (SSI)* and *PLASTID TRANSCRIPTIONALLY ACTIVE CHROMOSOME 7 (PTAC7)* (also known as *PIGMENT DEFECTIVE EMBYRO 225* or *PDE225*), which are linked in a common high-diversity region and the *BELL-LIKE HOMEODOMAIN 10 (BLH10*).

On the basis of homology *SSI* was previously identified as a member of the starch synthetic pathway and is highly conserved throughout the plant kingdom (Ral et al. 2004). Aberrant chain length distributions in amylopectin from plants lacking functional *SSI* indicates that this gene is a determinant of branching structure of amylopectin starch molecules (Delvalle et al. 2005). The protein product of the second candidate gene, *PTAC7*, was purified from the *A. thaliana* chloroplasts in association with the nuclear-

encoded single-subunit RNA polymerase (NEP), which suggests a function in the regulation of chloroplast gene expression (Pfalz et al. 2006). Consistent with this expectation, a pigment-defective knockout phenotype for this locus has also been observed (Budziszewski et al. 2001). The final candidate balanced polymorphism, *BLH10*, is a member of the BELL-like homeodomain family of transcription factors, which have conserved DNA-binding homeodomain and protein-interacting BELL domains (Becker et al. 2002). Members of this gene family have been shown to play important roles in various aspects of plant development, including ovule development (Ray et al. 1994), branching (Smith and Hake 2003) and internode patterning (Bao et al. 2004). Phylogenetic analysis indicates that *BLH10* is most closely related to an Arabidopsis-specific paralog, *BLH3* (Becker et al. 2002). Knockout mutant alleles of *BHLH10* and *BHLH3* have not yet been associated with any specific phenotypes, although these genes do show differences in expression patterns and no evidence for a relaxation of constraint (Duarte et al. 2006) suggesting the potential for non-redundant functions.

We examine levels and patterns of nucleotide variation across the sequences of these three candidate genes and examine the geographic distribution of alleles at each locus. Protein-level variation and differences in levels and patterns of allele specific gene expression are also investigated. Our findings support a hypothesis of balancing selection at these candidate loci by demonstrating the potential for functional differences and geographic patterning that may underlie the selective maintenance of allelic variants of these genes.

MATERIALS AND METHODS

Isolation and sequencing of alleles. Genomic DNA was extracted from young leaves of *A. thaliana* accessions chosen to span the natural historic geographic range of the species (Supplementary table 1) using the plant DNeasy mini kit (QIAGEN, Valencia, CA). PCR and sequencing primers were designed using Primer 3.0 software (http://fokker.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) with either the Col-0 genome sequence or conserved regions of existing partial gene alignments as a template (Supplementary Text 1). PCR was performed using either Taq (Roche, Indianapolis) or ExTaq (Takara, Madison, WI) and PCR products were cleaned using Qiagen PCR purification or gel extraction kits. PCR products were sequenced directly using Big Dye terminators and were run on ABI 3700 96 capillary automated sequencers (Applied Biosystems, Foster City, CA) at North Carolina State University's Genome Research Lab. PHRED and PHRAP functions (Ewing and Green 1998; Ewing et al. 1998) of BioLign (Tom Hall, North Carolina State University) were used for base calling and sequence contig formation. All polymorphisms were visually inspected and alignments were edited by hand.

Population genetic analysis. Parsimony trees used for haplotype network construction were generated using PAUP (Sinauer, Sunderland, MA). DNAsp version 4.10.7 (Rozas et al. 2003) was implemented to calculate summary statistics, measures of diversity and to conduct tests of selection. Nucleotide diversity was estimated using both total and silent sites as π (Tajima 1983) and as θ_w (Watterson 1975)for total sites only. Tajima's *D*

(Tajima 1993), haplotype number, haplotype diversity, and the intragenic linkage disequilibrium coefficient, Z_{nS} (Kelly 1997), were also calculated. Significance of Tajima's *D*, haplotype number, haplotype diversity, and Z_{nS} was determined by coalescent simulation under the conservative assumption of no recombination, using 10,000 runs and conditioning on the number of segregating sites. Sliding window analysis of nucleotide diversity was performed for each gene and the proportion of nonsynonsymous to synonymous nucleotide diversity was similarly investigated for *BLH10*. A window size of 100 bps and a step size of 20 bps was used for *BLH10*, while a window of 75 bps and a step size of 15 bps was used for *SSI* and *PTAC7/PDE225*. PolyPhen (http://genetics.bwh.harvard.edu/pph/) was used to assess the potential effects of replacement polymorphisms.

Allele specific expression. For each candidate balanced polymorphism, three heterozygous F1 lines were generated by crossing three different B allele (Ler-0 like) accessions to Col-0 A allele plants. In the case of the tri-allelic *PTAC7* two sets of three crosses were included to represent both A/B and the A/C allele comparisons (see Table S3). F_1 progeny were grown under long day (16-hour day length, 20^o C) conditions in 16-cell flats with subirrigation following a 3 week vernalization period at 4^o C. For each cross, three individuals were sampled for leaf, mixed-stage bud, and mixed-stage silique tissues by flash freezing in liquid nitrogen immediately after harvesting. Tissue samples were collected 21 days after planting or at the first point in development thereafter when all required tissues were available. mRNA extractions were performed for each of the three tissue samples for each individual using the Qiagen Plant RNA Extraction kit. 30

μL of each RNA prep (ranging in concentration between 500 and 800 ng/μL) were DNase treated using the DNA free kit according to manufacturer suggestions for rigorous treatment conditions (Ambion, Austin, TX). Maximum amounts of RNA (10 μL) from each DNase treated sample were used in each of two replicate cDNA reactions using Ambion's Retroscript kit (Ambion, Austin, TX) according to manufacturer specifications. Replicate cDNA samples were pooled and 1 μL of the combined cDNA preparation was used in each PCR reaction with 7.5 pmol of each, ExTaq polymerase, and ExTaq PCR reagents (Takara, Madison, WI). To test for the possibility of DNA contamination, several random DNase treated RNA samples were similarly subjected to PCR amplification. No product was observed in any of the tested samples, indicating the absence of DNA contamination in RNA samples subject to analysis. Four PCR and pyrosequencing reactions were included as technical replicates to assay the allele-differentiating SNPs of interest. Supplementary Text 2 lists the PCR and sequencing primers used in the analysis.

The proportion of alternate SNP nucleotides present in the sample was analyzed using a PSQ96 MA Pyrosequencer and Pyro Gold chemistry (Biotage, Uppsala, Sweden). SNP proportions were standardized by a within-sample monomorphic peak and then by an identical measure obtained from DNA. Since both alleles are present at 1:1 proportions in the DNA, any resulting differences in the observed quantity of alleles will reflect a bias in amplification during PCR amplification. By standardizing our cDNA measurements by a common DNA measure, we are assured that any observed differences in the proportion of transcripts are the result of a difference in expression only. Standardized measures of allele-specific expression were log-transformed normalize the

distributions. Tests for allele-specific expression were conducted using ANOVA, with the model $Y = \mu + T + L + A + T*L + L*A + A*T + A*L*T + I[L]random + A*I[L]random + T*I[L]random + A*T*I[L]random + error$, where Y is the log-transformed standardized allele expression level, T is the tissue type, L is the line or cross, A is the SNP allele, and I is the sampled individual.

Spatial distribution of alleles. Approximately 250 *A. thaliana* accessions were genotyped at allele-differentiating SNPs for each of the three candidate genes (Table S2). *A. thaliana* accessions from Siberia and Central Asia were provided by K. Schmidt (Leibniz-Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany). DNA was isolated from young plant leaves using the plant DNeasy mini kit (QIAGEN, Valencia, CA) and genotyped using cleaved amplified polymorphic sequences (CAPS) markers. CAPS markers were chosen by comparison of allele sequences in NEBCutter v2.0 (http://tools.neb.com/NEBcutter2/index.php). Genotyping primers were designed using Primer 3.0 software and gene products were subsequently digested with appropriate restriction enzymes (New England Biolabs, Ipswich, MA). Allele-specific digest patterns were resolved by agarose gel electrophoresis. The genotyping primers and restriction enzymes are listed in Supplementary Text 3.

Data for average January and July temperature and precipitation are as previously reported (Stinchcombe et al. 2004). Tests for local adaptation of alleles at candidate loci to annual high and low temperature and precipitation were performed by stepwise multiple logistic regression using JMP 5.0 (SAS, Cary, NC). Tests for significant geographic patterning of alternate alleles according to latitude and/or longitude were

similarly applied. The inverse distance weighted function of the Geographic Information System (GIS program, ArcMAP) was applied to the data according to the neighborhood method using 10-15 points to generate maps depicting predicted allele frequencies across Europe. The mean value was used if more than two points occur at the same map position.

RESULTS

Diversity across candidate genes. Full gene sequence data was collected from ~15 *A. thaliana* accessions (Table S2) to investigate levels and patterns of nucleotide variation across each of the three putatively selected loci. Table 1 summarizes data on measures of nucleotide diversity and tests of selection based on the site frequency spectrum. In general, levels of silent site nucleotide diversity (π_s) are in accord with previous findings (Cork and Purugganan 2005) and are consistent with expectations for a balanced polymorphism: (i) increased polymorphism relative to the *A. thaliana* genome average (~0.01) and (ii) a trend for variants to segregate at intermediate frequency (detected as a significantly positive Tajima's *D* value) were observed at all candidate loci.

The standardized intragenic linkage disequilibrium (LD) averaged across all pairwise comparisons, Z_{ns}, is expected to be significantly higher at loci subject to balancing selection (Kelly 1997). We observe significant elevations at both of the bi-allelic loci (*SSI* and *BLH10*), while the lack of significance at the trimorphic *PTAC7* likely reflects decreased power due to the small sample size and extreme variability of the locus. With the exception of Z_{nS}, deviations from neutrality (or from the genome

average) are generally more extreme than reported in our previous analyses (Cork and Purugganan 2005) due in part to our current analysis being based on full-gene sequence data instead of only ~1 kb gene fragments.

Consistent with previous findings and with observations of high intragenic linkage disequilibrium, the intermediate frequency variants at *SSI* and *BLH10* form two highly distinct allele classes, herein referred to as alleles A and B, where A refers to the Col-0 type and B to the Ler-0 type allele. A trimorphic haplotype pattern was observed at the *PTAC7* locus and allele categories for this gene are as previously described with a third C-type allele category (see Figure 1).

Interestingly, an ~810-bp polymorphic insertion/deletion (indel) in the *BLH10* gene, representing an insertion in several *A. thaliana* accessions relative to the Col-0 genome sequence, was identified in the intergenic region between BLH10 and its flanking gene, At1g19690. The indel was found to be associated with the *BLH10* B allele in all but two of the sampled accessions (Bs-1 from Switzerland and Ita-0 from Morocco). Sequence analysis reveals that this insertion shares 95% identity to a *copia*-like retrotransposon on chromosome 3 of the Col-0 reference sequence.

Variation in nucleotide polymorphism levels across gene regions. A sliding window analysis of nucleotide diversity was conducted to determine the distribution of nucleotide variation across the high-diversity region containing *SSI* and *PTAC7* (see Figure 2). Patterns in fluctuating levels of nucleotide diversity across these genes do not differ notably when only silent sites are considered. While both genes display several characteristics suggestive of balancing selection, it is most reasonable to hypothesize that

only one of these genes is the true target of selection in this genomic region. Indeed, significant intergenic linkage disequilibrium was previously observed between these two loci and suggests that they share correlated evolutionary histories, a finding in line with expectations of hitchhiking of neutral mutations with selected variants (Cork and Purugganan 2005). Prior analysis revealed decreased levels of variation and a non-significant HKA test result at the locus, At5g24310, which is located between these two candidate genes and it was therefore excluded from the present analysis.

Nucleotide variation across *SSI* appears to be lowest at the 5' end and increases to a ~3-fold higher level towards the 3'end of the gene. Levels of nucleotide diversity are much higher at the downstream *PTAC7* locus, which has a silent site diversity level greater than 4-fold higher than the genome-wide average (see Table 1)**.** Two peaks of diversity were localized to the upstream region and within the third intron of the *PTAC7* locus. Together, these data implicate *PTAC7* as a more likely target of selection in this genomic region (Figure 2) and opens the possibility that selection might operate on *cis*-acting regulatory elements that could be located in the upstream and/or intronic regions of this gene.

The variability in levels of nucleotide diversity was also assessed across a ~4 kb region that includes the *BLH10* gene, the 3' intergenic sequence including the identified retrotransposon insertion and the flanking gene, At1g19690 (see Figure 3). Peaks of nucleotide diversity are localized to the *BLH10* locus within which two peaks of nucleotide polymorphism levels are observed, one located over the portion of the gene encoding the BELL domain and another centered in the second intron. As gaps are ignored in our sliding window analysis, nucleotide diversity is depicted as being zero

across the section of the intergenic region corresponding to the large indel polymorphism representing the *copia*-like retroransposon. This insertion contains 35 segregating sites, and estimates of nucleotide diversity in this indel ($\pi = 0.017$) are slightly higher than the genome average (see Table 1) and suggest that this indel, although present at only intermediate frequencies in *A. thaliana*, does not represent a recent insertion.

Nonsynonymous variation at candidate genes. We examined nonsynonymous variation across each candidate gene to determine the potential for selective maintenance of differentially functioning proteins. Only two A/B allele differentiating nonsynonymous mutations were observed in *SSI* – a Gln (Q) to Glu (E) change in exon 2, and a Thr (T) to Ala (A) change in exon 13 (see Fig. 1). To see if these variants represent potentially functional amino acid substitutions, we used the online prediction program, PolyPhen, which uses structural data and/or an alignment-based methodology to determine if a substitution is likely to affect protein structure and function. Both substitutions that differentiate the A and B alleles of *SSI* are predicted to have no functional effects and are therefore unlikely to produce differential phenotypes that might be selectively maintained. The single allele differentiating replacement polymorphism observed in *PTAC7*, a Gly (G) to Arg (R) polymorphism in exon 2 that differentiates allele B from alleles A and C, is also predicted to have no functional effects.

At the *BLH10* locus, however, there is a peak of nonysynonymous/synonymous nucleotiode diversity centered at exon 1. We observed 5 allele-differentiating nonsynonymous substitutions and one 3 bp in-frame indel clustered within and 32 amino acids upstream of the protein interacting BELL domain (see Fig. 4). Interestingly, one of

these changes, a glutamine (Q) to histidine (H) substitution occurring within the conserved BELL domain is conserved among many (but not all) BELL-like homeodomain gene family members, and PolyPhen indicates that this may be a functional polymorphism. This suggests that alternate alleles of *BLH10* might possess variable functions that are dependent, in part, on protein structure/function differences.

Allele-specific expression of candidate genes. Differences in levels or patterns of expression among alternate alleles could be maintained by selection acting on divergent *cis*- regulatory elements; we thus tested for allele-specific expression of each of the three candidate genes. For each gene, three B allele-containing accessions (as well as three C allele accessions for the trimorphic *PTAC7* locus) were crossed to Col-0 (which had the A type allele) to provide three different heterozygous F1 progeny backgrounds. By examining the relative expression levels of alternate alleles within multiple genetic backgrounds we should be able to disentangle expression differences that result from *cis*- and *trans*-acting effects since only *cis*-acting changes should yield reproducible differences in allelic abundance across F1 lines. The proportion of transcripts attributed to each specific allele of a candidate locus in leaf, mixed-stage floral bud and silique tissues were assayed through detection of allele-differentiating SNPs via pyrosequencing.

No tissue-specific differences in allelic expression were observed, consistent with the ubiquitous expression patterns of these genes as reported by MPSS analysis (Meyers et al. 2004). Significant differences in transcript levels from each allele were detected, however, in two of the three candidate loci. The *SSI* locus shows a significant allele effect, with the Col-0 type A allele being expressed at higher levels throughout the plant

than the Ler-0 type B allele ($P < 0.0013$, ANOVA; see Table 2). Both allelic combinations (A vs. B and A vs. C) at the *PTAC7* locus also displayed significant allele-specific differences in expression, with the A vs. B contrast revealing a significant line x allele interaction ($P < 0.0166$, ANOXA; see Table 2) and the A vs. C comparison demonstrating an overall allele effect ($P < 0.0186$, ANOVA; see Table 2). The significant line x allele interaction for the A vs. B allelic comparison at *PTAC7* indicates that alternate alleles may behave differently in different genetic backgrounds, suggesting epistatic interaction of alternate *cis*- alleles with *trans*-acting factors. For both genes, however, there is observed variability in allelic abundance that can be accounted for solely by *cis*-acting polymorphism at the candidate loci, indicating regulatory differences between alternate alleles that could be selectively maintained.

The geographic distribution of alleles. Geographic patterns in the distribution of alleles at selected loci can provide insights into the evolutionary forces that act on genes. In *A. thaliana*, the distribution of alleles within and between populations is confounded by human disturbance and its homogenizing effects on variation among populations, particularly in agricultural regions of Central Western Europe (Mitchell-Olds and Schmitt 2006). Despite the lack of historically distinct populations across most of the geographic range of this species, several studies have demonstrated population structure and suggest that the present day species range was populated through expansions from asian and mediterranean pleistocene refugia (Sharbel, Haubold, and Mitchell-Olds 2000; Nordborg et al. 2005; Schmid et al. 2005; Ostrowski et al. 2006).

We analyzed the geographic distribution of alleles at *SSI*, *PTAC7* and *BLH10* by genotyping a set of ~250 *A. thaliana* accessions distributed largely across Eurasia. Multiple logistic regression revealed that allele identity was significantly correlated with latitude ($P < 0.01$) and longitude ($P < 0.003$) for *PTAC7*, and to longitude at *SSI* and *BLH10* ($P < 0.006$ and $P < 0.0001$, respectively). This latitudinal and longitudinal pattern patterning may result from a variety of factors including alternate refugial origins and subsequent spread of alleles, or local adaptation to different environments. Stepwise multiple logistic regression of allele type on average summer and winter temperature and precipitation shows no correlation of candidate gene alleles with these climatic variables.

A GIS analysis allows us to map the distribution of alleles across the *A. thaliana* geographic range in Eurasia (Figure 4). A longitudinal cline is evident across this geographic range, a pattern consistent with previous reports of a general trend for loss of allelic richness in the eastern portion of the geographic range for this species. This pattern is especially marked for the *BLH10* locus, whose allelic distribution is similar to previous indications of isolation by distance among *A. thaliana* accessions (Sharbel, Haubold, and Mitchell-Olds 2000; Nordborg et al. 2005; Schmid et al. 2005). The longitudinal cline, however, is less apparent for the linked *SSI* and *PTAC7* than that observed from an analysis using genomic-wide data (Nordborg et al. 2005; Schmid et al. 2005). Alternate alleles of these loci show an admixed pattern across a broader portion of the geographic range, and the observed longitudinal cline is not significant for either locus when peripheral accessions from Siberia are excluded from the regression analysis (unpublished data). Interestingly, both *SSI* alleles and two of the three *PTAC7* alleles, are observed across the full east-west extent of the geographic range, and ~50% of Siberian

populations contained multiple alleles of both genes even with the small sample sizes (2-5) investigated here. These observations suggest that evolutionary forces other than population structure, possibly selection acting to maintain variation within local geographic locations, might help to explain the spatial distribution of variants in the *SSI*/*PTAC7* high-diversity region.

DISCUSSION

High-diversity genes represent an important category of loci in organismal genomes since they may be targets of balancing selection. The relative importance of balancing selection in maintaining intraspecific variation is not yet well understood and is still debated among those who believe it may play a major role and those who favor a mutation-drift balance explanation (Ferreira and Amos 2006; Charlesworth, Miyo, and Borthwick 2007). In *A. thaliana* several genomic regions with levels and patterns of sequence diversity consistent with the presence of a balanced polymorphism have been identified and suggest a role for selection in maintaining diversity at these genes (Kroymann et al. 2003; Cork and Purugganan 2005; Kroymann and Mitchell-Olds 2005; Bakker et al. 2006). If selection is indeed acting at these loci then we would expect to find evidence for functional variation in addition to simply detecting molecular footprints of selection. In all three cases studied here, we observe either differences in allele-specific gene expression levels or differentiated protein variants that provide further support for the possibility of balancing selection acting at these candidate genes.

The *PTAC7* gene was identified as a high-diversity gene on chromosome 5 (Cork and Purugganan 2005) and our analysis indicates that it has high levels of variation in the promoter region and third intron that could potentially harbor *cis*-regulatory elements. Consistent with this analysis, our expression study reveals differences in the level of allele-specific expression at *PTAC7*. This gene encodes a protein that is associated with the plastid RNA polymerase, suggesting a role in chloroplast gene expression; mutant alleles of this gene produces a pigment defective phenotype (Budziszewski et al. 2001). There is some evidence that *PTAC7*expression might be upregulated in response to metabolite cues; MPSS signatures associated with this gene indicate a nearly two-fold increase in expression in leaves following treatment with salicylic acid, which may also suggest a possible involvement in a disease resistance (Ward et al. 1991) or more generalized stress response. It would be attractive to consider a role for this gene in the transcriptional reprogramming that is known to occur during disease resistance (Nimchuk et al. 2003), especially given the pre-established role of balancing selection at other disease resistance loci.

We also found a difference in the level of expression of alternate alleles at the *SSI* locus, which is linked to the *PTAC7* gene and also displays high levels of nucleotide polymorphism. Although the possible functional consequence of this expression variation is unknown, it is conceivable that the amount of *SSI* product could impact amylopectin branching patterns and consequently alter aspects of resource availability in the plant. The molecular signal of selection seems stronger, however, at the downstream *PTAC7* locus; it may be that *SSI* shares regulatory elements with *PTAC7*, or that the

observed variation in the level of expression is neutral, and thereby invisible to selection, but is the result of hitchhiking of neutral *cis*-regulatory variants at *SSI*.

In the case of *BHL10*, a sliding window analysis reveals a peak of diversity centered on multiple replacement polymorphisms located within and just upstream of the protein-protein interacting BELL domain. A recent yeast two-hybrid screen defined the protein interaction network involving the BELL-like gene family members and their interactors (Hackbusch et al. 2005). The *BLH10* protein was shown to be one of the most highly connected members of the network, interacting with 13 other known proteins. The binding affinity and specificity of these transcription factors are largely determined by the protein interactions they form. Our results, including identifying at least one radical gultamine to histidine amino acid substitution in the BELL domain, provide a basis for future investigations into allele-specific phenotypic effects once the precise function of *BLH10* is defined.

A polymorphic copia-like retrotransposon is located in the intergenic region downstream of the *BLH10* candidate locus, and it is known that large indel polymorphisms and repetitive elements can suppress recombination in local regions of the genome and facilitate the maintenance of differentiated alleles (Charlesworth, Langley, and Stephan 1986; Jaarola, Martin, and Ashley 1998; Casselman et al. 2000). This may produce a signature similar to that of a balanced polymorphism especially if alternate alleles are maintained in isolated populations for long evolutionary periods, as suggested in this case by the strong longitudinal cline at *BLH10*. It is unclear, however, why, in the absence of selection, so much excess variation would accumulate in this region and why it would be confined entirely to the *BLH10* locus. It is thought that

transposition events often accompany gene duplication (Bailey, Liu, and Eichler 2003)and can underlie adaptive evolution (Michalak et al. 2001). For this reason, transposable element insertions that occur at high frequency within a species are often considered candidate adaptive trait loci, particularly when they are observed in close proximity to known genes (Franchini, Ganko, and McDonald 2004).

We see no evidence for local adaptation that could maintain diversity at these candidate loci over broad geographic scales as determined by winter and summer temperatures or precipitation. There are, of course, many other environments to which plants could be locally adapted, although they cannot all be investigated without some prior knowledge about gene function. The lack of discrete historical populations in *A.thaliana* precludes traditional population genetic analyses to compare levels of variation within and between populations. We were however able to qualitatively compare the Eurasian allelic distribution of each of our candidate loci with genome-wide population structure predictions (Nordborg et al. 2005; Schmid et al. 2005). *BLH10* showed a similar pattern to genome-wide markers, while *SSI* and *PTAC7* seemed less differentiated along the east to west Eurasian axis than random gene markers.

Understanding the relative contribution of balancing selection to the adaptive process will be important if the maintenance of within species diversity is to be adequately explained. Genome scanning may be a useful tool for identifying new sets of genes subject to balancing selection and this methodology is advanced if the functional mechanisms responsible can be further characterized. Determining a possible molecular basis for allelic differentiation of putative balanced polymorphisms and assessments of

the spatial distribution of alleles are essential if we are to reveal the functional and ecological significance of these genes.

ACKNOWLEDGEMENTS

We would like to thank Trudy F. C. Mackay and Ian Dworkin for advice with statiscal analysis and Lisa Goering for her pyrosequencing protocols. We are greatful to the North Carolina State Univerity Genome Research Lab and Phytotron personnel for their assistance. We would also like to thank Karl Schmid for his generosity in sharing recent collections of *A. thaliana* seed and Jen Modliszewski for her help with RNA isolation. This work was supported by a NIH pre-doctoral training grant fellowship to J.M.R. and in part by a grant from the NSF Frontiers in Integrative Biological Research Program to M.D.P.

REFERENCES

Abbot, R. J., and M. F. Gomes, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

Aidoo, M., D. J. Terlouw, M. S. Kolczak, P. D. McElroy, F. O. ter Kuile, S. Kariuki, B. L. Nahlen, A. A. Lal, and V. Udhayakumar. 2002. Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet **359**:1311-1312.

Asthana, S., S. Schmidt, and S. Sunyaev. 2005. A limited role for balancing selection. Trends Genet **21**:30-32.

Bailey, J. A., G. Liu, and E. E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet **73**:823-834.

Bakker, E. G., C. Toomajian, M. Kreitman, and J. Bergelson. 2006. A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell **18**:1803-1818.

Bamshad, M. J., S. Mummidi, E. Gonzalez, S. S. Ahuja, D. M. Dunn, W. S. Watkins, S. Wooding, A. C. Stone, L. B. Jorde, R. B. Weiss, and S. K. Ahuja. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A **99**:10539-10544.

Bao, X., R. G. Franks, J. Z. Levin, and Z. Liu. 2004. Repression of AGAMOUS by BELLRINGER in floral and inflorescence meristems. Plant Cell **16**:1478-1489.

Becker, A., M. Bey, T. R. Burglin, H. Saedler, and G. Theissen. 2002. Ancestry and diversity of BEL1-like homeobox genes revealed by gymnosperm (Gnetum gnemon) homologs. Dev Genes Evol **212**:452-457.

Bubb, K. L., D. Bovee, D. Buckley, E. Haugen, M. Kibukawa, M. Paddock, A. Palmieri, S. Subramanian, Y. Zhou, R. Kaul, P. Green, and M. V. Olson. 2006. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics **173**:2165-2177.

Budziszewski, G. J., S. P. Lewis, L. W. Glover, J. Reineke, G. Jones, L. S. Ziemnik, J. Lonowski, B. Nyfeler, G. Aux, Q. Zhou, J. McElver, D. A. Patton, R. Martienssen, U. Grossniklaus, H. Ma, M. Law, and J. Z. Levin. 2001. Arabidopsis genes essential for seedling viability: isolation of insertional mutants and molecular cloning. Genetics **159**:1765-1778.

Casselman, A. L., J. Vrebalov, J. A. Conner, A. Singhal, J. Giovannoni, M. E. Nasrallah, and J. B. Nasrallah. 2000. Determining the physical limits of the Brassica S locus by recombinational analysis. Plant Cell **12**:23-33.

Charbonnel, N., and J. Pemberton. 2005. A long-term genetic survey of an ungulate population reveals balancing selection acting on MHC through spatial and temporal fluctuations in selection. Heredity **95**:377-388.

Charlesworth, B., C. H. Langley, and W. Stephan. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics **112**:947-962.

Charlesworth, B., T. Miyo, and H. Borthwick. 2007. Selection responses of means and inbreeding depression for female fecundity in Drosophila melanogaster suggest contributions from intermediate-frequency alleles to quantitative trait variation. Genet Res:1-7.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet **2**:e64.

Charlesworth, D., and P. Awadalla. 1998. Flowering plant self-incompatibility: the molecular population genetics of Brassica S-loci. Heredity **81 (Pt 1)**:1-9.

Cork, J. M., and M. D. Purugganan. 2005. High-diversity genes in the Arabidopsis genome. Genetics **170**:1897-1911.

Crawford, D. L., J. A. Segal, and J. L. Barnett. 1999. Evolutionary analysis of TATA-less proximal promoter function. Mol Biol Evol **16**:194-207.

Delvalle, D., S. Dumez, F. Wattebled, I. Roldan, V. Planchot, P. Berbezy, P. Colonna, D. Vyas, M. Chatterjee, S. Ball, A. Merida, and C. D'Hulst. 2005. Soluble starch synthase I: a major determinant for the synthesis of amylopectin in Arabidopsis thaliana leaves. Plant J **43**:398-412.

Duarte, J. M., L. Cui, P. K. Wall, Q. Zhang, X. Zhang, J. Leebens-Mack, H. Ma, N. Altman, and C. W. dePamphilis. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol **23**:469-478.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8**:186-194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8**:175-185.

Ferreira, A. G., and W. Amos. 2006. Inbreeding depression and multiple regions showing heterozygote advantage in Drosophila melanogaster exposed to stress. Mol Ecol **15**:3885-3893.

Franchini, L. F., E. W. Ganko, and J. F. McDonald. 2004. Retrotransposon-gene associations are widespread among D. melanogaster populations. Mol Biol Evol **21**:1323-1331.

Hackbusch, J., K. Richter, J. Muller, F. Salamini, and J. F. Uhrig. 2005. A central role of Arabidopsis thaliana ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. Proc Natl Acad Sci U S A **102**:4908-4912.

Harr, B., M. Kauer, and C. Schlotterer. 2002. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. Proc Natl Acad Sci U S A **99**:12949-12954.

Jaarola, M., R. H. Martin, and T. Ashley. 1998. Direct evidence for suppression of recombination within two pericentric inversions in humans: a new sperm-FISH technique. Am J Hum Genet **63**:218-224.

Kayser, M., S. Brauer, and M. Stoneking. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol **20**:893-900.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146**:1197-1206.

Kroymann, J., S. Donnerhacke, D. Schnabelrauch, and T. Mitchell-Olds. 2003. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci U S A **100 Suppl 2**:14587-14592.

Kroymann, J., and T. Mitchell-Olds. 2005. Epistasis and balanced polymorphism influencing complex trait variation. Nature **435**:95-98.

Lerman, D. N., P. Michalak, A. B. Helin, B. R. Bettencourt, and M. E. Feder. 2003. Modification of heat-shock gene expression in Drosophila melanogaster populations via transposable elements. Mol Biol Evol **20**:135-144.

Loisel, D. A., M. V. Rockman, G. A. Wray, J. Altmann, and S. C. Alberts. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5' cis-regulatory region. Proc Natl Acad Sci U S A **103**:16331-16336.

Meyers, B. C., D. K. Lee, T. H. Vu, S. S. Tej, S. B. Edberg, M. Matvienko, and L. D. Tindell. 2004. Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol **135**:801-813.

Michalak, P., I. Minkov, A. Helin, D. N. Lerman, B. R. Bettencourt, M. E. Feder, A. B. Korol, and E. Nevo. 2001. Genetic evidence for adaptation-driven incipient

speciation of Drosophila melanogaster along a microclimatic contrast in "Evolution Canyon," Israel. Proc Natl Acad Sci U S A **98**:13195-13200.

Mitchell-Olds, T., and J. Schmitt. 2006. Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature **441**:947-952.

Nimchuk, Z., T. Eulgem, B. F. Holt, 3rd, and J. L. Dangl. 2003. Recognition and response in the plant immune system. Annu Rev Genet **37**:579-609.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N. A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. 2005. The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3**:e196.

Ostrowski, M. F., J. David, S. Santoni, H. McKhann, X. Reboud, V. Le Corre, C. Camilleri, D. Brunel, D. Bouchez, B. Faure, and T. Bataillon. 2006. Evidence for a large-scale population structure among accessions of Arabidopsis thaliana: possible causes and consequences for the distribution of linkage disequilibrium. Mol Ecol **15**:1507-1517.

Pfalz, J., K. Liere, A. Kandlbinder, K. J. Dietz, and R. Oelmuller. 2006. pTAC2, -6, and -12 are components of the transcriptionally active plastid chromosome that are required for plastid gene expression. Plant Cell **18**:176-197.

Prigoda, N. L., A. Nassuth, and B. K. Mable. 2005. Phenotypic and genotypic expression of self-incompatibility haplotypes in Arabidopsis lyrata suggests unique origin of alleles in different dominance classes. Mol Biol Evol **22**:1609-1620.

Ral, J. P., E. Derelle, C. Ferraz, F. Wattebled, B. Farinas, F. Corellou, A. Buleon, M. C. Slomianny, D. Delvalle, C. d'Hulst, S. Rombauts, H. Moreau, and S. Ball. 2004. Starch division and partitioning. A mechanism for granule propagation and maintenance in the picophytoplanktonic green alga Ostreococcus tauri. Plant Physiol **136**:3333-3340.

Ray, A., K. Robinson-Beers, S. Ray, S. C. Baker, J. D. Lang, D. Preuss, S. B. Milligan, and C. S. Gasser. 1994. Arabidopsis floral homeotic gene BELL (BEL1) controls ovule development through negative regulation of AGAMOUS gene (AG). Proc Natl Acad Sci U S A **91**:5761-5765.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol **19**:1991-2004.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**:2496-2497.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds. 2005. A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169**:1601-1615.

Schulte, P. M., H. C. Glemet, A. A. Fiebig, and D. A. Powers. 2000. Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. Proc Natl Acad Sci U S A **97**:6597-6602.

Sharbel, T. F., B. Haubold, and T. Mitchell-Olds. 2000. Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol **9**:2109-2118.

Smith, H. M., and S. Hake. 2003. The interaction of two homeobox genes, BREVIPEDICELLUS and PENNYWISE, regulates internode patterning in the Arabidopsis inflorescence. Plant Cell **15**:1717-1727.

Stinchcombe, J. R., C. Weinig, M. Ungerer, K. M. Olsen, C. Mays, S. S. Halldorsdottir, M. D. Purugganan, and J. Schmitt. 2004. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci U S A **101**:4712-4717.

Storz, J. F., B. A. Payseur, and M. W. Nachman. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol **21**:1800-1811.

Storz, J. F., S. J. Sabatino, F. G. Hoffmann, E. J. Gering, H. Moriyama, N. Ferrand, B. Monteiro, and M. W. Nachman. 2007. The molecular basis of high-altitude adaptation in deer mice. PLoS Genet **3**:e45.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437-460.

Tajima, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol **10**:677-688.

Tian, D., H. Araki, E. Stahl, J. Bergelson, and M. Kreitman. 2002. Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A **99**:11525-11530.

Tian, D., M. B. Traw, J. Q. Chen, M. Kreitman, and J. Bergelson. 2003. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature **423**:74-77.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams, and A. G. Clark. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293**:455-462.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol **4**:e72.

Ward, E. R., S. J. Uknes, S. C. Williams, S. S. Dincher, D. L. Wiederhold, D. C. Alexander, P. Ahl-Goy, J. P. Metraux, and J. A. Ryals. 1991. Coordinate Gene Activity in Response to Agents That Induce Systemic Acquired Resistance. Plant Cell **3**:1085-1094.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7**:256-276.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen. 2007. Localizing Recent Adaptive Evolution in the Human Genome. PLoS Genet **3**:e90.

Table 1. Measures of diversity for each candidate gene.

Gene	n^a	Length (bp) ^{b}	S^c	$S(silent)_d$	π_s ^{d}	π ^{e}	θ_w ^{e}	Z_{nS}	No. of Haplotypes	Haplotype Diversity	Tajima's D
BLH10	14	1856	62	53	0.04182	0.01623	0.0105	0.8245***	8	0.833*	2.4083**
Copia-like transposon	9	684	35	35	0.01698	0.01698	0.0188	nd	6	0.89	0.4962
PTAC7/PDE225	15	1332	114	112	0.04940	0.03594	0.0263	0.4216	12	0.971	1.6057*
SSI	16	3803	101	96	0.01764	0.01118	0.0080	0.8662***	13	0.967	1.7100*

nd, not determined. *** P< 0.001, ** P< 0.01, * P<0.05

[a] Number of samples

[b] Length of sequence region excluding gaps

[c] Number of segregating sites

[d] Estimates are based on silent sites

[e] Estimates are based on total sites

Table 2. Mixed-Model ANOVA.

	SSI					*BLH10*				
	SS[a]	MS[b]	DF[c]	F-Ratio	P	SS[a]	MS[b]	DF[c]	F-Rratio	P
Fixed:										
Allele	0.3498	0.3498	1	31.8978	0.0013	0.1216	0.1216	1	4.2217	0.0857
Line	0.0058	0.0029	2	1.1630	0.3742	0.0108	0.0054	2	2.2762	0.1838
Tissue	0.0019	0.0010	2	0.7342	0.5002	0.0040	0.0020	2	0.8725	0.4428
Line x Tissue	0.1773	0.0044	4	3.4317	0.0432	0.0062	0.0015	4	0.6692	0.6256
Line x Allele	0.0095	0.0047	2	0.4322	0.6678	0.0804	0.0402	2	1.3944	0.3182
Tissue x Allele	0.2305	0.0115	2	3.0590	0.0844	0.0474	0.0237	2	1.3299	0.3008
Line x Tissue x Allele	0.0429	0.1072	4	2.8427	0.0718	0.1241	0.0310	4	1.7400	0.2058
Random:										
Individual [Line]	0.0150	0.0025	6	0.2942	0.9081	0.0142	0.0024	6	0.1775	0.9461
Individual x Allele [Line]	0.0659	0.0110	6	2.9119	0.0544	0.0277	0.0023	6	0.1293	0.9994
Individual x Tissue [Line]	0.0155	0.0013	12	0.3423	0.9623	0.1730	0.0288	12	1.6164	0.2253
Individual x Allele x Tissue [Line]	0.0453	0.0038	12	7.1064	<0.0001	0.2142	0.0179	12	11.1565	<0.0001
	PTAC7 A/B					*PTAC7* A/C				
	SS[a]	MS[b]	DF[c]	F-Ratio	P	SS[a]	MS[b]	DF[c]	F-Ratio	P
Fixed:										
Allele	0.0001	0.0001	1	0.0054	0.9441	0.2194	0.2194	1	10.250	0.0186
Line	0.0124	0.0062	2	2.2461	0.1870	0.0055	0.0028	2	0.6852	0.5395
Tissue	0.0027	0.0014	2	0.8078	0.4687	0.0059	0.0029	2	1.5680	0.2483
Line x Tissue	0.0015	0.0004	4	0.2188	0.9228	0.0044	0.0011	4	0.5899	0.6764
Line x Allele	0.0684	0.0342	2	8.7578	0.0166	0.0225	0.0112	2	0.5250	0.6164
Tissue x Allele	0.0040	0.0020	2	0.9317	0.4206	0.0458	0.0229	2	1.8463	0.1999
Line x Tissue x Allele	0.0086	0.0021	4	0.9866	0.4513	0.1080	0.0270	4	2.1752	0.1336
Random:										
Individual [Line]	0.0166	0.0028	6	0.8053	0.6168	0.0241	0.0040	6	0.3696	0.8499
Individual x Allele [Line]	0.0234	0.0039	6	1.8005	0.1817	0.1284	0.0019	6	1.7247	0.1985
Individual x Tissue [Line]	0.0203	0.0017	12	0.7797	0.6633	0.0224	0.0214	12	0.1504	0.9987
Individual x Allele x Tissue [Line]	0.0260	0.0022	12	1.8699	0.0416	0.1489	0.0124	12	3.2592	0.0003

[a] SS, sum of squares
[b] MS, mean square
[c] DF, degrees of freedom

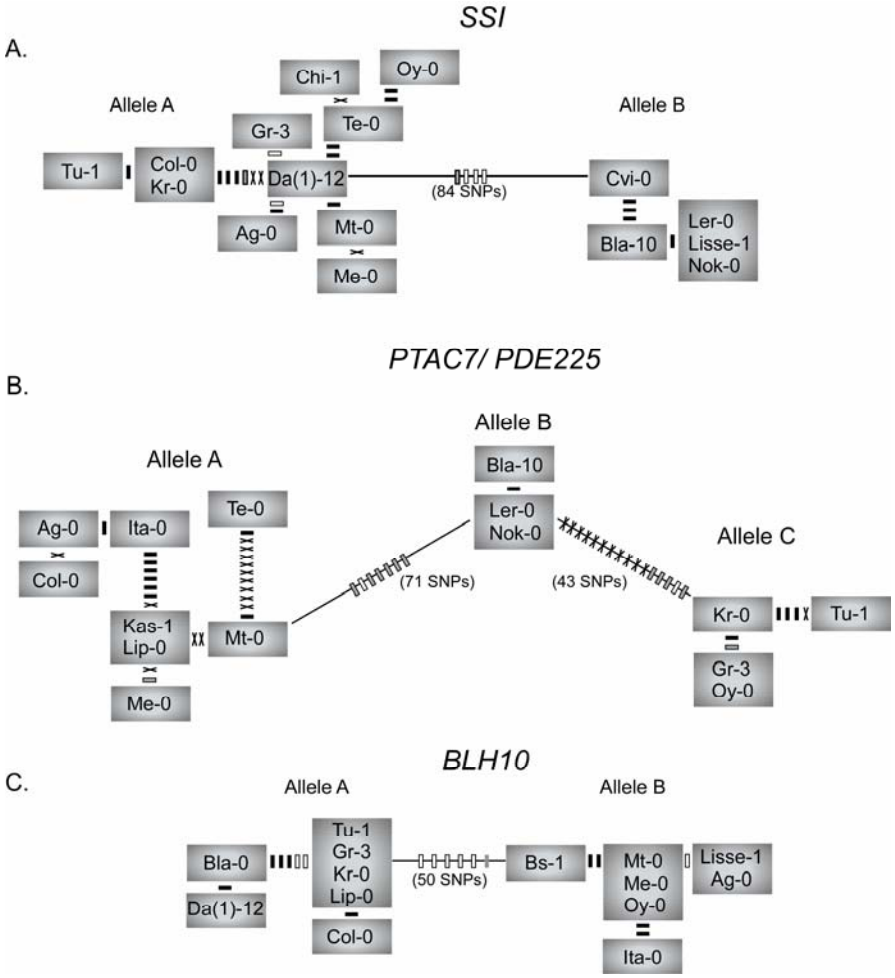


Figure 1. Haplotype networks

Haplotype networks showing the relationship between allele classes among sequenced *A. thaliana* accessions at *SSI, PTAC7* and *BLH10*. Bars in black mark noncoding SNPs, white bars for nonsynonymous changes, gray bars with no border for nonsynonymous indels, gray bars borders for tri-allelic SNPs, and X for homoplasies. Longer branches are marked by a solid line and all SNPs that are not silent or noncoding are indicated by their appropriate bar. The number of total allele-differentiating SNPs that contribute to these branches are also indicated.

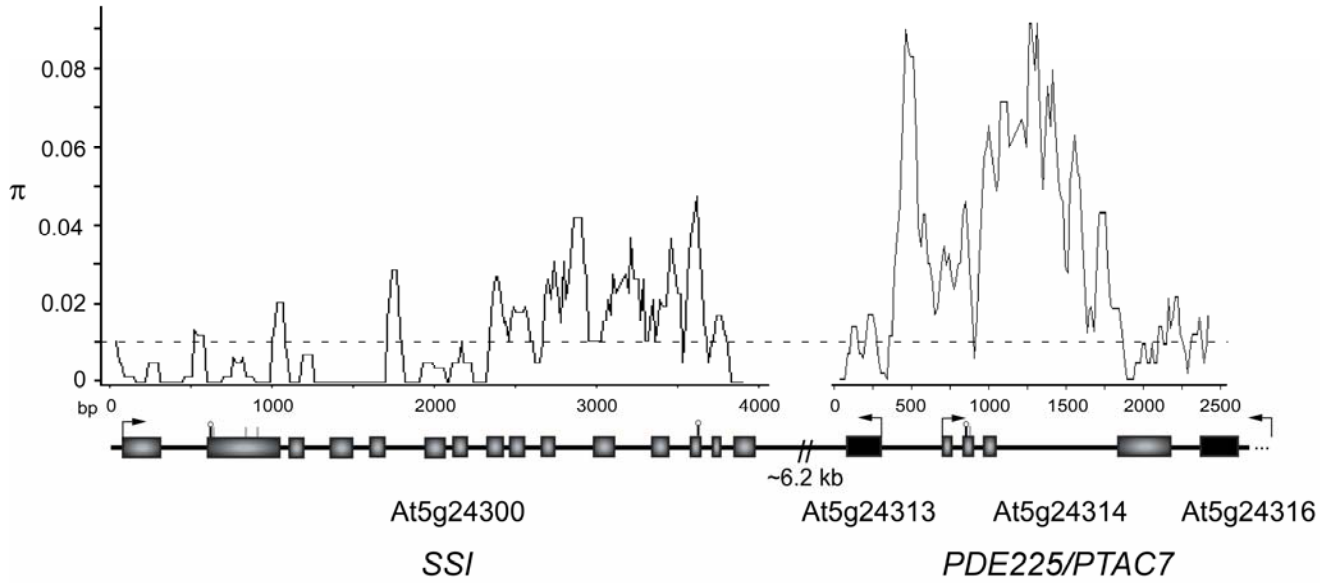


Figure 2. Sliding window analysis of *SSI* and *PTAC7*

Sliding window analysis of total nucleotide diversity (π) across *SSI* and *PTAC7* (*PDE225*). Focal genes are diagramed in dark grey, while flanking genes are in black. The dashed line represents average genome-wide nucleotide diversity in *A. thaliana.* All observed nonsynonymous variants at candidate loci are indicated as flags on the gene diagram as follows: grey flags represent singletons and knobbed black flags indicate allele-differentiating variants. Black flags indicate replacement polymorphisms that occur with a frequency > 10% that do not differentiate allele classes.

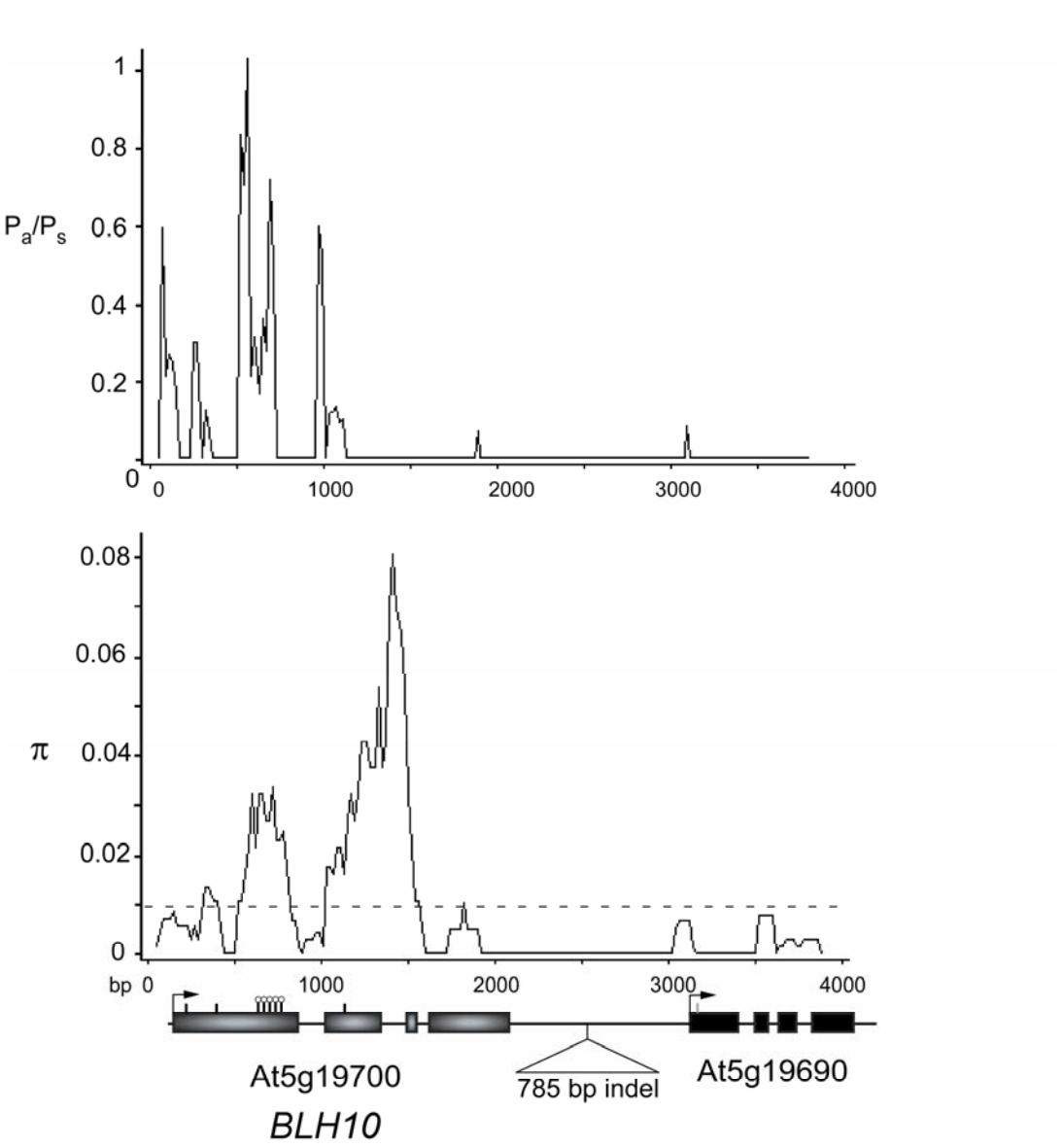


Figure 3. Sliding window analysis of *BLH10*

Sliding window analysis of total nucleotide diversity (π) and the proportion of replacement to synonymous nucleotide diversity, Pa/Ps, across the *BLH10* gene region. The dashed line represents average genome-wide nucleotide diversity in *A. thaliana*. Replacement polymorphisms are as described for Figure 1.

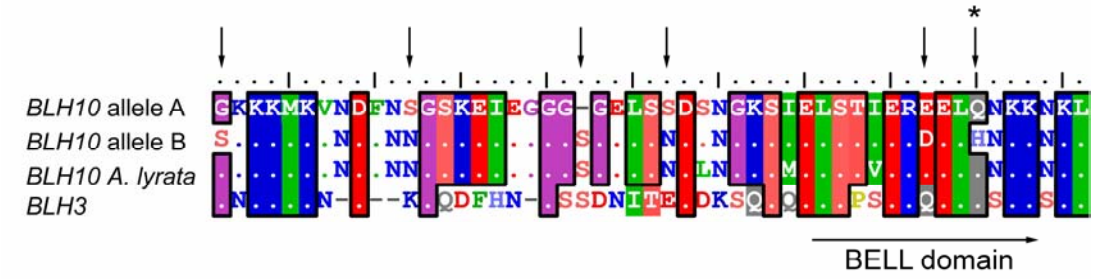
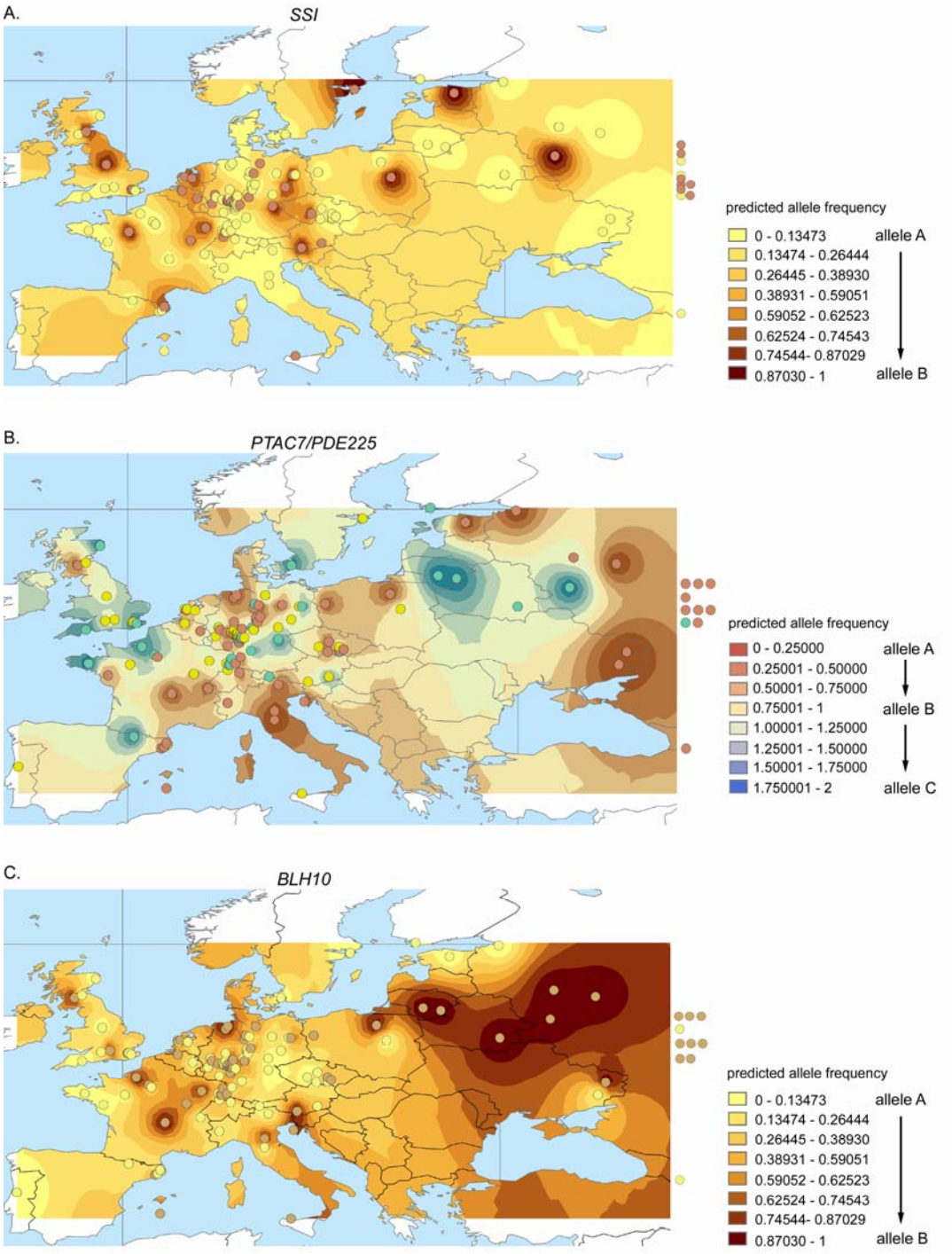


Figure 4. ***BLH10* BELL domain protein sequence alignment.**

Alignment of the region of the BLH10 protein showing the clustered replacement polymorphisms (arrows) observed within and just upstream of the BELL protein-interacting domain. Both A and B *BLH10* alleles are indicated and are aligned to the homologous *Arabidopsis lyrata BLH10* sequence and the *A. thaliana BLH3* paralog. The radical amino acid substitution predicted by PolyPhen is indicated by an asterisk.

Figure 5. Geographic allele frequency distributions

Inverse distance weighted map of observed and predicted allele frequencies for each candidate locus across *A. thaliana's* natural historic geographic range. Data points to the right of the map represent allele identities for ~30 accessions from eastern Siberia and Central Asia. Where populations are monomorphic only one data point describing the identity of accessions collected from that locale or population is depicted.


A.
SSI
predicted allele frequency
0 - 0.13473
0.13474 - 0.26444
0.26445 - 0.38930
0.38931 - 0.59051
0.59052 - 0.62523
0.62524 - 0.74543
0.74544- 0.87029
0.87030 - 1
allele A
allele B
B.
PTAC7/PDE225
predicted allele frequency
0 - 0.25000
0.25001 - 0.50000
0.50001 - 0.75000
0.75001 - 1
1.00001 - 1.25000
1.25001 - 1.50000
1.50001 - 1.75000
1.750001 - 2
allele A
allele B
allele C
C.
BLH10
predicted allele frequency
0 - 0.13473
0.13474 - 0.26444
0.26445 - 0.38930
0.38931 - 0.59051
0.59052 - 0.62523
0.62524 - 0.74543
0.74544- 0.87029
0.87030 - 1
allele A
allele B

Supplementary Table 1. Sequenced *A. thaliana* accessions.

Seed stock	Ecotype	Origin
CS1244	Ita-0	Ibel Tazekka - Morocco
CS1364	Me-0	Mechtshausen/Hessen - Germany
CS6090	Lisse-1[a]	Lisse, Keukenhof - Netherlands
CS6622	Bla-10	Blanes/Gerona - Spain
CS6764	Kr-0	Krefeld - Germany
CS6780	Lip-0	Lipowiec/Chrzanow - Poland
CS6799	Mt-0	Martuba/Cyrenaika - Libya
CS6807	Nok-0	Noordwijk - Netherlands
CS901	Ag-0	Argentat - France
CS902	Cvi-0	Cape Verdi Islands
CS903	Kas-1	Kashmir - India
CS20	Ler-0	Landsberg erecta
CS1202	Gr-3	Graz - Austria
CS6824	Oy-0	Oystese - Norway
CS6876	Tu-1	Turin - Italy
CS6918	Te-0	Tenela - Finland
CS917	Da(1)-12	Czechoslovakia
CS996	Bs-1	Switzerland - Basel
CS6665	Chi-1	Chisdra - Russia

[a] This designation does not represent an official ecotype name.

Supplementary Table 2. Genotypes of candidate genes.

Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele	Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele
910	B	B	B	6611	A	C	A
913	B	C	A	6612	B	C	A
915	B	A	A	6614	B	A	A
919	A	A	A	6615	B	A	A
920	A	A	A	6616	A	B	B
923	A	A	A	6618	A	A	B
924	A	A	A	6626	A	A	A
926	A	A	A	6631	A	A	A
927	B	A	A	6632	B	B	B
928	B	A	A	6633	B	A	A
932	A	C	A	6634	B	B	B
1316	A	C	A	6636	B	B	B
1602	B	C	A	6637	B	A	A
3109	A	C	A	6639	B	B	B
3180	A	B	A	6640	A	A	A
6044	A	A	A	6641	B	C	A
6047	A	C	A	6646	B	A	A
6048	A	C	A	6650	A	B	B
6058	A	C	A	6651	A	?	B
6060	A	C	A	6655	A	B	B
6062	A	C	A	6656	A	B	B
6065	A	B	B	6659	A	B	B
6073	A	C	A	6665	B	C	B
6076	A	B	A	6676	A	C	A
6078	A	C	A	6677	B	B	B
6087	B	B	A	6678	B	B	B
6090	B	B	B	6681	B	B	B
6093	B	B	B	6683	A	B	A
6095	A	C	A	6685	B	C	A
6097	A	A	A	6686	B	B	A
6098	A	A	A	6688	A	B	B
6100	B	A	A	6689	A	B	B
6102	B	C	A	6690	B	A	A
6104	A	A	A	6692	A	C	A
6181	A	A	A	6693	A	B	B
6600	A	C	A	6694	A	A	A
6602	A	B	A	6695	B	C	A
6603	A	A	A	6700	A	A	B
6604	B	B	B	6703	B	C	A
6605	A	A	B	6704	A	B	B
6606	A	C	A	6708	B	A	A
6608	A	C	B	6709	B	C	A
6609	B	C	A	6710	A	A	A
6610	B	B	B	6712	A	A	A

Supplementary Table 2 continued.

Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele	Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele
6713	B	A	A	6809	B	B	B
6718	A	A	A	6810	B	A	A
6720	B	B	B	6812	B	C	A
6722	A	C	A	6813	A	A	A
6725	A	C	A	6816	B	B	?
6727	B	C	A	6818	B	C	A
6728	A	B	B	6820	B	A	A
6730	A	B	B	6821	B	A	A
6731	A	C	A	6822	A	B	B
6732	B	C	A	6823	B	A	A
6733	B	A	A	6825	B	B	B
6736	A	B	B	6826	B	A	A
6738	A	A	A	6827	B	A	A
6739	A	A	A	6831	A	A	A
6741	A	A	A	6832	A	A	A
6742	A	B	A	6833	A	C	A
6744	B	A	A	6834	A	B	B
6747	A	A	A	6835	A	B	B
6748	A	A	A	6837	A	C	A
6752	B	B	B	6838	A	C	A
6753	B	B	B	6839	A	C	A
6756	A	B	B	6840	B	A	A
6757	A	C	A	6841	B	A	A
6760	B	C	A	6845	A	A	A
6761	B	C	A	6846	A	C	A
6762	B	C	A	6847	B	C	A
6765	B	A	A	6850	B	A	A
6766	A	B	B	6853	A	A	A
6767	A	A	A	6855	A	C	A
6768	B	A	A	6856	A	A	A
6770	A	A	A	6859	B	A	A
6772	A	A	A	6860	B	A	A
6775	B	C	A	6862	A	B	B
6777	A	B	B	6863	A	B	B
6778	A	A	B	6864	A	A	A
6781	A	A	B	6867	A	B	B
6782	A	A	B	6872	B	A	A
6783	A	A	A	6873	A	C	A
6784	A	B	B	6875	A	C	A
6785	A	B	B	6876	A	A	A
6788	B	A	A	6879	A	A	A
6789	B	C	A	6880	A	C	A
6790	B	A	A	6881	B	C	A
6793	B	A	A	6882	B	?	B
6797	B	A	A	6885	A	B	B

Supplementary Table 2 continued.

Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele	Accession	*BLH10* allele	*PTAC7* allele	*SSI* allele
6798	B	A	A	6886	A	?	B
6800	B	C	A	6887	A	A	A
6801	A	A	A	6888	B	C	A
6802	B	B	B	6889	B	C	A
6804	A	A	A	6892	B	C	A
6893	A	C	A	9628	B	A	B
6895	A	A	A	9629	B	A	A
6896	A	A	A	9630	B	C	A
6903	A	A	A	9631	B	C	A
6906	B	A	A	9632	B	A	B
6918	A	C	A	9633	B	A	B
6920	A	C	A	9634	B	C	A
6923	B	B	B	9635	B	A	B
6928	A	?	A	9636	B	C	A
6619	B	?	A	9637	?	A	B
6621	A	B	B	9638	B	C	A
6622	A	B	B	9639	B	A	B
6623	A	B	B	9640	B	C	A
6624	A	A	A	9641	B	A	B
6669	A	B	B	9642	B	A	B
6671	A	A	A	9643	B	A	B
6672	A	B	A	9644	B	A	B
6836	A	B	B	9645	B	A	B
6857	A	?	B	9648	B	A	B
6868	B	A	B	9649	B	A	B
6869	B	?	B	9650	B	A	B
6870	B	A	?	9651	B	A	B
6871	A	?	B	9652	B	A	B
6917	B	A	A	9653	A	A	A
9624	B	A	B	9654	A	A	A
9625	B	A	B	9655	A	A	A
9626	B	A	B	9656	A	A	A
9627	A	A	A				

Supplementary Text 1. Gene PCR and sequencing primers.

BLH10:

BLH1F[1]- CCAAACTCATCGAACTAAGAACTCA

BLH2F- CACCGTTTCATGTGGATGAG

BLH3F- AATGAGACATCAGGAGGGTTT

BLH4F- AAGAAGAACAAGCTTTTAACAATGG

BLH1R- TGCAACCATTAACACATCCA

BLH5F- CATTTGAGATAGTAGCAGGACTTGG

BLH6F[2]- AGAACCCAACAGCACAAACC

BLH7F- GTGTAGGTTGCTTTTGTGTATAGTAGA

BLH8F- CGGTGGTATAAAGGATACTTGCA

BLH2R[1]- GTTGCTGTGGGTCGTTATCA

BLH3R- CTTTCTTGCTGGTTGCTTCC

BLH4R- GCGCTTCAAGAATCATCTCC

BLH5R- TCGGAGGACTGAGAGACTCC

BLH6R- CCATCATTGAAGATCTGATTAAAAC

BLH7R[2]- CATGAACAAAAGAGATGAAACTTGG

SSI:

SSI_1F[1]- CGAAGAAACTCTCACTCTTTAATTG

SSI_2F- TTTTGCTTGTGGTCCTCAGA

SSI_3F- GGGCTAAGGATTTGGGGATA

SSI_1R- CATGCCAGTCATTGACAAGG

SSI_2R- ACTGCTCCATACCATTCCGA

SSI_3R- CACCGGAATATCAGCTCAGA

SSI_4F[2]- CAAGATGGCACTACAAAAGGAATT

SSI_5F- GCCTGAATGTCCTATGGTGA

SSI_4R[1]- CCGTTCAATAGACTGACTTACGAAT

SSI_5R- CCAACTCGGTTTGGTTCAGT

SSI_6R[2]- AATTCTATACCTCCCCAAAACAAAC

Supplementary Text 1 continued.

SSI_6F- AGTTGTTCATGGCACTGGAG

SSI_7R- AGTTCGGGTCATTCCTCTT

SSI_8R- AAATTCTCAACCGTATCCTGAAAA

PTAC7:

PTAC7_1F[1]- CCGATTCTCACCATCATCA

PTAC7_1R[1]- CTAAGACTTCCCAAGCCATC

PTAC7_2F- GGGTTCCCACTAACTATTATTAGATTG

PTAC7_2R- TCCAACAAATAGCCAGGACC

PTAC7_3F- TGTTCTTCTCCATCTTCGATTTTAC

PTAC7_4R- GTAACGCAACATATCGTCCTTCTT

PCR primer sets are denoted by [1] and [2].

Supplementary Text 2.

Pyrosequencing SNPs and Primers:

PTAC7: A/B allele comparison

Expected sequence: (**A/G**)TCAAAGCCGAGATCG…

Sequencing primer: 5- CTACATAAGCCTCAGGGCG -3’

PCR primers: PTAC7pF: 5’- /Bio/TGACGAAGAAGGACGATATG –3’

PTAC7pR: 5’- CTACATAAGCCTCAGGGCG-3’

PTAC7: A/C allele comparison

Expected sequence: (**T/C**)ACTTGTTCTTCCACAAA…

Sequencing primer: 5’- TTCCCAACTTCTCTTATCTTCCT -3’

PCR primers: PTAC7pF: 5’- /Bio/TGACGAAGAAGGACGATATG -3’

PTAC7pR: 5’- CTACATAAGCCTCAGGGCG -3’

SSI:

Expected sequence: ATTCCG(**T/C**)GGTTGGGTTGGG…

Sequencing primer: 5’- TGGAGGAAACATACAGAGACAA -3’

PCR primers: SSIpF: 5’- ATGCTTGGTTCGGGTGA -3’

SSIpR: 5’- /Bio/GGCTGTGATTCGATGAGAGA -3’

BLH10:

Expected sequence: (**G/T**)ATCATCTCG…

Sequencing primer: 5’- CAGATTGACCCACAA -3’

PCR primers: BLH10pF: 5' - /Bio/TCTCTGTTGCTCCGTTTCAC - 3'

BLH10pR: 5’- GCTGCTCATAAACCCTCCTG- 3

Supplementary Text 3. Genotyping PCR primers and restriction enzymes.

SOLUBLE STARCH SYNTHASE I

Primers:

SSIgF- CAGCCGGGTAAAAACAAATC

SSIgR- TTCTGCATAAGGGTTGAAATTCT

An indel with visible product size differences was genotyped. No was enzyme used.

PLASTID TRANSCRIPTIONALLY ACTIVE 7

Primers:

PTAC7gF- TGTTGAGGTTCCATCTTGTAATG

PTAC7gr- GGAACTCTTTGCATTTTGTAACG

Restriction enzyme: BstYI

Product description: PCR product is ~1485 bp. BstYI cuts allele A once (322 bp and 1163 bp), allele B twice (322 bp, 789 bp, and 374 bp), and does not cut allele C (1485 bp).

BELL-LIKE HOMEODOMAIN 10

Primers:

BLH10gF- CCTGCTCAGAATTTGCTTGA

BLH10gr- GCTTCCATTTGATGGTAATATTGG

Restriction enzyme: BfaI

Product description: PCR product is ~370 bp. BfaI cuts allele B once (253 bp and 117 bp) and does not cut allele A (370 bp).

CHAPTER FOUR

Nucleotide variation across a rice centromere

Nucleotide variation across a rice centromere

Jennifer M. Reininga[1], Philip Awadalla[1], Susan McCouch[2] and Michael D. Purugganan[3*]

[1]Department of Genetics, Box 7614
North Carolina State University, Raleigh, NC 27695

[2]Department of Plant Breeding and Genetics, Cornell University, 161 Emerson Hall, Ithaca, NY 14850, USA.

[3]Department of Biology and Center for Comparative Functional Genomics, New York University, 1009 Silver, 100 Washington Square East, New York, NY 10003

*CORRESPONDING AUTHOR: Michael Purugganan. Telephone (212) 992 9628 and email: mp132@nyu.edu

Prepared for submission to *Genome Research*.

ABSTRACT

Rice centromere 8 contains actively expressed genes and relatively few satellite repeats leading to the suggestion that it represents an intermediate stage of centromere evolution. We sequenced fragments from ~18 expressed single copy genes distributed across a ~3.3 Mb region encompassing rice centromere 8 in ~ 20 individuals each of three domesticated *O. sativa* varieties (*indica*, *temperate japonica*, and *tropical japonica*) as well as from the wild progenitor, *O. rufipogon* and compared population data to the outgroup *O. meridionalis* to assess the evolutionary forces shaping genetic variation in this region. Estimates of the historical population recombination rate for each rice group corroborate genetic map data indicating that this region is largely non-recombining among domesticated varieties. Our empirical findings contradict theoretical predictions and empirical data from animal models which predict that nucleotide diversity should be low in this non-recombining region due to hitchhiking with selected variants. We find that both *temperate japonica* and *tropical japonica* centromere 8 regions possess significantly elevated levels of nucleotide variation when compared to an empirical genome-wide distribution. While *indica* and *O. rufipogon* show decreased diversity, HKA tests do not support the hypothesis that natural selection is responsible. Demography and/or selective interference seem most likely to explain patterns of diversity and divergence at rice centromere 8.

INTRODUCTION

Centromeres are a defining feature of eukaryotic chromosomes, functioning to mediate sister chromatid cohesion and microtubule attachment during chromosomal segregation in both mitosis and meiosis. All higher eukaryotic centromeres consist of 100-200 bp satellite DNA repeats interspersed or flanked by a large number of transposable elements (Dawe 2005) that together span broad regions of a chromosome (>1Mb) (Ma and Bennetzen 2006). The functional centromere, or kinetochore, is associated with a centromere-specific histone variant (CENH3) and other centromere-specific proteins (Talbert et al. 2002). CENH3-affiliated DNA exists within heterochromatin, which has a high proportion of epigenetic marks, including histone modifications like dimethylated H3 lysine-9 (H3K9me2). Recent work in yeast, humans, and rice , however, suggests that kinetochores may also possess euchromatic or other distinctive methylation features (Cam et al. 2005; Sullivan and Karpen 2004; Wong et al. 2006; Yan and Jiang 2007). Centromeres are also known to have reduced levels of meiotic recombination, a characteristic that has been attributed to the repetitive nature of these regions and their associated heterochromatin, but may also be epigenetically regulated (Yan et al. 2005). Despite a lack of homologous recombination in the region, unequal crossing over is prevalent in centromere cores, providing a mechanism for rapid satellite expansion and contraction (Ma and Bennetzen 2006; Ma et al. 2007).

Centromeres present a paradox in that they are functionally conserved across higher eukaryotes, even though the repetitive DNA sequences that define them are highly variable even among closely related species (Hall et al. 2003; Henikoff et al. 2001; Jiang

et al. 2003; Malik and Henikoff 2002). Recent comparative evolutionary studies of pericentromeric sequences in rice and other plant species have highlighted these regions as highly dynamic in terms of size and structural divergence (Hall et al. 2006; Ma and Bennetzen 2006) but also have also demonstrated similar genomic features among homologous centromeres from different species (Hall et al. 2006).

While there have been numerous comparative studies of centromere evolution between species, there have been a relatively limited number of empirical studies on the microevolutionary processes acting at centromeres or pericentromeric regions. A few studies have focused on the rapid evolution and homogenization of centromere satellite repeats within species, including the model plant *A. thaliana* and its relatives (Hall et al. 2003; Kawabe and Charlesworth 2007; Kawabe and Nasuda 2006; Kawabe et al. 2006). There have also been several studies on the levels and patterns of nucleotide polymorphism at or near centromeres (Cabot et al. 1993; Hall et al. 2003; Nachman 1997; Stephan and Langley 1989; Stephan and Mitchell 1992; Wright et al. 2005; Wright et al. 2006). One focus of attention in these latter studies has been the low levels of recombination observed in centromeres, since theoretical predictions based on models of genetic hitchhiking (Charlesworth et al. 1993; Hudson and Kaplan 1995; Innan and Stephan 2003) leads to the prediction that these low recombination levels should result in decreased intraspecific diversity. Selective scenarios, including selective sweeps (Maynard Smith and Haigh 1974), background (Charlesworth et al. 1993) balancing selection (Charlesworth et al. 1997), and selective interference (Hill and Robertson 1966; Otto and Barton 1997) have greater impact in regions of low recombination due to hitchhiking of selected variants. Often, selective events may be population specific, and

manifest locally in the genome as regions of elevated divergence between selected and non-selected populations. Alternatively, species-wide demographic effects could also have pronounced effects on diversity in low recombining regions since genetic drift will play a stronger role in regions of the genome with a smaller effective population size (*Ne*) (Charlesworth et al. 1997).

In *Drosophila*, reduced levels of variation in low-recombining centromere and sub-telomeric regions and positive correlations between recombination and nucleotide diversity (but not divergence) have been attributed to recurrent selective sweeps and their associated hitchhiking effects (Begun and Aquadro 1992; Langley et al. 2000; Przeworski et al. 2001). Positive correlations between diversity and recombination rates and have also been observed in other animal models including mice (Nachman 1997; Takahashi et al. 2004) and humans (Nachman 2001). Data from plants, however, are more equivocal, and no strong relationship between recombination and nucleotide variation has been observed (Baudry et al. 2001; Bevan et al. 2001; Nordborg et al. 2005; Roselius et al. 2005; Tenaillon et al. 2002; Wright et al. 2005; Wright et al. 2006). In maize, the correlation between recombination and diversity seems to depend on the measure of recombination employed (Tenaillon et al. 2002), although a significant clustering of genes with the signature of positive selection have been observed near centromeres (Wright et al. 2005). Polymorphism data from the pericentromeres of six *A. lyrata* populations also show contrasting patterns, with only one population showing the decreased diversity consistent with theoretical models of selective hitchhiking (Wright et al. 2006). Contrary to data from *Drosophila* pericentromeres (Stephan and Langley

1989; Stephan and Mitchell 1992), data from plants do not seem to show any general trend for levels of diversity in centromeric regions.

A major limitation of previous studies on centromere variation has been the lack of detailed sequence information spanning eukaryotic centromeres. In this regard, domesticated rice (*Oryza sativa*) has emerged as a model for the study of centromere structure, function and evolution (Yan and Jiang 2007), largely as a consequence of the completion of full length centromere sequences for chromosomes 4, 5 and 8 (Matsumoto et al. 2005; Nagaki et al. 2004; Wu et al. 2004; Zhang et al. 2004). Sequence analysis reveals that rice centromeres are characterized by 155-bp Cen0 satellite repeats (extending over 60 kb to 2Mb) and centromere-specific CRR retrotransposons, either or both of which are required for associations with centromeric histone 3 (CENH3) (Nagaki et al. 2004).

The most noteworthy feature of these rice centromeres is the presence of expressed, functional genes that exist within the heterochromatic CENH3 binding domain (Nagaki et al. 2004; Wu et al. 2004; Zhang et al. 2004), which had previously been observed only in pericentomeres and human neocentromeres (Saffery et al. 2003). These genes appear to be primarily associated with islands of euchromatic epigenetic features, including enrichment in H4 acetylation and H3K4me2 (Yan and Jiang 2007), that are embedded within the largely heterochromatic centromere. The presence of actively expressed genes and the reduced occurrence of repetitive elements in rice centromere 8 have led to suggestions that it represents an intermediate stage in centromere evolution (Nagaki et al. 2004; Yan and Jiang 2007).

Here we report on a molecular population genetic study of expressed genes in rice centromere 8, including genes located within the functionally defined kinetochore. Patterns of polymorphism and divergence are compared among three major domesticated rice varieties (*indica*, *temperate* and *tropical japonica*) and the wild progenitor, *Oryza rufipogon* using the outgroup *O. meridonalis.*. We show contrasting patterns of variation between rice groups across the region and observe no evidence of hitchhiking with selected variants, despite the obvious role of selection in shaping genomic diversity in this domesticated species. Genetic drift and/or selective interference seem most likely to play a role in shaping the observed nucleotide diversity at genes in the rice centromere 8 region.

MATERIALS AND METHODS

Population Sampling. A collection of 96 rice accessions were included to represent a diverse sample of domesticated *Oryza. sativa* and the wild ancestor, *O. rufipogon*. The *O. sativa* sample comprises 18 tropical japonica, 21 temperate japonica, and 21 indica landraces (with the exception of five accessions, which are modern cultivars) collected from diverse Asian locales. Twenty-one *O. rufipogon* accessions were included and consist of 8 samples from both Nepal and China, and one each from Malaysia, Laos, India and Papua New Guinea. Also included in the collection is one sample from each of two outgroup species, *O. barthii* and *O. meridionalis*, and one *O. nivara* accession (see Olsen et al. 2006 for a detailed sample listing).

Seed was obtained from germplasm repositories (International Rice Germplasm Collection (IRGC), Philippines; National Small Grains Collection, Idaho; and from the Rice Genetic Resources Unit, Japan) with the exception of Nepalese *O. rufipogon* accessions, which were obtained from Hee Jong Koh (Seoul National University). DNA was isolated from single plants as described in McCouch et al. 1988 with slight modification. All *O. sativa* accessions were purified by self-fertilization for two generations prior to initiation of the study.

PCR and sequencing. Approximately 1 kb of nucleotide sequence data were obtained from 18 of 20 genes chosen to span a 3.3 Mb region including centromere 8. PCR primers were designed using Primer 3 software (http://frodo.wi.mit.edu/) and the Nipponbare (*temperate japonica*) genomic sequence and annotation (www.gramene.com) as a reference. Primers were anchored in exons and designed to generate ~500 bp PCR products including both coding and intronic DNA whenever possible. Primers were blasted against the Nipponbare standard to ensure that only the gene of interest would be amplified. See table S1 for a list of primer sequences. Two separate PCR and sequencing reactions were performed at each locus to obtain ~1 kb total sequence. In most cases the amplicons are overlapping, although this was not possible for several loci (Os08g19850, Os08g21590, Os08g21840, Os08g22160, Os08g23100, Os08g23150) due to undesirable GC content and repetitive/non-unique sequences. DNA sequencing was performed by Genaissance (New Haven, CT) as previously described (Olsen et al. 2006).

Phred and Phrap (Codon Code, Dedham, MA) were used in conjunction with Biolign Version 2.0.9 (Tom Hall, NC State University) for sequence editing and

alignment. Heterozygous sites were identified using Polyphred (Deborah Nickerson, University of Washington) and manually by visual inspection. As expected, heterozygotes were uncommon in *O. sativa* with the exception of the Indonesian *indica* accession POPOT-165 which was excluded from further analysis. Heterozygous *O. rufipogon* positions were assigned ambiguity codes.

Population genetic analysis. Intra- and inter- specific measures of diversity and divergence were estimated independently for each examined locus in the centromere 8 region. DNAsp 4.10 (Rozas and Rozas 1999) was used to calculate silent site nucleotide diversity as both π (Tajima 1983) and θ_W (Watterson 1975), Tajima's *D* (Tajima 1989), and the number of nonsynonymous (*Ka*) and silent (*Ks*) substitutions per site (Nei 1987) for *O. sativa* and *O. rufipogon* comparisons with *O. meridionalis*. All analyses were preformed separately in each population (*indica*, *temperate japonica*, and *tropical japonica*) and in each species as a whole (*O. sativa* and *O. rufipogon*). Tests for significant differences among the above mentioned parameters and statistics between the *Cen8* region and a random genomic sample (Caicedo et al., submitted) were conducted using non-parametric Mann-Whitney tests (web tool available at http://faculty.vassar.edu/lowry/utest.html). TASSEL 2.0 (http://www.maizegenetics.net/index.php?page=bioinformatics/tassel/index.html) was used to estimate and describe linkage disequilibrium between informative variants with frequencies >10% in each rice group across the Cen8 region, as well as to generate a minimum spanning tree which was used as a guide during manual construction of the presented haplotype network. HKA tests of selection (Hudson et al. 1987) were

conducted using the multiloucus application available from Jody Hey (Hey and Nielsen 2004) (http://lifesci.rutgers.edu/_heylab/HeylabSoftware.html). Since the rate of recombination across centromere 8 was not significantly different from zero we concatenated sequence fragments to generate a single centromere locus which was then compared to a collection of 5 random genomic STS loci (Caicedo et al., submitted).

LD was quantified as $r^2 = (P_{11}P_{22}–P_{12}P_{21})^2/(p_1(1–p_1)p_2(1–p_2))$ (Hill and Robertson 1966), where p_1 and p_2 are the allele frequencies at the first and second locus respectively, and P_{ij} are the haplotype frequencies at loci i and j. Estimates of r^2 and recombination were calculated using LDhat (McVean et al. 2002) software. Since pairwise LD estimates in a given dataset are not independent we used a permutation test to assess whether there was a significant decline of LD with physical distance (Charlesworth and Awadalla 1998). This test randomizes the positions of segregating sites, and for each randomization computes the correlation coefficient for physical distance between pairs of sites and LD. Estimates of $4Nr$ were calculated for SNP data only using the composite likelihood approach in LDhat. Estimates were made using models of single crossover and gene conversion (Awadalla and Charlesworth 1999; McVean et al. 2002). The relationship between physical distance and recombination rate with gene conversion is $R_{ij} = 2ct(1\text{-}e\text{-}^{\text{dij/t}})$, where c is the per base rate of initiation of gene conversion and t is the average gene conversion tract length (assuming an exponential distribution (Andolfatto and Nordborg 1998)). We fixed the average tract length and estimated the compound parameter $\gamma = 8N_ect$, i.e., the population rate of recombination between two distantly linked loci caused by gene conversion.

RESULTS AND DISCUSSION

Levels and patterns of diversity within and among rice varieties at rice centromere 8. Population sequence data from 18 loci scattered across a 3.3 Mb region encompassing rice centromere 8 (Table 1) was collected from three major domesticated rice varieties (*indica*, and *temperate* and *tropical japonicas*) and the wild progenitor species, *O. rufipogon*. The sampled region includes a previously defined 2.3 Mb non-recombining region (Yan et al. 2005) and represents the first polymorphism data for active genes physically located within the kinetochore of a centromere.

Levels of per site nucleotide diversity across centromere 8 vary widely across gene fragments, ranging from $\pi = 0.0000$-0.0047 for domesticated rice, and $\pi = 0.0003$-0.0075 for *O. rufipogon* (Fig. 1 and Table 2). Levels of silent nucleotide variation are much higher in the *tropical* and *temperate japonica* variety groups relative to *indica*, which differs from genome-wide estimates of silent diversity where *indica* is the most polymorphic (Caicedo et al., submitted) (Fig. 2).

Diversity is reduced by about half in the centromeric region of *O. rufipogon* and *O. sativa* ssp. *indica* compared to a genome-wide sample (Fig. 2). The distributions of silent site nucleotide polymorphism levels between genome-wide vs. centromere loci for these two groups are significant using a non-parametric Mann-Whitney test ($p < 0.031$ for *O. rufipogon* and $p < 0.023$ for *indica*). Surprisingly, and in contrast, both *temperate* and *tropical japonica* display an increase in diversity at the centromere relative to subspecies genome averages; these differences are also significant (Mann-Whitney tests, $p<0.0001$for *temperate japonica* and $p<0.009$ for *tropical japonica*). The significant

differences in the distributions reflect both a change in median nucleotide levels between groups and the variances in the centromere 8 regions (Fig. 2).

Levels of silent nucleotide polymorphism also vary across the centromere 8 region. Interestingly, levels of nucleotide variation are quite high at several genes sampled in the kinetochore, with average kinetochore diversity around 0.0020 in domesticated rice and 0.0024 in the wild ancestor, *O. rufipogon* (Fig. 1 and Table 2). Additionally, we observed a local decrease in nucleotide diversity ($\pi_{(s)}$) between positions 12.5 and 13.3 Mb along the chromosome, with the valley of reduced variation including and just upstream of the primary 38.3 kb Cent0 repeat array (Nagaki et al. 2004)(Fig 1). This pattern, however, is not as marked when $\theta_{(s)}$ is considered (Fig. 1).

Tajima's *D* for genes in a rice centromere. We calculated Tajima's *D* (Tajima 1989), which is sensitive to the skew in allele frequencies relative to those observed under a neutral-equilibrium model, at loci of the centromere 8 region and in comparison to genome-wide estimates (Caicedo et al., submitted). Negative Tajima's *D* values indicate the presence of low frequency variants in the sample, while positive values suggest intermediate frequency polymorphism. Centromere 8 Tajima's *D* values for all but one group are largely negative, with the exception of *tropical japonica*, which has Tajima's *D* values that tend toward positive values (Fig. 2 and Table 2). We tested if the distribution of Tajima's *D* between centromere and genome-wide loci are significantly different, and only two significant results were observed. The centromere 8 *tropical japonica* sample has significantly higher Tajima's *D* compared to the genomic sample (Caicedo et al., submitted) ($p<0.0047$), while *indica* values were significantly lower ($p<0.0006$) (Fig. 2).

Recombination and linkage disequilibrium. Several studies have suggested that patterns of variation in pericentromeric regions are different from euchromatic genomic regions, a difference that has been primarily attributed to the reduced frequency of crossing-over events near centromeres. To investigate the contribution of recombination in shaping patterns of variation in the centromere 8 region we inferred population recombination rates from single nucleotide polymorphisms for centromeric regions of the largely selfing domesticated *O. sativa* variety groups and compared these with the outcrossing progenitor species *O. rufipogon* (Fig. 3).

Genetic maps of *O. sativa* suggest that crossing over is reduced in pericentromeric regions on chromosome 8, and was the basis for defining a 2.3 Mb non-recombining region associated with the centromere (Harushima et al. 1998). Composite likelihood-based inferences of the historical population recombination rates ($\rho = 4Nr$), confirm this observation and appear to be greatly reduced in rice centromere 8 of all domesticated subspecies and the outcrossing progenitor. Across a 3.3 Mb region, estimates of population recombination rates were not significantly different from zero for *O. sativa* subspecies (Fig. 3), and sites are in extremely high LD (Fig. S1). Similarly, LD does not significantly decay with distance ($p>0.2$, see methods). For the same region in *O. rufipogon*, however, the inferred rate was low ($\rho \sim 2$ -5), yet significantly different from zero (Fig. 3), perhaps arising from the larger effective population size in this outcrossing species. LD decays significantly with distance ($p<0.02$) in *O. rufipogon*. Similar patterns were observed if a model of gene conversion was implemented rather than 4N*r*. Altogether, these observations suggest that recombination has occurred in the history of

this region and likewise in the ancestor of this group of species. However, the signature of recombination has been lost over time in the domesticated varieties, possibly as a consequence of both reduced crossing over due to proximity to a centromere and genetic drift associated with a reduction in the effective population size.

Relationships between haplotypes at centromere 8. To examine the evolutionary relationhips among centromere haplotypes, we constructed a minimum spanning haplotype network of rice centromere 8 that includes the three main domesticated rice variety groups as well as the wild progenitor, *O. rufipogon* (Fig. 4). Core haplotypes were identified using all informative SNP's identified among sampled loci across the *Cen8* region. The three core haplotypes represented primarily by *temperate japonica, tropical japonica* and *indica* are defined as haplogroups A, B and C, respectively. In general, these haplotype groups distinguish the three domesticated variety groups, although more extrensive haplotype sharing is observed between *temperate* and *tropical japonicas*. *O. rufipogon* haplotypes are distributed throughout the network and connect the *O. sativa* core haplotype groups, a pattern that is consistent with *O. rufipogon* being ancestral to domesticated rice.

Approximately 83 percent of sampled *temperate japonicas* have rice centromere 8 haplotypes that belong to haplogroup A, while approximately 6 percent belong to both haplogroups B and C. One unique and highly divergent *temperate japonica* haplotype was also observed. Sixty-five percent of *tropical japonica* centromere haplotypes are haplogroup B variants with shared *tropical japonica* haplotypes confined to haplogroup A and a smaller uniquely *tropical japonica* haplotype cluster branching off the core A

grouping. The observation that *temperate japonica* centromere 8 haplotypes belong to additional network haplogroups and are more diverse than those found in *tropical japonica* is somewhat surprising since the former is known be the result of a secondary domestication from *tropical japonica* (Caicedo et al., submitted). While it is possible that this may simply reflect chance sampling, it may also suggest that ancestral haplotypes absent in the *tropical* group still persist in the *temperate japonicas*. *Indica* haplotypes show a less diverse haplotype composition, with most individuals belonging to haplogroup C (82%) although several distinct haplotypes are also observed (Table 3).

Evidence for selection at rice centromere 8. The relationship between polymorphism and divergence at rice centromere 8 was investigated using multilocus HKA tests (Hey and Nielsen 2004; Hudson et al. 1987). No significant deviations were observed among genes of the centromere (*indica*: $X^2 = 16.30$, $P < 0.295$; *temperate japonica*: $X^2 = 17.33$, $P < 0.239$; *tropical japonica*: $X^2 = 16.28$, $P < 0.297$; *O. rufipogon*: $X^2 = 12.33$, $P < 0.580$), or due to centromeric deviations when centromere genes were included in a comparison with 5 random euchromatic genes. Significant HKA test results were obtained in the latter comparison for the three domesticated varieties, however, due to deviations occurring among the random euchromatic loci used for comparison (*indica*: $X^2 = 24.07$, $P <0.0002$; *temperate japonica*: $X^2 = 13.05$, $P <0.023$; *tropical japonica*: $X^2 = 20.15$, $P < 0.001$; *O. rufipogon*: $X^2 = 9.45$, $P < 0.092$). Contrary to theory and empirical findings in *Drosophila* and maize there is no evidence for reduced variation due to hitchhiking with selected variants at this rice centromere.

Divergence and mutation rates at centromere 8. Population genetic theory suggests that non-recombining regions should be less successful at purging slightly deleterious mutations, than regions with higher recombination rates (Nordborg et al. 1996), resulting in the accumulation of more fixed nonsynonymous mutations in such regions. Furthermore, prior research has suggested either a more recent common ancestor or a lower mutation rate for the centromere relative to the rest of the genome (Ma and Bennetzen 2006) due to an observed regional deficiency in silent site divergence between two domesticated varieties.

To investigate patterns of divergence at *Cen8,* synonymous and nonsynonomous divergence between the outgroup, *O. meridonalis*, and both *O. sativa* and *O. rufipogon* sequences were compared with a genomic sample (Caicedo et al., submitted). Consistent with theory, average *Ka* estimates at the centromere are indeed slightly higher than the genomic average (Caicedo et al., submitted) in both the *O. sativa* and *O. rufipogon* comparisons, although this effect is not significant (Mann-Whitney one-tailed $p<0.1146$). We did not observe evidence, however, for decreased synonymous divergence at this centromere along the longer evolutionary timeline (~2 mya) investigated here, suggesting that a decreased mutation rate is unlikely to explain observed levels and patterns of variation in this genomic region.

Evidence of gene degradation and loss. It has been previously suggested that rice centromere 8 is at an intermediate stage of evolution, presently in transition from a neocentromere to a fully mature centromere. This arises from observations of expressed genes in the kinetochore region and a deficiency in the number of repetitive elements in

rice centromere 8 relative to other studied centromeres (Nagaki et al. 2004). Active genes are believed to be lost from centromeres as the number of centromere-specific repeats in the region increase, accompanied by meiotic drive from the strong selective advantage of more efficient centromeres during chromatid separation (Henikoff et al. 2001; Ma et al. 2007; Malik and Henikoff 2002) . The shift from euchromatic to heterochromatic genomic sequences, with the epigenetic modifications that accompany this change, might also result in the degeneration of genes in or near centromeres (Hall et al. 2006).

To test whether there is evidence of gene loss or degradation, we calculated the proportion of fixed nonsynonomous to synonomous differences (*Ka*/*Ks*) between *O. sativa* and the outgroup *O. meridionalis* for 15 expressed genes in rice centromere 8. Purifying selection will reduce *Ka*/*Ks* values less than one, *Ka*/*Ks* ~ 1 suggests neutral evolution of protein sequence arising from a reduction in functional constraint, while a value greater than 1 may indicate diversifying selection. Thirteen of the fifteen sampled genes in the *Cen8* region appear to be subject to purifying selection (*Ka*/*Ks*<1), indicating that selection acts to maintain gene function at the majority of loci across this region.

Coding region insertion/deletion (indel) polymorphisms can also be indicative of a loss in gene functionality, particularly when they result in frameshift and/or premature stop mutations that are often trademarks of pseudogene formation. Our results do suggest that several genes contain indel mutations that may compromise or abolish gene function. Four coding region indels, ranging from 13 to 204 bps in length, were observed at low to intermediate frequency in four rice centromere 8 loci. Two adjacent genes, designated Os08g23120 and Os08g23150, have a 13 bp and 108 bp deletion, respectively; the former results in a frameshift mutation. Both of these mutations, however, are singletons found

in only one individual in the sample (in *indica* and *O. rufipogon* groups, respectively). Another low-frequency coding region deletion (37 bps in length) leads to a frameshift mutation in Os08g22160, located in the kinetochore region. This variant was found in two individuals, one from the *tropical japonica* group and the other from *O. nivara*, a wild rice species that may be a subspecies of *O. rufipogon*.

The only indel found at intermediate (>10%) frequency is a 204-bp coding region indel in Os08g22780, located just outside the kinetochore. This indel is fixed in indica, and is at 90 percent frequency in *temperate japonica*, 33 percent in *tropical japonica*, and 59 percent in *O. rufipogon*. This variant represents an insertion relative to the standard Nipponbare sequence and results in a frameshift mutation that leads to a premature stop codon. The whole genome rice sequence was searched to see if other copies of this inserted 204-bp fragment were present elsewhere. The most similar sequence was identified in a hypothetical gene on chromosome 3, which shares 83 percent identity with the insertion and comparable similarity over other portions of the gene as well.

The microevolution of a plant centromere. Centromeres have for some time been an evolutionary enigma (Henikoff et al. 2001). They are functionally conserved across species, yet some of their sequences have been known to evolve quite rapidly, leading to suggestions that the diversification in their sequences and associated centromere binding proteins may underlie reproductive incompatibilities that accompany speciation (Hall et al. 2005; Henikoff et al. 2001; Stump et al. 2005). Past investigations on intraspecific variation in centromere-associated sequences have been focused on readily identifiable satellite repeat sequences found within centromeres (Cabot et al. 1993; Hall et al. 2003;

Waye and Willard 1987). Studies on these centromeric satellite repeats reveal substantial polymorphism; for example, intraspecific variation in a 178-bp centromere satellite repeat has been demonstrated in *A. thaliana* (Hall et al. 2003) and has led to suggestions that these may cause differential ability to be recognized by repeat binding proteins.

Population genetic studies in *Drosophila* species and other animal models have documented low levels of variation in genomic regions associated with centromere sequences (Begun and Aquadro 1992; Langley et al. 2000; Nachman 1997; Nachman 2001; Przeworski et al. 2001; Takahashi et al. 2004) These have validated population genetic theories that link reduced recombination and with low levels of nucleotide polymorphisms (Nordborg et al. 1996), which is thought to arise from genetic hitchhiking of neutral polymorphisms to genes subject to positive or background selection. The link between recombination and polymorphism, however, has been shown to be weak in plants (Baudry et al. 2001; Bevan et al. 2001; Nordborg et al. 2005; Roselius et al. 2005; Tenaillon et al. 2002; Wright et al. 2005; Wright et al. 2006). Pericentromeric sequences surveyed across multiple populations of *Arabidopsis lyrata* (Wright et al. 2006) and various species of tomato (Roselius et al. 2005), for example, do not show the consistently low levels of variation observed in other species.

There has been substantial interest in determining the levels and patterns of variation across centromeres, although the lack of complete sequence information of centromeres has precluded an analysis of the molecular population genetics of these key genomic regions and have restricted studies to either the satellite repeats or to pericentromeric genes. The availability of complete genomic sequences of several rice centromeres, including one for chromosome 8 (Nagaki et al. 2004), for the first time

permit us to probe the levels and patterns of nucleotide variation across centromere regions in functional genes outside the satellite repeats.

Our analysis of variation in these sequences in *O, sativa* and its wild relative *O. rufipogon* have confirmed that these centromeres have low historical rates of recombination, with the selfing cultivated rice having ρ equivalent to 0, while those of the outcrossing *O, rufipogon* have low but non-zero recombination levels. This is correlated with genetic map information (Harushima et al. 1998), that indicates reduced crossing over at the centromere region of chromosome 8. Given the low recombination rates at this centromere region, there is thus the expectation of lower variation in centromere sequences. This prediction is borne out half of the time; for the rice variety groups *indica* and and the wild rice *O, rufipogon*, levels of nucleotide polymorphism among gene fragments in the centromere are on average 2- to 4-fold lower than that observed in the genome-wide average of variation. While nucleotide variation levels are lower in these groups, the reduction is not as dramatic as observed in centromeres studied in other species. Moreover, the *O. sativa* variety groups *temperate* and *tropical japonica* actually harbor a higher mean level of nucleotide variation, ~2-fold higher than the genome-wide average, in direct contradiction to both theoretical expectations and previous work in other species.

There are several possible explanations for these observations. First, rice centromere 8 appears to be at a relatively early stage of centromere formation (Nagaki et al. 2004), which may consequently affect its evolutionary dynamics. The important genetic feature relevant to theoretical predictions, however, is the level of recombination, and both genetic map data (Harushima et al. 1998) and our results on historical

recombination indicate a low level of recombination, which should lead to lower polymorphism levels if hitchhiking accompanying either positive or background selection occurs.

The second possibility is that there has been no recent selection (either positive or background selection) on genes near this centromere. It is unclear, however, why this would be the case, especially since centromere 8 contains actively expressed single copy genes that could serve as selective targets. The presence of at least some mutations that could negatively impact protein function, for example, indicates that deleterious mutations can occur in this genomic region that may be subject to background selection. Nevertheless, it remains possible that selection has not occurred recently in this region in both cultivated and wild rice and this could be explained by selective interference due to the linkage of many functional genes in a common non-recombining region (Hill and Robertson 1966; Otto and Barton 1997). Indeed, the existence of different haplotypes indicates that positive selection, including centromere drive, may be rare in rice centromere 8 and has not yet led to the fixation of a single centromere type.

The third possibility is that the dynamics of selection at low-recombining regions may differ in rice, possibly because of greater population structure as well as departures from equilibrium due to bottlenecks and subsequent expansion that accompanies crop domestication. Gene flow between variety groups (and between cultivated and wild rice) could account for the polymorphisms in centromere 8 regions, and indeed our analysis indicates that haplotypes can be found in multiple groups (see Fig 4). Domestication may also lead to a reduction in the efficiency of purifying selection, both through the permissiveness of agricultural environments and the rapid expansion of crop populations

since the Neolithic 10,000-12,000 years ago. *Temperate japonica*, which has been subject to a very recent bottleneck as a consequence of its secondary derivation from *tropical japonica*, also has levels of nucleotide variation in centromere 8 that are higher than the genome-wide average for this group.

The precise interplay between recombination, selection and demography in shaping variation in these and other eukaryotic centromeres requires further consideration. What appears to be emerging is that in many plant species, unlike in *Drosophila* and other species, the correlation between recombination and polymorphism may not be as strong, and this results in centromeres that show distinct sequences within populations or species. It would be interesting to determine if functional differentiation is associated with these various centromere haplotypes; such differentiation may place a role in incipient speciation between individuals and/or populations by creating cytogenetic incompatibilities. It would also suggest that the macroevolutionary dynamics of centromere structure and function may differ between plants and animals, and further comparative analysis within and between species may shed light on the mechanisms by which centromeres across eukaryotic systems evolve.

ACKNOWLEDGEMENTS

We would like to thank Ana Caicedo for helpful comments and for providing genome-wide data for comparisons. This work was funded in part by the NSF Plant Genome Research Program and by a NIH pre-doctoral training grant to JMR.

REFERENCES

Andolfatto, P. and M. Nordborg. 1998. The effect of gene conversion on intralocus associations. *Genetics* **148:** 1397-1399.

Awadalla, P. and D. Charlesworth. 1999. Recombination and selection at Brassica self-incompatibility loci. *Genetics* **152:** 413-425.

Baudry, E., C. Kerdelhue, H. Innan, and W. Stephan. 2001. Species and recombination effects on DNA variability in the tomato genus. *Genetics* **158:** 1725-1735.

Begun, D.J. and C.F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356:** 519-520.

Bevan, M., K. Mayer, O. White, J.A. Eisen, D. Preuss, T. Bureau, S.L. Salzberg, and H.W. Mewes. 2001. Sequence and analysis of the Arabidopsis genome. *Curr Opin Plant Biol* **4:** 105-110.

Cabot, E.L., P. Doshi, M.L. Wu, and C.I. Wu. 1993. Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the Responder locus of Drosophila melanogaster. *Genetics* **135:** 477-487.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Fledel-Alon, A., York, T.L., Polato, N.R., Olsen, K.M., Nielsen, R., McCouch, S.R., Bustamante, C.D., and Purugganan, M.D. Genome-wide patterns of nucleotide polymorphism in domesticated rice. Submitted to PLOS Genetics.

Cam, H.P., T. Sugiyama, E.S. Chen, X. Chen, P.C. FitzGerald, and S.I. Grewal. 2005. Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nat Genet* **37:** 809-819.

Charlesworth, B., M.T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289-1303.

Charlesworth, B., M. Nordborg, and D. Charlesworth. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res* **70:** 155-174.

Charlesworth, D. and P. Awadalla. 1998. Flowering plant self-incompatibility: the molecular population genetics of Brassica S-loci. *Heredity* **81 (Pt 1):** 1-9.

Dawe, R.K. 2005. Centromere renewal and replacement in the plant kingdom. *Proc Natl Acad Sci U S A* **102:** 11573-11574.

Hall, A.E., G.C. Kettler, and D. Preuss. 2006. Dynamic evolution at pericentromeres. *Genome Res* **16:** 355-364.

Hall, S.E., G. Kettler, and D. Preuss. 2003. Centromere satellites from Arabidopsis populations: maintenance of conserved and variable domains. *Genome Res* **13:** 195-205.

Hall, S.E., S. Luo, A.E. Hall, and D. Preuss. 2005. Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. *Genetics* **170:** 1913-1927.

Harushima, Y., M. Yano, A. Shomura, M. Sato, T. Shimano, Y. Kuboki, T. Yamamoto, S.Y. Lin, B.A. Antonio, A. Parco, H. Kajiya, N. Huang, K. Yamamoto, Y. Nagamura, N. Kurata, G.S. Khush, and T. Sasaki. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* **148:** 479-494.

Henikoff, S., K. Ahmad, and H.S. Malik. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293:** 1098-1102.

Hey, J. and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics* **167:** 747-760.

Hill, W.G. and A. Robertson. 1966. The effect of linkage on the limits to natural selection. *Genetical Research* **8:** 269-294.

Hudson, R.R. and N.L. Kaplan. 1995. Deleterious background selection with recombination. *Genetics* **141:** 1605-1617.

Hudson, R.R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153-159.

Innan, H. and W. Stephan. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* **165:** 2307-2312.

Jiang, J., J.A. Birchler, W.A. Parrott, and R.K. Dawe. 2003. A molecular view of plant centromeres. *Trends Plant Sci* **8:** 570-575.

Kawabe, A. and D. Charlesworth. 2007. Patterns of DNA variation among three centromere satellite families in Arabidopsis halleri and A. lyrata. *J Mol Evol* **64:** 237-247.

Kawabe, A. and S. Nasuda. 2006. Polymorphic chromosomal specificity of centromere satellite families in Arabidopsis halleri ssp. gemmifera. *Genetica* **126:** 335-342.

Kawabe, A., S. Nasuda, and D. Charlesworth. 2006. Duplication of centromeric histone H3 (HTR12) gene in Arabidopsis halleri and A. lyrata, plant species with multiple centromeric satellite sequences. *Genetics* **174:** 2021-2032.

Langley, C.H., B.P. Lazzaro, W. Phillips, E. Heikkinen, and J.M. Braverman. 2000. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. *Genetics* **156:** 1837-1852.

Ma, J. and J.L. Bennetzen. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci U S A* **103:** 383-388.

Ma, J., R.A. Wing, J.L. Bennetzen, and S.A. Jackson. 2007. Plant centromere organization: a dynamic structure with conserved functions. *Trends Genet* **23:** 134-139.

Malik, H.S. and S. Henikoff. 2002. Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev* **12:** 711-718.

Matsumoto, T., J.Z. Wu, H. Kanamori, and e. al. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793-800.

Maynard Smith, J. and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. *Genetical Research* **23:** 23-35.

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160:** 1231-1241.

Nachman, M.W. 1997. Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics* **147:** 1303-1316.

Nachman, M.W. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **17:** 481-485.

Nagaki, K., Z. Cheng, S. Ouyang, P.B. Talbert, M. Kim, K.M. Jones, S. Henikoff, C.R. Buell, and J. Jiang. 2004. Sequencing of a rice centromere uncovers active genes. *Nat Genet* **36:** 138-145.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia Unversity Press, New York.

Nordborg, M., B. Charlesworth, and D. Charlesworth. 1996. The effect of recombination on background selection. *Genet Res* **67:** 159-174.

Nordborg, M., T.T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N.A. Rosenberg, C. Shah, J.D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. 2005. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol* **3:** e196.

Olsen, K.M., A.L. Caicedo, N. Polato, A. McClung, S. McCouch, and M.D. Purugganan. 2006. Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* **173:** 975-983.

Otto, S.P. and N.H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. *Genetics* **147:** 879-906.

Przeworski, M., J.D. Wall, and P. Andolfatto. 2001. Recombination and the frequency spectrum in Drosophila melanogaster and Drosophila simulans. *Mol Biol Evol* **18:** 291-298.

Roselius, K., W. Stephan, and T. Stadler. 2005. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171:** 753-763.

Rozas, J. and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174-175.

Saffery, R., H. Sumer, S. Hassan, L.H. Wong, J.M. Craig, K. Todokoro, M. Anderson, A. Stafford, and K.H. Choo. 2003. Transcription within a functional human centromere. *Mol Cell* **12:** 509-516.

Stephan, W. and C.H. Langley. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. *Genetics* **121:** 89-99.

Stephan, W. and S.J. Mitchell. 1992. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of Drosophila ananassae. *Genetics* **132:** 1039-1045.

Stump, A.D., M.C. Fitzpatrick, N.F. Lobo, S. Traore, N. Sagnon, C. Costantini, F.H. Collins, and N.J. Besansky. 2005. Centromere-proximal differentiation and speciation in Anopheles gambiae. *Proc Natl Acad Sci U S A* **102:** 15930-15935.

Sullivan, B.A. and G.H. Karpen. 2004. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat Struct Mol Biol* **11:** 1076-1083.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:** 437-460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-595.

Takahashi, A., Y.H. Liu, and N. Saitou. 2004. Genetic variation versus recombination rate in a structured population of mice. *Mol Biol Evol* **21:** 404-409.

Talbert, P.B., R. Masuelli, A.P. Tyagi, L. Comai, and S. Henikoff. 2002. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell* **14:** 1053-1066.

Tenaillon, M.I., M.C. Sawkins, L.K. Anderson, S.M. Stack, J. Doebley, and B.S. Gaut. 2002. Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). *Genetics* **162:** 1401-1413.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7:** 256-276.

Waye, J.S. and H.F. Willard. 1987. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* **15:** 7549-7569.

Wong, N.C., L.H. Wong, J.M. Quach, P. Canham, J.M. Craig, J.Z. Song, S.J. Clark, and K.H. Choo. 2006. Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet* **2:** e17.

Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, and B.S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* **308:** 1310-1314.

Wright, S.I., J.P. Foxe, L. DeRose-Wilson, A. Kawabe, M. Looseley, B.S. Gaut, and D. Charlesworth. 2006. Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. *Genetics* **174:** 1421-1430.

Wu, J., H. Yamagata, M. Hayashi-Tsugane, S. Hijishita, M. Fujisawa, M. Shibata, Y. Ito, M. Nakamura, M. Sakaguchi, R. Yoshihara, H. Kobayashi, K. Ito, W. Karasawa, M. Yamamoto, S. Saji, S. Katagiri, H. Kanamori, N. Namiki, Y. Katayose, T. Matsumoto, and T. Sasaki. 2004. Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* **16:** 967-976.

Yan, H. and J. Jiang. 2007. Rice as a model for centromere and heterochromatin research. *Chromosome Res* **15:** 77-84.

Yan, H., W. Jin, K. Nagaki, S. Tian, S. Ouyang, C.R. Buell, P.B. Talbert, S. Henikoff, and J. Jiang. 2005. Transcription and histone modifications in the recombination-free region spanning a rice centromere. *Plant Cell* **17:** 3227-3238.

Zhang, Y., Y. Huang, L. Zhang, Y. Li, T. Lu, Y. Lu, Q. Feng, Q. Zhao, Z. Cheng, Y. Xue, R.A. Wing, and B. Han. 2004. Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res* **32:** 2023-2030.

Table 1. Sampled genes in the centromere 8 region.

Gene ID	Function	Chromosomal position	Evidence for expression [c]
Os08g19320	Similar to At5g11030	11551085 - 11557349 bp	EST: CB636198, CI087475 , CB651023
Os08g19850	Putative threonyl-tRNA synthetase	11881741 - 11890954 bp	EST: CI218827, CF323173, CB682441
Os08g20200	Fatty acyl CoA reductase	12109374 - 12114857 bp	EST: CI711007, CI096803
Os08g20530, 6828.t00009[a]	Putative Ca2+ dependant lipid binding	12310949 - 12315180 bp	Gene not currently annotated in Gramene; RT-PCR positive [b]; EST: C28110
Os08g20660	Putative sucrose-phosphate synthase	12405930 - 12415165 bp	EST: CI002385, CF309892, CB667221, CB638627, CB678310
Os08g21330	formamidopyrimidine-DNA glycosylase family protein, expressed	12705811 - 12714086 bp	EST: CI071890, CI258417
Os08g21350	Putative cytokinin inducible	12719187 - 12727482 bp	EST: CI084744, AU225342, AU173140, CI613016
Os08g21590	Phosphatidylinositol 3-kinase, root isoform, putative, expressed	12849330 - 12866265 bp	EST: CI458537, CI272853, CB662215, CI695317
Os08g21840	Putative ribosomal protein L15	13032191 - 13035551 bp	EST: CI218141, CI766789
Os08g22160, 6730.t00011[a]	Putative CBS domain containing	13231274 - 13234987 bp	Gene not currently annotated in Gramene; RT-PCR positive [b];
Os08g22350, 6729.t00009[a]	Putative poly A binding	13339368 - 13342408 bp	Gene not currently annotated in Gramene; RT-PCR positive [b]; EST: AU075435
Os08g22360, 6729.t00010[a]	Putative poly A binding	13344848 - 13346370 bp	Gene not currently annotated in Gramene; RT-PCR positive [b]; EST: TC84123[d]
Os08g22780	Putative paramycin	13557735 - 13563792 bp	EST: TC248617 [d]
Os08g23120	Ribonucleoprotein 1	13797728 - 13801842 bp	EST: CK069205, CI084542
Os08g23150	Indole-3-phosphate synthase	13821693 - 13825573 bp	EST: CI222740, CB618959, CI639010
Os08g23790	Polygalacturonase family protein, expressed	14271325 - 14273729 bp	EST: CI141711, AU172573
Os08g24310	LRR protein kinase, expressed	14540216 - 14546582 bp	EST: CI679268, CI615166, CI379618
Os08g24760	Pre-mRNA splicing factor RNA helicase, putative, expressed	14848613 - 14859241 bp	EST: CI120845, CB622777, CI678771

[a] Gene identifiers from Nagaki et al. 2004 are included for genes whose annotations are not presently annotated in the Gramene database.
[b] RT-PCR evidence of expression from Nagaki et al. 2004.
[c] Genbank accession numbers are listed for ESTs unless otherwise noted.
[d] TIGR EST cluste

Table 2. Measures of diversity and divergence for sampled genes in the centromere 8 region.

	Similar to At5g11030	Putative threonyl-tRNA synthetase	Fatty acyl CoA reductase	Putative Ca2+ dependant lipid binding	Putative sucrose-phosphate synthase	formamidopyrimidine-DNA glycosylase	Putative cytokinin inducible	Phosphatidylinositol 3-kinase	Putative ribosomal protein L15	Putative CBS domain containing	Putative poly A binding	Putative poly A binding	Putative paramycin	Ribonucleoprotein 1	Indole-3-phosphate synthase	Polygalacturonase family protein	LRR protein kinase	Putative Pre-mRNA splicing factor, RNA helicase
O. sativa																		
$\pi_{(s)}$	0.0028	0.0035	0.0034	0.0021	0.0014	0.0006	0.0002	0.0011	0.0010	0.0013	0.0039	0.0019	0.0088	0.0007	0.0075	0.0001	0.0017	0.0041
$\theta_{(s)}$	0.0026	0.0038	0.0034	0.0040	0.0017	0.0022	0.0017	0.0034	0.0001	0.0021	0.0041	0.0019	0.0104	0.0007	0.0022	0.0071	0.0004	0.0027
Tajima's *D*	-0.08	-0.22	0.60	-1.19	-0.47	-0.68	-0.39	-0.29	0.30	-0.94	-0.52	-0.47	-1.13	-0.01	-0.10	-1.42	-1.07	0.69
Ka	0.0132	0.0000	0.0061	0.0006	0.0000	0.0061	0.0028	0.0035	0.0000	na	0.0087	0.0000	na	0.0029	0.0000	0.0025	0.0015	na
Ks	0.0173	0.0242	0.0029	0.0010	0.0000	0.0006	0.0314	0.0017	0.0179	na	0.0442	0.0270	na	0.0000	0.0856	0.0186	0.0086	na
indica																		
$\pi_{(s)}$	0.0002	0.0016	0.0002	0.0014	0.0000	0.0010	0.0000	0.0002	0.0000	0.0000	0.0008	0.0003	0.0017	0.0006	0.0004	0.0002	0.0000	0.0011
$\theta_{(s)}$	0.0006	0.0045	0.0006	0.0028	0.0000	0.0020	0.0000	0.0005	0.0000	0.0000	0.0015	0.0009	0.0017	0.0009	0.0009	0.0000	0.0006	0.0000
Tajima's *D*	-1.16	-2.30	-1.51	-1.72	na	-1.14	-0.11	-1.51	na	na	-1.21	-1.16	-1.44	-0.81	-0.93	-1.16	na	-1.14
temperate japonica																		
$\pi_{(s)}$	0.0016	0.0014	0.0012	0.0011	0.0002	0.0000	0.0008	0.0007	0.0004	0.0011	0.0020	0.0008	0.0045	0.0002	0.0010	0.0000	0.0031	0.0030
$\theta_{(s)}$	0.0029	0.0022	0.0023	0.0032	0.0006	0.0000	0.0023	0.0016	0.0006	0.0024	0.0027	0.0017	0.0120	0.0004	0.0024	0.0000	0.0000	0.0037
Tajima's *D*	-1.30	-1.15	-1.74	-2.19	-1.16	-1.31	-1.98	-1.07	-0.64	-1.74	-1.03	-1.18	-2.29	-1.16	-1.90	-1.16	-0.80	0.95
tropical japonica																		
$\pi_{(s)}$	0.0024	0.0047	0.0045	0.0023	0.0013	0.0004	0.0000	0.0002	0.0010	0.0011	0.0020	0.0007	0.0065	0.0000	0.0020	0.0000	0.0019	0.0039
$\theta_{(s)}$	0.0018	0.0045	0.0042	0.0022	0.0012	0.0010	0.0000	0.0005	0.0006	0.0007	0.0020	0.0018	0.0054	0.0000	0.0012	0.0000	0.0000	0.0013
Tajima's *D*	0.73	0.13	0.20	0.03	0.14	0.43	na	-0.19	1.17	1.51	0.02	-1.51	0.30	na	0.93	na	0.87	1.87
O. rufipogon																		
$\pi_{(s)}$	0.0027	0.0075	0.0040	0.0026	0.0013	0.0012	0.0013	0.0010	0.0016	0.0019	0.0037	0.0024	0.0113	0.0009	0.0026	0.0003	0.0021	0.0033
$\theta_{(s)}$	0.0029	0.0051	0.0044	0.0047	0.0015	0.0017	0.0040	0.0009	0.0025	0.0028	0.0048	0.0052	0.0178	0.0016	0.0022	0.0000	0.0005	0.0032
Tajima's *D*	-0.24	1.79	-0.14	-1.20	-0.36	-0.24	-1.15	-0.35	-0.89	-0.87	-0.87	-1.53	-1.13	-0.93	0.25	0.24	-0.99	-0.40
Ka	0.0130	0.0000	0.0054	0.0011	0.0000	0.0064	0.0020	0.0021	0.0000	na	0.0083	0.0001	na	0.0034	0.0003	0.0019	0.0016	na
Ks	0.0095	0.0181	0.0161	0.0097	0.0050	0.0156	0.0215	0.0126	0.0133	na	0.0298	0.0135	na	0.0102	0.0115	0.0011	0.0183	na

Table 3. Rice variety frequencies in each haplogroup.

Variety	Hap A	Hap B	Hap C	Other
indica	0	0	0.824	0.176
temperate japonica	0.833	0.056	0.056	0.056
tropical japonica	0.118	0.647	0	0.235
O. rufipogon	0.300	0.200	0	0.500

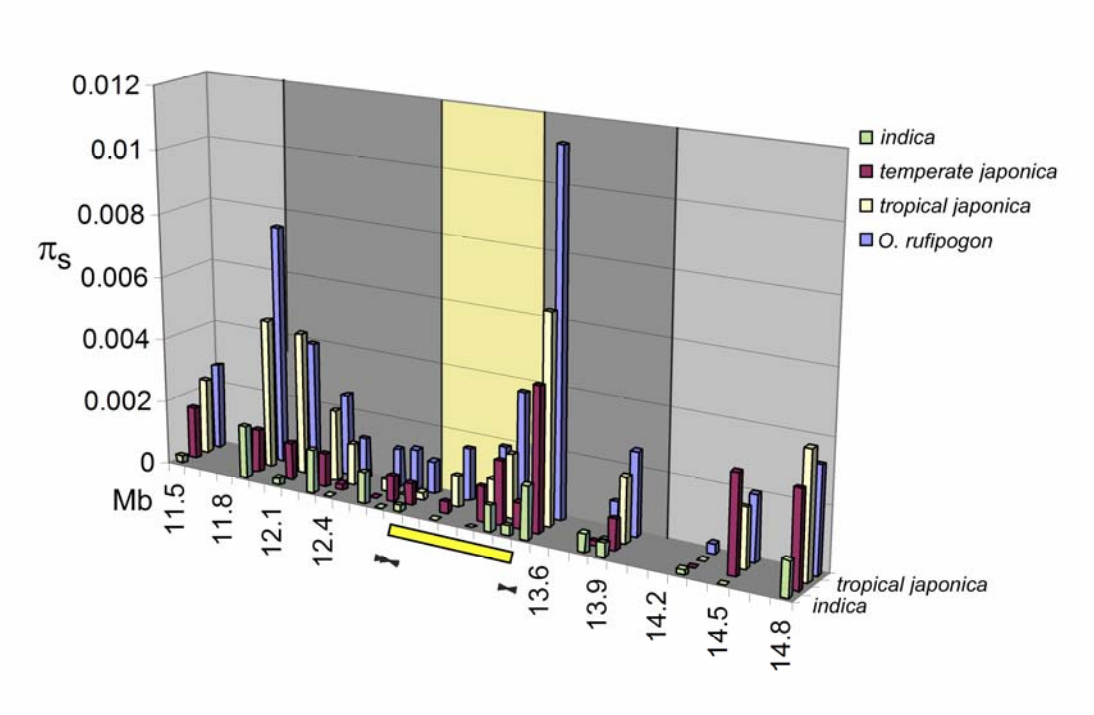


Figure 1. Silent site diversity across centromere 8

Levels of per gene silent site diversity across centromere 8 for domesticated rice varieties and the wild progenitor *O. rufipogon*. The yellow shaded background and bar along the x-axis indicate genes that are located within the kinetochore while the dark grey background represents the previously defined non-recombining region (Harushima et al. 1998). Arrow clusters correspond to the physical locations of the two groups of CRR repeats (Nagaki et al. 2004).

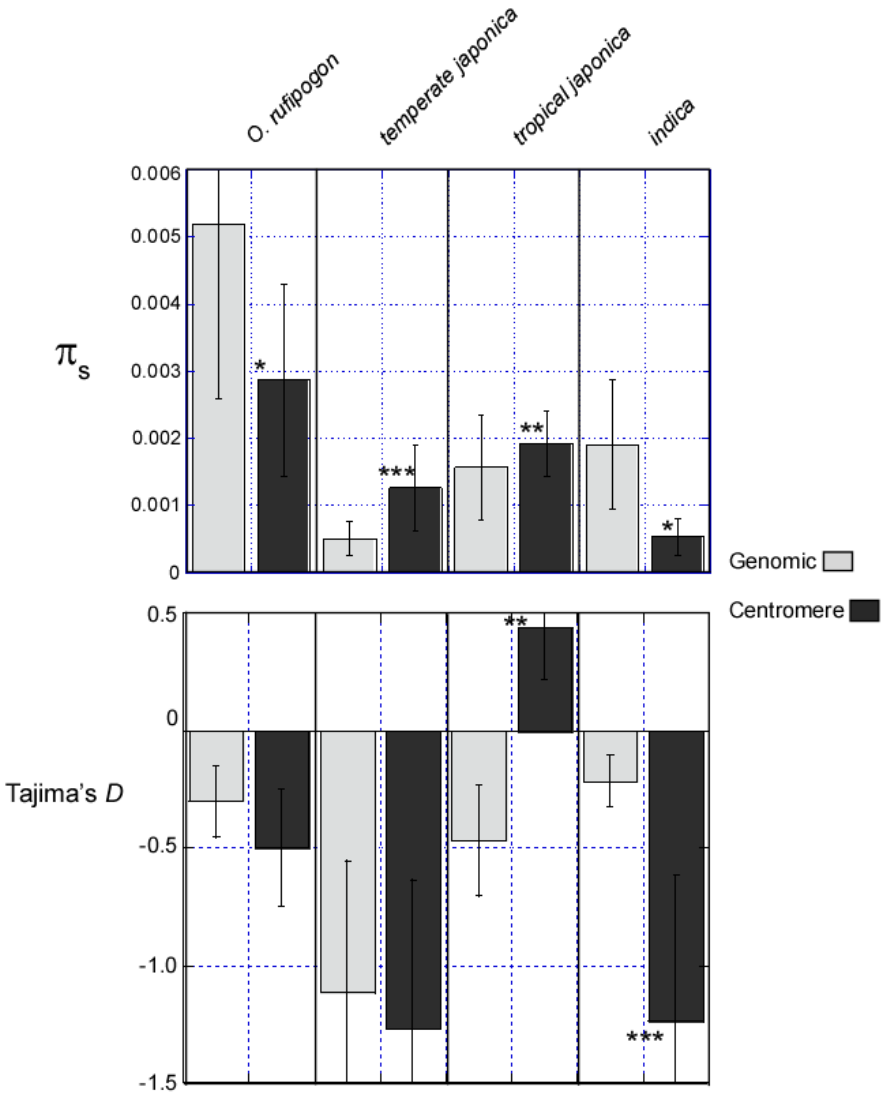


Figure 2. **Centromere vs. genome-wide diversity**

Mean genomic (Caicedo et al., submitted) and centromere silent site diversity and Tajima's *D* for major domesticated rice varieties and the wild progenitor, *O. rufipogon*. Bars represent the range of 50% of the values. Significant differences between levels of genome-wide and centromere diversity (Mann-Whitney test) are indicated by asterisks as follows: P<0.05*, P<0.01**, P<0.001***.

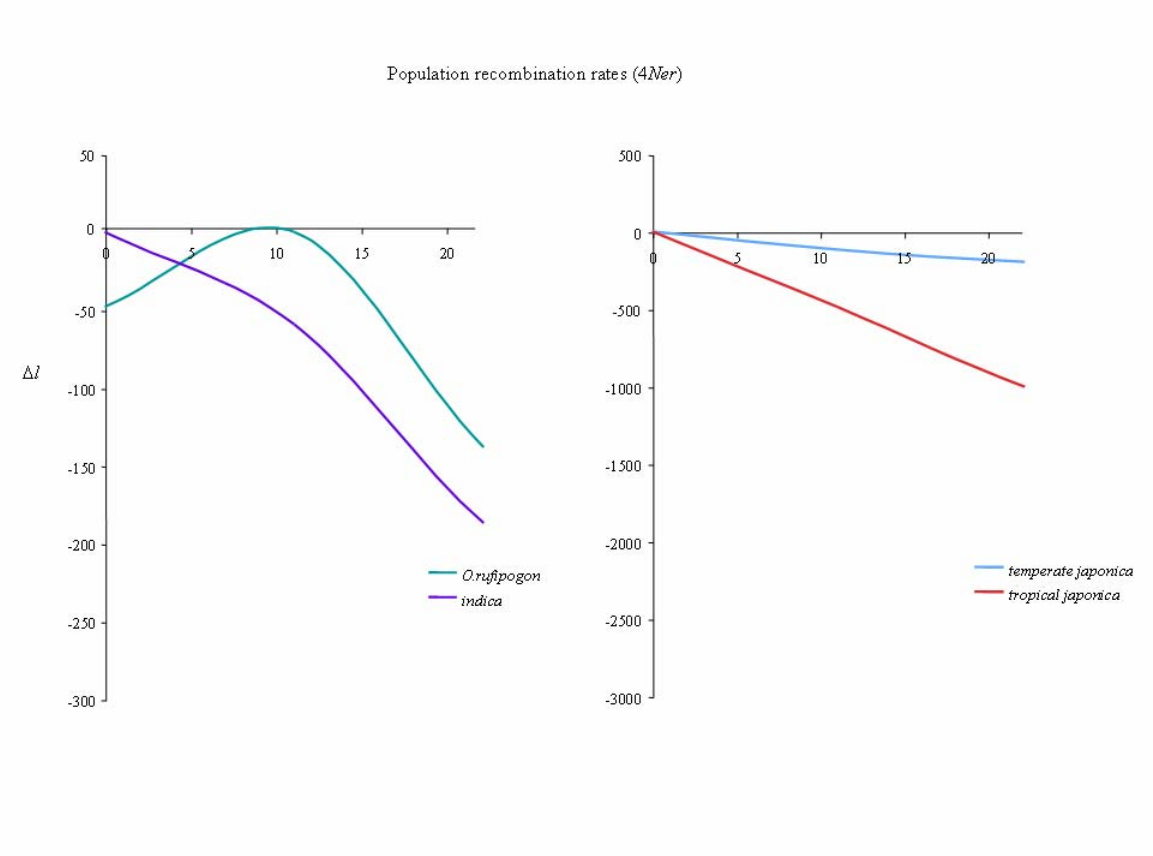


Figure 3. Historical population recombination rates

Plots of the historical population recombination rates ($\rho = 4Nr$) across a 3.3 Mb region encompassing centromere 8, based on composite likelihood. Estimates are significantly different from zero for *O. rufipogon* only.

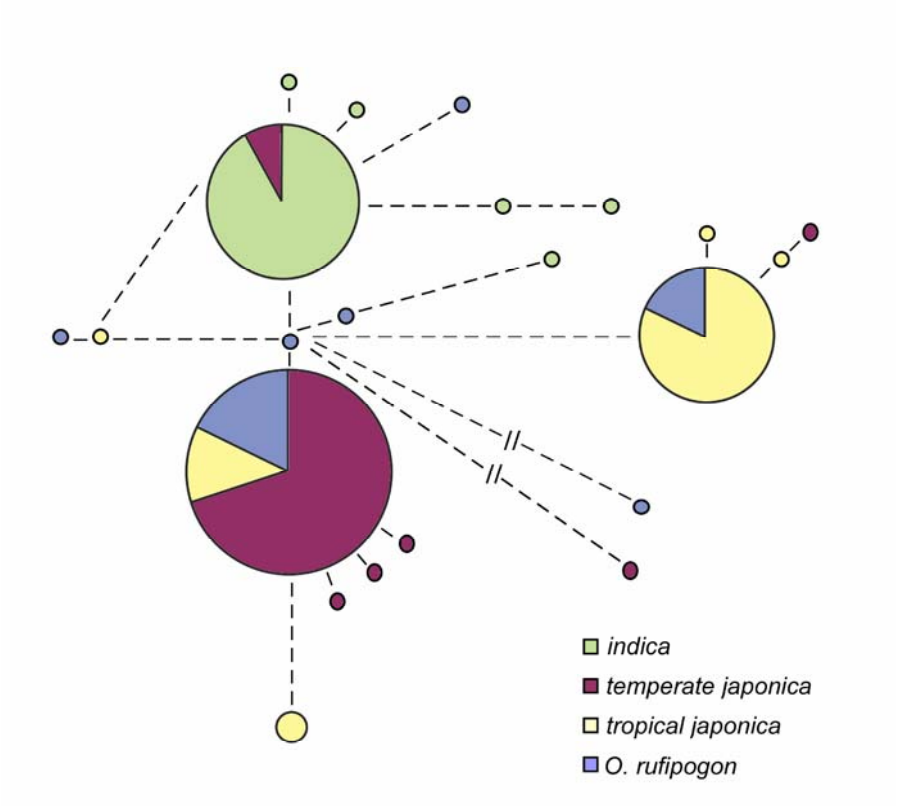


Figure 4. Centromere 8 haplotype network

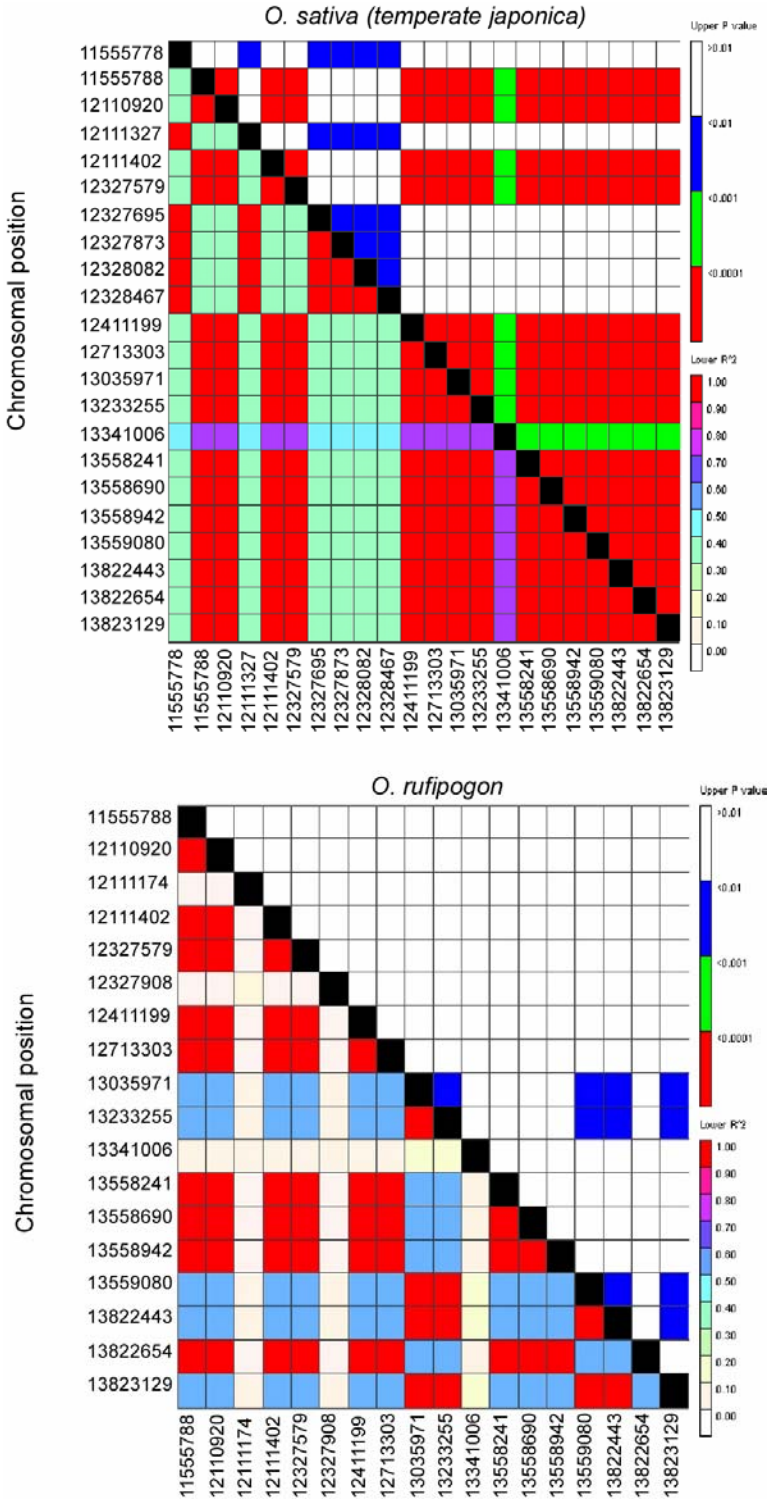
A minimum spanning haplotype network depicting the relationship of haplotypes within and among domesticated rice groups and *O. rufipogon*. Dashed connections represent individual SNPs contributing to between haplotype differences and double slash marks (//) represent 10 SNPs each.

Supplementary Table 1. PCR primers for sampled genes in the centromere 8 region.

Primer id	Gene	Oligo (L)	Oligo (R)
jc8_003	LOC_Os08g23150	GGAGAAGATCATTTGGGACAAGG	TTGAAGTTCTCCCTGAGCACACC
jc8_004	LOC_Os08g23150	CCTCCTGCAAGAGACTTCTACGG	GAGAGCGGGCATAATAGATCTGC
jc8_005	LOC_Os08g23120	GTTGTTGACAGAGTGATTGAGG	CTGCATCAAAGACTACAAATCC
jc8_006	LOC_Os08g23120	GGTGGTGGAACATGAAATCATTCG	CAGGTGATCTATAGGGGCCAAAGC
jc8_009	LOC_Os08g22780	GCTTCAAGCTCTTTCCTTTCACG	CAACTCCTGCCATTCCTGTAAGC
jc8_010	LOC_Os08g22780	CTTACAGGAATGGCAGGAGTTGG	TGCTCTATGGAAGAGTTCGTGAGC
jc8_011	LOC_Os08g22360	ATGCTTCCAAGGGGTAGTCG	GCTGATCAGTTGGGGTATTTGC
jc8_012	LOC_Os08g22360	GCTTCCAGCAACATCTCATTCC	TCATATGGGGATGGCACTAAACC
jc8_013	LOC_Os08g22350	GTGATGGGAAATCTAGATGCTTTGG	GCCAATCATCTTACCATTCATAGCC
jc8_014	LOC_Os08g22350	CATTTAAGTCTGCTGAAGATGC	GAAGGAGGCATCACCATAGG
jc8_015	LOC_Os08g22160	GGTCTCTGGTTTGATGGTTGTACG	GACTTCACAACATGGAGTTCTTCC
jc8_040	LOC_Os08g22160	CAGCTGATGAACTGGTGAACTGC	CAGTTTCTCCATTTGGTCACTGC
jc8_019	LOC_Os08g21590	TTGCATGTGATCCTGAGATTGG	TCCACATTTAAAGTGCGTGTGG
jc8_020	LOC_Os08g21590	CAACTTAAAACTGGGAGGCTGAGG	CTGAAAGACAACTCTGTGTTCAAAGC
jc8_021	LOC_Os08g21350	GCCAGCCATTCAATGAAGACC	CGATTTGTTCGCCTCAATGG
jc8_022	LOC_Os08g21350	CGAAATTTAACAGCGCCTTCC	CTGCAGCCTTTGAATTTGTTGG
jc8_023	LOC_Os08g21330	CCTCATCCTCTTGGATCTCATGG	GGTTGCAACAGCAGTCAGTATCC
jc8_024	LOC_Os08g21330	CTTCCTCTGTGGCCTCTTTGC	ATATGTGCCAGAGTTGCAGAAGC
jc8_025	LOC_Os08g20660	GGCTGAGGAACTTTGTCTTGATGC	GCGAGGGTTGGAGAAGAATCG
jc8_026	LOC_Os08g20660	AAGCTAATGAAGATGGTTCAGG	CAAAACAGCTGAATTTGTGC
jc8_027	LOC_Os08g20530	CTCCAGGCATCTGAGCTTATTGG	GACTACAAATGGGTCAGCCTTGC
jc8_028	LOC_Os08g20530	CCATTGAGGATCAGTTGACATGG	CAGGCTGAAGATCAGACAAGTCC
jc8_029	LOC_Os08g20200	ATGGGTCAAGCACTAAGGGAAGG	CTTGAGGATGCTGGTTATGATGC
jc8_030	LOC_Os08g20200	AAAAGCATGGTGATGGGTTCG	GATATCCAAATGCCTGCCTTCC
jc8_031	LOC_Os08g19850	ATGCTCACTGTGGCTAGTTCAGC	CATTTTGGTTTTGCCCATGC
jc8_032	LOC_Os08g19850	GTCCAAAACAGCCTTAACTTCG	TGATTTTGGAGTCCTTCACAGG
jc8_033	LOC_Os08g19320	GGTTGCCCTTAACATCTTTCAGC	CTTGTGGTTCAGTTGTTCTTCAGC
jc8_034	LOC_Os08g19320	TGCTGAAGAACAACTGAACC	GCTGTTATACCAGAATGTGACG
jc8_035	LOC_Os08g23790	GAACTCGACCCAGTCGTTGC	GTCAGGAACTACGGCGCTAAGG
jc8_036	LOC_Os08g23790	CCACGTCGATGTTGTCGTTCC	GAAGAGGTTCGGCAACGACTGG
jc8_037	LOC_Os08g24310	GTTCCAACAGCCTTAGTGGTTCG	GAAAAGTTGCCCAGAGATGAAGG
jc8_038	LOC_Os08g24310	AATGGTCTGACAGGCACCATACC	GACAATACCTCCAACCGTGTGC
jc8_039	LOC_Os08g24760	GATGGTTTCAGCATCGTTGTCC	ATGTCCCTCCTCGGCTACTCC
jc8_041	LOC_Os08g21840	GTATCAAGGACAACTGCTAGGG	ACTCGAGCAGCTTCACGTTTTGC
jc8_042	LOC_Os08g21840	CTAGACGATACTTCACGCGTTGC	CAGTTTCTCCATTTGGTCACTGC

Supplementary Figure 1. **Linkage disequilibrium across rice centromere 8**

Linkage disequilibrium (LD) given as r^2 across the centromere 8 region. Significance of LD is depicted on the top right portions of each graph and is correlated with r^2. LD between informative SNPs with frequencies >10% are shown for *O. sativa* ssp. *temperate japonica* and *O. rufipogon*. Other rice varieties are omitted since most variants occur with <10% frequency.


O. sativa (temperate japonica)
Chromosomal position
11555778
11555788
12110920
12111327
12111402
12327579
12327695
12327873
12328082
12328467
12411199
12713303
13035971
13233255
13341006
13558241
13558690
13558942
13559080
13822443
13822654
13823129
Upper P value
>0.01
<0.01
<0.001
<0.0001
Lower R^2
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
O. rufipogon
Chromosomal position
11555788
12110920
12111174
12111402
12327579
12327908
12411199
12713303
13035971
13233255
13341006
13558241
13558690
13558942
13559080
13822443
13822654
13823129
Upper P value
>0.01
<0.01
<0.001
<0.0001
Lower R^2
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00

CHAPTER FIVE

Conclusions and Future Directions

CONCLUSION AND FUTURE DIRECTIONS

Understanding the genetic basis of plant adaptation has been a goal of evolutionary biology for over 20 years (KREITMAN 1983), yet few examples demonstrating a connection between adaptive phenotypes and genotypes exist (WRIGHT and GAUT 2005). We have used molecular population genetics to survey patterns of nucleotide variation for evidence of selection in unique regions of the *Arabidopsis* and rice genomes for which theoretical expectations are well established. This research has identified several novel candidate balanced polymorphisms in *Arabidopsis thaliana* and empirical deviations from selection theory in rice highlight an important area where research efforts could be focused in the future.

We conducted a preliminary genome-wide scan for high-diversity regions in the *A. thaliana* genome by comparing ~4,300 shotgun sequence fragments from the L*er* accession genome to the fully sequenced Col-0 genome. Levels and patterns of polymorphism across broader chromosomal regions flanking several of the identified high-diversity fragments revealed significant deviations from neutrality and were consistent with both theoretical (CHARLESWORTH *et al.* 1997; CHARLESWORTH 2006; HUDSON and KAPLAN 1988) and empirical (TIAN *et al.* 2002) expectations for the molecular footprint of balancing selection (chapter 1). Observations of potentially functional non-synonymous polymorphism at *BLH10* and differential expression phenotypes among alleles at *SSI* and *PTAC7* provide additional support for the hypothesis that balancing selection might act to maintain functional variation at these candidate loci (chapter 2). Our results demonstrate the utility of genome scanning for novel targets of

balancing selection in *A. thaliana* and provide the basis for similar more comprehensive studies in the future.

The ability to complement traditional forward genetic screens for targets of natural selection with reverse approaches, such as genome scanning, represents an important advancement for the field of evolutionary genomics. The utilization of genome scanning has aided in the identification of new candidate adaptive trait loci and provides a foundation for the study of broader evolutionary questions such as: How prevalent are various forms of selection in organismal genomes? and which biological functions are most often targeted? Indeed comprehensive genome-wide scans for evidence of selection in the human HapMap data have already revealed that natural selection is likely to have played an important role as human populations adapted to new environments outside of Africa (VOIGHT *et al.* 2006; WILLIAMSON *et al.* 2007).

One caveat with genome scanning is that often long lists of candidate genes are generated, while the challenge of linking genetic variants with their functional consequences remains largely unmet. Connecting adaptive phenotypes to patterns of genomic variation will likely be easiest in model genetic systems like *Arabidopsis* and *Drosophila* since functional genomics will necessarily be a key component of this effort. In our study, a precise function had been previously established for only one of our three identified candidate genes (*SSI*), and patterns of variability seemed most consistent with selection acting at a linked downstream locus (*PTAC7*) instead, thus prohibiting an allele specific functional assay of the known end-point phenotype. Alternatively, allele specific expression differences can be easily measured in the absence of functional knowledge and many efforts, including ours, have placed a focus on using expression phenotypes to

infer functional genetic variation (DE MEAUX *et al.* 2005; DE MEAUX *et al.* 2006; LOISEL *et al.* 2006; PRIGODA *et al.* 2005; WITTKOPP *et al.* 2004). Once more gene functions are clearly defined it will be possible to design a variety of fitness assays for multiple top candidate genes that are specifically tailored to appropriately challenge the phenotypes in question. This would greatly progress the utility of the genome scanning approach and the goal of the *Arabidopsis* community to identify the functions of every gene in the genome by 2010 (CHORY *et al.* 2000) places such possibilities in the near future.

The potential functional (and fitness) effects of differential allelic expression at *SSI* and *PTAC7* and of the nonsynonymous variation at *BLH10* warrant further investigation. It would be interesting to see if the signature of selection could be detected at these loci in other related plant species, for instance, *A. thaliana's* sister species, *Arabidopsis lyrata*. If evidence of selection were to replicate, further insights could be gleaned from population genetic investigations in an outcrossing species with a well known ecology and clearly defined populations. This would allow for a direct assessment of the distribution of alleles using measures such as F_{st} to determine whether functional variants are shared within or distributed between historical populations of a species and would also permit finer scale mapping to identify the specific functional variant(s). Understanding the mechanism responsible for the maintenance of diversity at these loci will be essential if the functional and ecological significance of these genes is to be fully explained.

We also investigated the evolutionary dynamics of rice centromere 8 with particular attention to the relationship between recombination and diversity since population genetic theory predicts that both directional selection on adaptive variants

(MAYNARD SMITH and HAIGH 1974) and background selection against deleterious mutations (CHARLESWORTH *et al.* 1993) will produce extreme reductions in diversity across such low-recombining regions (chapter 3). DNA sequence data from ~18 loci distributed across a ~3.3 Mb region encompassing rice centromere 8 revealed no evidence of selection in the three investigated domestic varieties (*indica*, *temperate japonica*, and *tropical japonica*) or the wild progenitor (*O. rufipogon*). Surprisingly, we observed significantly elevated levels of nucleotide variation in *temperate japonica* and *tropical japonica* centromere 8 regions relative to empirical genome-wide distributions. Although *indica* and *O. rufipogon* displayed the expected reductions in diversity, demography or selective interference seem most likely to explain patterns of diversity and divergence at rice centromere 8.

While theoretical predictions for a positive correlation between recombination and diversity are consistently upheld in animal models including *Drosophila* (AGUADE *et al.* 1994; AGUADE *et al.* 1989; LANGLEY *et al.* 2000; STEPHAN and LANGLEY 1989), mouse (NACHMAN 1997; TAKAHASHI *et al.* 2004) , and humans (NACHMAN 2001), this correlation seems to be less strong in many plant species (BAUDRY *et al.* 2001; NORDBORG *et al.* 2005; STEPHAN and LANGLEY 1998; TENAILLON *et al.* 2002; WRIGHT *et al.* 2005; WRIGHT *et al.* 2006). One intriguing question is whether deviations from theoretical expectations at rice centromere 8 represent centromere specific evolutionary dynamics, or whether the lack of correlation between recombination and diversity occurs more broadly throughout the genome. Given the prevalence of artificial selection during the domestication process and observations of extreme reductions in genome-wide diversity among domesticated varieties relative to the wild ancestor it is somewhat

surprising that directional hitchhiking does not leave its footprint in this genomic region. This may lend some support to the idea that these findings describe the specific evolutionary processes occurring at centromere 8; indeed no other region of the rice genome harbors so many functional genes linked in a common non-recombining region and our data suggest that Hill-Robertson interference might play a major role in shaping patterns of nucleotide diversity during the intermediate stages of centromere formation. Future work to assess the level of diversity across regions of the genome with respect to local population recombination rates would help to distinguish between these two possibilities.

It is clear that more studies outlining general relationship between recombination and diversity are needed. This will provide a deeper understanding of the prevalence of different forms of selection in organismal genomes. However, if such investigations are to yield useful mechanistic insight it will be important that they have the power to distinguish between the stochastic and deterministic processes that act to shape patterns of intraspecific diversity. Since selection is a stochastic process and is likely to act only in certain populations under certain conditions, sampling within and between populations or groups of related species should aid in this effort.

REFERENCES

AGUADE, M., W. MEYERS, A. D. LONG and C. H. LANGLEY, 1994 Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the su(s) and su(wa) regions of the Drosophila melanogaster X chromosome. Proc Natl Acad Sci U S A **91:** 4658-4662.

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced Variation in the Yellow-Achaete-Scute Region in Natural Populations of Drosophila Melanogaster. Genetics **122:** 607-615.

BAUDRY, E., C. KERDELHUE, H. INNAN and W. STEPHAN, 2001 Species and recombination effects on DNA variability in the tomato genus. Genetics **158:** 1725-1735.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, B., M. NORDBORG and D. CHARLESWORTH, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res **70:** 155-174.

CHARLESWORTH, D., 2006 Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet **2:** e64.

CHORY, J., J. R. ECKER, S. BRIGGS, M. CABOCHE, G. M. CORUZZI *et al.*, 2000 National Science Foundation-Sponsored Workshop Report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. Plant Physiol **123:** 423-426

DE MEAUX, J., U. GOEBEL, A. POP and T. MITCHELL-OLDS, 2005 Allele-specific assay reveals functional variation in the chalcone synthase promoter of Arabidopsis thaliana that is compatible with neutral evolution. Plant Cell **17:** 676-690.

DE MEAUX, J., A. POP and T. MITCHELL-OLDS, 2006 Cis-regulatory evolution of chalcone-synthase expression in the genus Arabidopsis. Genetics **174:** 2181-2202.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831-840.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature **304:** 412-417.

LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics **156:** 1837-1852.

LOISEL, D. A., M. V. ROCKMAN, G. A. WRAY, J. ALTMANN and S. C. ALBERTS, 2006 Ancient polymorphism and functional variation in the primate MHC-DQA1 5' cis-regulatory region. Proc Natl Acad Sci U S A **103:** 16331-16336.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genetical Research **23:** 23-35.

NACHMAN, M. W., 1997 Patterns of DNA variability at X-linked loci in Mus domesticus. Genetics **147:** 1303-1316.

NACHMAN, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet **17:** 481-485.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3:** e196.

PRIGODA, N. L., A. NASSUTH and B. K. MABLE, 2005 Phenotypic and genotypic expression of self-incompatibility haplotypes in Arabidopsis lyrata suggests unique origin of alleles in different dominance classes. Mol Biol Evol **22:** 1609-1620.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. Genetics **121:** 89-99.

STEPHAN, W., and C. H. LANGLEY, 1998 DNA polymorphism in lycopersicon and crossing-over per physical length. Genetics **150:** 1585-1593.

TAKAHASHI, A., Y. H. LIU and N. SAITOU, 2004 Genetic variation versus recombination rate in a structured population of mice. Mol Biol Evol **21:** 404-409.

TENAILLON, M. I., M. C. SAWKINS, L. K. ANDERSON, S. M. STACK, J. DOEBLEY *et al.*, 2002 Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). Genetics **162:** 1401-1413.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A **99:** 11525-11530.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. PLoS Biol **4:** e72.

WILLIAMSON, S. H., M. J. HUBISZ, A. G. CLARK, B. A. PAYSEUR, C. D. BUSTAMANTE *et al.*, 2007 Localizing Recent Adaptive Evolution in the Human Genome. PLoS Genet **3:** e90.

WITTKOPP, P. J., B. K. HAERUM and A. G. CLARK, 2004 Evolutionary changes in cis and trans gene regulation. Nature **430:** 85-88.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310-1314.

WRIGHT, S. I., J. P. FOXE, L. DEROSE-WILSON, A. KAWABE, M. LOOSELEY *et al.*, 2006 Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. Genetics **174:** 1421-1430.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol **22:** 506-519.

APPENDIX

The evolution of molecular genetic pathways and networks

Jennifer M. Cork[1]* and Michael D. Purugganan[1]

[1]Department of Genetics, Box 7614

North Carolina State University, Raleigh, NC 27695

*Name change; this author is now Jennifer M. Reininga

REFERENCE:

Cork, J. M., and M. D. Purugganan. 2004. The evolution of molecular genetic pathways and networks. *BioEssays*. **26**:479-484.

The evolution of molecular genetic pathways and networks

Jennifer M. Cork and Michael D. Purugganan*

Summary

There is growing interest in the evolutionary dynamics of molecular genetic pathways and networks, and the extent to which the molecular evolution of a gene depends on its position within a pathway or network, as well as over-all network topology. Investigations on the relationships between network organization, topological architecture and evolutionary dynamics provide intriguing hints as to how networks evolve. Recent studies also suggest that genetic pathway and network structures may influence the action of evolutionary forces, and may play a role in maintaining phenotypic robustness in organisms.
***BioEssays* 26:479–484, 2004.**

Introduction

Most biochemistry teaching laboratories (and indeed also research laboratories) display a wall-sized chart that depicts the complexities of organismal metabolism. This chart, known as the official International Union of Biochemsitry and Molecular Biology-Nicholson Chart of Metabolic Pathways, had modest beginnings nearly half a century ago.[1] Don Nicholson, a University of Leeds biochemist, sketched out all the known metabolic pathways and their intricate linkages in 1960 to use as a teaching aid; the 21st edition of the chart published in 2000 incorporates 550 enzymatic transformations. It is arguably the first large-scale molecular network assembled, a prelude to many of the molecular genetic networks currently under construction with contemporary genomics studies.

The metabolic pathways chart illustrates a central feature of biological organization at all hierarchical levels—no biological entity within our biosphere, whether metabolites, proteins, genes, cells or even species, exists in isolation. Instead, they are found as components of complex networks and pathways that together constitute organization at every level of biology.[2–5] Molecular genetic networks and pathways describe the molecular and/or genetic components that underlie cellular and organismal processes, and the functional interactions among them (see Fig. 1).[4,5] Pathways can be thought of as small linear components of larger global networks, and it might be useful to consider them as local networks in which interactions are depicted linearly or which possess simple unreticulated branch points. Network patterns are often more complex than simple linear pathways, and usually involve functional cross-connections that create webs of interactions.

It is clearly one of the objectives of biological research to identify and characterize these networks,[4,5] a task that is daunting, but increasingly possible. New high-throughput techniques of genomic analysis provide tools towards achieving these goals at molecular and cellular levels.[6] As the number of characterized molecular genetic networks increases, there are growing opportunities to dissect patterns of gene evolution within a network context.

Molecular evolutionary analyses have previously focused on individual genes outside of the interaction context. It is now clear that there is a need to understand how evolutionary forces act on multiple interacting genes that are components of molecular genetic pathways and networks. The network organization inherent to molecular genetic systems raises many questions regarding the evolutionary dynamics of interacting molecular systems.[7–9] How does the organization of genes as members of interacting pathways affect the rates of evolutionary change? How does the topology of molecular genetic networks constrain evolutionary forces acting on component genes? How do networks as entities evolve? Here we describe some recent studies on the molecular evolution of genetic networks, and how network structures can constrain or channel evolutionary processes, both at the microevolutionary and macroevolutionary levels.

Targets of selection in molecular genetic pathways

The organization of a molecular genetic pathway can provide a useful framework for the study of gene evolution, and the study of gene evolution in this context has been an important first step towards understanding how evolution acts differently on genes in a locally interacting system. Since pathways are, in essence, modular components of networks (see Fig. 1A), the study of pathways also provides a foundation for the study of global networks.

An area of interest has been how the position of genes in molecular genetic pathways can impact the levels and types of selective forces that act on these loci. One hypothesis

Department of Genetics, North Carolina State University.
Funding agency: In part by the US National Science Foundation.
*Correspondence to: Michael D. Purugganan, Department of Genetics, North Carolina State University, Box 7614, Raleigh, NC 27695.
E-mail: michaelp@unity.ncsu.edu
DOI 10.1002/bies.20026
Published online in Wiley InterScience (www.interscience.wiley.com).

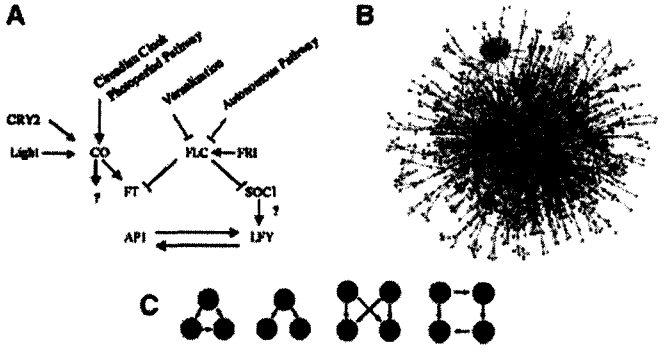


Figure 1. Three levels of organization at which the evolution of genes and their protein products have been studied in their interaction context. **A:** The *Arabidopsis thaliana* flowering-time pathway, shown as an example of a modular component of a genetic interaction network. Activation and repression interactions are depicted, along with several environmental signal inputs. **B:** A map of the yeast protein–protein interaction network from Ref. 31, as an example of global network organization (reproduced from Jeong H, Mason SP, Barabasi AL, Oltavi ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42 with permission of Nature Publishing Group). This network is represented by nodes (proteins) connected to one another by edges (interactions). Some proteins in the network have a large number of protein–protein interactions, while the majority of proteins have very few, resulting in a scale-free topology. **C:** Examples of network motifs (Refs. 43 and 44). Several three- and four-protein interactions motifs show types of recurring themes in network organization. Circles represents proteins and arrows show their interactions.

predicts that genes functioning early in a genetic pathway are subject to stronger stabilizing selection than downstream loci, since mutations in these genes are likely to have greater pleiotropic effects and affect all downstream phenotypes.

In a study on the rates of gene evolution in the plant anthocyanin biosynthetic pathway, differential rates of molecular evolution have been noted for enzymatic genes.[10] In comparisons between monocots and dicots, the nonsynonymous substitution rates of the downstream genes DFR, ANS and UF3GT evolve faster than the upstream loci CHS-D, CHI and F3H. The correlation between the protein evolutionary rate and the position of the enzyme in the anthocyanin pathway is significant ($P < 0.01$). This pattern has also been confirmed at the intragenic level between species in the genus *Ipomoea*.[11] In this pathway, upstream enzymes are positioned above major branch points, and mutational changes in these enzymatic loci are likely to have pleiotropic effects on the synthesis not only of anthocyanins but of flavonoids as well.[10] The pleiotropy of the upstream genes, and the greater specialization in biosynthetic functions of downstream enzymatic loci, appears to result in greater stabilizing selection on genes that act earlier in this biosynthetic pathway.

Greater constraint on earlier acting genes is also observed in studies of molecular variation in regulatory and signal transduction genes within populations. The *Ras*-mediated signal transduction pathway is an evolutionarily conserved genetic pathway that plays a central role in cell differentiation.[12] A molecular population genetic study suggests differential evolution among signal transduction loci in the *Ras* pathway in *Drosophila melanogaster*, indicating strong purifying selection on *Ras*, *Drk* and *polehole*, three upstream genes in the signaling pathway.[13] These three genes have only a few low-frequency replacement polymorphisms within *D. melanogaster* and no fixed amino acid differences between *D. melanogaster* and *D. simulans*. In contrast, there are several interspecific amino acid differences in downstream genes *Dsor1*, *corkscrew* and *Ksr*. In *Ras*-mediated signal transduction, the upstream genes act as control points, which may explain the greater degree of stabilizing selection on these loci. This study also suggests that the downstream components, which act as modifiers of *Ras* kinase signaling, are the most-likely source of quantitative phenotypic variation in this regulatory pathway.[13]

Other studies do not support the hypothesis that earlier-acting genes in genetic pathways are under strong stabilizing selection. In the *Arabidopsis thaliana* floral developmental pathway, upstream genes appear to be the targets of positive, not stabilizing, selection.[9] A molecular population genetic study of six genes in the *Arabidopsis* floral developmental pathway suggests that four downstream transcription factor genes (the floral meristem identity genes *AP1* and *CAL*, and the floral organ identity genes *AP3* and *PI*) have neutral patterns of molecular evolution. In contrast, the two earliest-acting genes in the study, the inflorescence architecture gene *TFL1* and the floral meristem identity gene *LFY*, show a significant reduction in silent site nucleotide variation consistent with a recent adaptive sweep.[9] Moreover, the promoter alleles of *TFL1* are differentiated into two distinct haplotype groups that

may be maintained by selective forces.[9] These results indicate that, in the floral developmental pathway, it is the earliest-acting genes that are likely to be the targets of recent selective forces.[9]

The difference in selective patterns between the *A. thaliana* floral regulatory genes and the *D. melanogaster Ras* signaling loci is noteworthy, but may be explicable in the context of the functions of these two developmental pathways. Mutations in the latter have potentially strong pleiotropic effects on fly development, since these loci affect several developmental processes from oogenesis to appendage morphogenesis.[12] In contrast, the floral developmental pathway is associated with a relatively specific and discrete developmental module (the flower), and the activity of *TFL1* and *LFY* function to specify when and where this entire module is established.[9] Thus, these genes form a discrete "gene net",[14] and any selection on evolutionary change in the spatial and temporal establishment of floral developmental modules would be more likely to act on these upstream loci in the regulatory hierarchy. Gene position within a molecular genetic pathway may indeed result in differential selection on component genes, but the specific nature of the selective forces on the component loci will largely depend on the function of the pathway.

Differential selective forces on different components of a pathway at a microevolutionary level may be taken to a macroevolutionary extreme in considering the diversification of sex-determination pathways.[15] In sex-determination cascades, downstream components appear to have conserved expression patterns (e.g. *Sox 9* between birds and mammals),[16] while upstream components, such as *Sry*, have a more circumscribed distribution even within mammals.[17] The observed plasticity of upstream components is also seen in other cases within the Diptera.[18,19] together, these observations suggest either retrograde evolution in pathway expansion or evolutionary flexibility in upstream (but not downstream) pathway genes.[15] The continued study of the evolution of sex-determination pathways may provide further insights into the nature of selective forces that not only impact on the evolution of genes, but also on the evolutionary construction of genetic pathways.

Metabolic networks and branch point evolution

The dichotomy between upstream and downstream genes in a pathway is a crude differentiation of function. For metabolic pathways, a theory of pathway fluxes provides a framework for developing more precise hypotheses regarding selection on component genes based on position within the pathway and/or specific biochemical functions. Metabolic control theory[20] describes how pathway architecture constrains evolutionary change by depicting how metabolic fluxes might be partitioned into alternate channels through selection.[21]

The evolution of genes in relation to their position in pathway branch points has been explored most explicitly in the context of population variation in the glycolytic pathway in *D. melanogaster*.[21] The PGM and G6PD enzymes sit at the top of the glycolytic pathway and partition glucose into glycogen and pentose shunt branches.[21] Nucleotide and allozyme variation at these two glycolytic loci show clinal variation across latitudes.[22,23] An excess of within-species amino acid changes is observed at the *PGM* gene, with 12 of 25 within-species coding region polymorphisms being replacement polymorphisms.[24] In contrast, none of the 19 fixed differences between *D. melanogaster* and *D. simulans* result in amino acid changes. This pattern suggests selection for the maintenance of protein variation in *D. melanogaster*. The *G6PD* locus shows a different pattern, with a significant excess of amino acid differences fixed between species. This pattern is consistent with positive selection favoring protein divergence between the two species.[25] Variation for these genes, as assayed by allozyme polymorphism, is also observed in species outside of *Drosophila*.[26,27]

Selection at enzymatic branch points is also observed in the evolution of starch biosynthesis in *Zea mays*.[28] Maize domestication has been accompanied by selection for altered starch content in kernels. A molecular population genetic study of six starch biosynthetic enzymes show reduced variation in three enzymatic loci—*bt2*, *ae1* and *su1*—associated with selective sweeps.[28] The latter two genes encode enzymes that function at branch points. Both are involved in amylopectin synthesis, partitioning ADP-glucose between amylose and amylopectin. As in *Drosophila* glycolytic pathway studies, selection on the maize starch biosynthetic pathway involved pathway branchpoints that function by partitioning substrates into alternate product pools.

Together, these results indicate that genes that act at metabolic pathway branch points are targets of adaptive forces. These studies also suggest multilocus responses to selection,[21] and provide clear examples of how a theoretical framework for pathway fluxes could yield insights into the nature of selective forces on molecular genetic pathways. More detailed studies on the evolution of molecular genetic networks will also benefit from more rigorous general theoretical models on network structure and dynamics, particularly on the phenotypic consequences of mutational change on network components. Development of these models for developmental regulatory systems,[8,29,30] for example, may provide the basis for a coherent study of the evolution of developmental phenotypes.

Global structure and the evolutionary growth of networks

Although there has been some progress in our understanding of pathway evolution, we are still left with several unanswered questions relating to the evolution of larger networks. Does

selection act differently at local (pathway) and global (network) scales? Is there selection for global network structure? How does the global network grow? Empirical studies of global network structure and evolution have only recently become possible and are beginning to address these and other related questions.

The identification of extensive protein–protein interaction networks in *Saccharomyces cerevisiae* has provided the basis for several early analyses of network evolution. One of the first examples of such a network consists of 1,870 proteins related to one another through a web of 2,240 interactions (see Fig. 1B). This protein interaction network is characterized by a scale-free topology, in which different proteins have varying numbers of interactions—a small number of proteins have a large number of interactions, while the majority of proteins have very few connections.[31] This scale-free topology has been shown to be a common characteristic of protein interaction networks,[31,32] metabolic networks[33] and *Caenorhabditis* expression networks determined from yeast, human, *Drosophila* and *C. elegans* microarray data.[34]

How does this scale-free network architecture originate? There have been some suggestions that this architecture may arise from selection for genetic robustness (see below). Others have suggested, however, that the emergence of a scale-free network topology can be accounted for without assuming the involvement of natural selection on global network structure.[35] Biological justification for the "growing network model"[35] reveals that there are only two requirements for the evolution of a scale-free network structure in the yeast protein interaction network: (1) the addition of new nodes to the network and (2) the preferential attachment of these new nodes to already highly connected nodes.[35,36,37] Although this model does not explicitly invoke positive selection in the large-scale organization of networks, it should be noted that selection may be implicitly involved since preferential attachment of new nodes may be driven by selective forces acting at the local level.

An analysis of divergent gene duplicates in the yeast protein interaction network provides some support for the preferential attachment assumption. The relationship between protein connectivity and the likelihood of gaining interaction partners was shown to be close to linear, indicating that proteins with larger numbers of interactions are more likely to gain new connections.[35] Examining the addition of new network interaction links through evolutionary time also permits testing of the preferential attachment hypothesis.[35] "Snapshots" in evolutionary time were generated for the yeast protein interaction network through genome-wide comparisons of yeast with *E. coli*, *A. thaliana* and *S. pombe*. Each protein was classified in one of four age groups based on its presence or absence within the comparison species, which allowed the observation of changes in protein connectivities over evolutionary time. It was demonstrated that new links in the network are more likely to be added through interactions with proteins that are already highly connected, which then leads to the emergence of network scale-free topology.[37]

Interestingly, Qin et al. advanced a slightly different view of network evolution for the yeast protein interaction network.[38] In their analysis, they suggested that proteins of similar ages (which they refer to as isotemporal categories) have a greater tendency to interact with each other than if they were in different age groups. Moreover, it appears that the networks tend to grow via addition of modules or groups of interacting proteins rather than single protein additions.[38] It would be instructive to examine whether network growth is driven by connectivity or evolutionary age, although it may be difficult to disentangle these two factors given that they may be related to some degree.

Gene evolution, network position and network motifs

The evolutionary rates of genes and/or proteins appear to be related in part to their network position. A negative correlation between connectivity and rate of protein evolution has been observed in the yeast protein interaction network (although this relationship has been debated[39]), with highly connected proteins in the network showing the slowest rates of evolution.[40] The authors suggest that this effect arises from highly connected proteins having most of their structure involved in functional interactions, and thus under greater selective constraint. Evolutionary changes at functionally important protein sites might thus be attributed largely to co-evolutionary diversification between interacting protein partners, and proteins that interact with one another have similar evolutionary rates.[40] Other correlations between protein age, connectivities and function have also been found.[37] Finally, essential proteins also appear to be older[37] and to have greater connectivities and slower rates of evolution;[31,40] however, this has also been questioned.[41]

These relationships are also buttressed by examining the evolution of network motifs, local patterns of interaction that recur at different positions within a network[42] (see Fig. 1C). Wuchty et al. classified topological motifs in a *S. cerevisiae* protein interaction database derived primarily from two-hybrid studies, which include 3,183 interacting yeast proteins.[43] These include all possible two-protein, three-protein, four-protein, and a few five-protein interaction motifs, and the database specifies between 10^3 and 10^6 copies of each motif. Conservation of motifs across evolutionary time was judged based on the presence of an orthologous protein across five eukaryotic species in plants, animals and fungi.[43]

The larger motifs with higher connectivities were conserved to a greater extent than smaller motifs; 47% of fully connected five-protein motifs were conserved, compared to less than 5% of linear three-protein motifs. The evolutionary retention rate of

every network motif was higher than expected by random chance, and the ratio of observed number of motifs over random expectation was higher the larger the motif. Interestingly, the composition and frequency of motifs is related to biochemical function. Large motifs had functionally homogenous components. In 95% of conserved fully connected five-protein motifs, all the protein components share at least one biochemical function class (i.e., regulation, protein fate, cell cycle, etc.). In contrast, only 10% of the two-protein motifs were functionally homogenous. Moreover, different functional classes have different characteristic motifs. Regulation, cellular transport and transport facilitation have only one or two characteristic motifs, while all 11 studied motifs were overrepresented in modules associated with subcellular localization, protein fate and transcription.[43] In addition, it should be noted that previous studies identified network motifs that also appear to be commonly involved with certain functional roles.[42,44,45]

Network connectivities and genetic robustness

The impact of the topological architectures of networks on their ability to withstand mutational perturbations is an exciting avenue of exploration, and could lead to refinement of hypotheses on selective targeting of genes during evolution. The observation that a large number of genetic changes may be buffered from expressing mutational variation has traditionally been ascribed to the presence of gene duplications that lead to genetic redundancies.[46–48] Recently, the role that genetic network organization may play in maintaining robustness is gaining increased attention.[46,49] It has been shown, for example, that highly connected proteins in the yeast protein interaction network are three times less likely to be tolerant to loss-of-function mutations than proteins with fewer connections.[31] This observation, coupled with the scale-free topologies in protein interaction networks, has led to suggestions that selection constructs a network that is phenotypically robust against mutations. If network structure does provide an organism with phenotypic robustness against mutation, then this may be one explanation for the phenomenon of genetic canalization.[50,51]

The identification of molecular genetic networks also provides avenues for exploring the molecular basis of evolutionary epistasis. It has always been recognized that network organization leads to genetic interactions that may result in epistasis.[8,52] Candidate epistatic interactions can now be systematically identified in the context of known genetic or physical interaction networks, and may provide a molecular basis for determining the extent and nature of evolutionary epistasis.

Conclusions

The study of the evolution of genes and genetic systems has entered a new phase. The advent of genomics technologies has provided opportunities for expanded analyses, which permit studies of evolutionary change at genome-wide scales. There is now the possibility of studying the evolution of whole genetic networks rather than single loci, and examining the implications of network organization on the dynamics of evolutionary change. Although much has been accomplished by our attempts to define molecular genetic networks, there is still a great deal of work to be done and many more details to consider.

There still remain several areas of network analysis that could help spur further evolutionary investigations, including the development of precise, quantitative models that relate network topologies and dynamics to their phenotypic consequences. Eventually the depiction of molecular genetic networks will have to allow for the incorporation of spatial and temporal regulatory mechanisms. The complete network of cellular interactions, after all, involves much more than just a depiction of who regulates and/or interacts with whom. The development of these models will provide evolutionary studies with a functional context and allow for the integration of molecular, population and quantitative genetic studies. Furthermore, analyses of global network structure will prove fruitful not only for understanding global evolutionary pressures, but also of forces acting at a local level and will allow us to consider pathway structure and evolution in a more realistic functional context.

Acknowledgments

We thank members of the Purugganan laboratory for a critical reading of the manuscript, as well as Dr. Adam Wilkins and two anonymous reviewers for insightful comments and ideas.

References

1. Dixon B. 1988. Metabolic pathways chart: an all-time bestseller. The Scientist 2:23.
2. Loomis WF, Sternberg PW. 1995. Genetic networks. Science 269:649.
3. Melian CJ, Bascompte J. 2002. Complex networks: two ways to be robust? Ecol Lett 5:705–708.
4. White KP. 2001. Functional genomics and the study of development, variation and evolution. Nat Rev Genet 2:528–537.
5. Slonim DK. 2002. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32:502–508.
6. DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680–686.
7. Wilkins AS. 2002. The evolution of developmental pathways. Sunderland, MA: Sinauer.
8. Frank SA. 1999. Population and quantitative genetics of regulatory networks. J Theor Biol 197:281–294.
9. Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the Arabidopsis thaliana floral developmental pathway. Genetics 161:1641–1650.
10. Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16:266–274.
11. Lu Y, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol 20:1844–1853.
12. Khosravi-Far R, Rossman K, Clark G, Der C. 1998. Increasing complexity of Ras signaling. Oncogene 17:1395–1413.

13. Riley RM, Jin W, Gibson G. 2003. Contrasting selection pressures on components of Ras-mediated signal transduction in Drosophila. Mol Ecol 12:1315–1324.
14. Bonner JT. 1988. The evolution of complexity. Princeton, NJ: Princeton University Press.
15. Stothard P, Pilgrim D. 2003. Sex-determination gene and pathway evolution in nematodes. BioEssays 25:221–231.
16. Morais da Silva S, Hacker A, Harley V, Goodfellow P, Swain A, Lovell-Badge P. 1996. Aoz9 expression during gonadal development implies a conserved role for the gene in testis defferentiation in mammals and birds. Nat Genet 14:62–68.
17. Just W, Rau W, Vogel W, Akhverdian M, Fredga K, et al. 1995. Absence of Sry in species of the vole *Ellobius*. Nat Genet 11:117–118.
18. Saccone G, Peluso I, Artiaco D, Giordano E, Bopp D, Polito LC. 1998. The *Ceratitis capitata* homologue of the *Drosophila* sex-determining gene Sex-lethal is structurally conserved, but not sex-specifically regulated. Development 125:1495–1500.
19. Meise M, Hilfiker-Kleiner D, Dubendorfer A, Brunner C, Nothinger R, Bopp D. 1998. Sex-lethal, the master sex-determining gene in Drosophila, is not sex specifically regulated in Musca domestica. Development 125:1487–1494.
20. Kacser H, Burns JA. 1973. The control of flux. Symp Soc Exp Biol 27:65–104.
21. Eanes WF. 1999. Analysis of selection on enzyme polymorphisms. Ann Rev Ecol Syst 30:301–326.
22. Verrelli BC, Eanes WF. 2001. Clinal variation for amino acid polymorphisms at the Pgm locus in Drosophila melanogaster. Genetics 157:1649–1663.
23. Oakeshott JG, Chambers GK, Gibson JB, Eanes WF, Wilcocks DA. 1983. Geographic variation in G6pd and Pgd allele frequencies in Drosophila melanogaster. Heredity 50:67–72.
24. Verrelli BC, Eanes WF. 2000. Extensive amino acid polymorphism at the Pgm locus is consistent with adaptive protein evolution in Drosophila melanogaster. Genetics 156:1737–1752.
25. Eanes WF, Kirchner M, Yoon J. 1993. Evidence for adaptive evolution of the G6pd gene in the Drosophila melanogaster and D. simulans lineages. Proc Natl Acad Sci USA 90:7475–7479.
26. Schulte PM, Gomez-Chiarri M, Powers DA. 1997. Structural and functional differences in the promoter and 5′ flanking region of Ldh-B within and between populations of the teleost Fundulus heteroclitus. Genetics 145:759–769.
27. Watt WB. 1977. Adaptation at specific loci. I. Natural selection on phosphoglucose isomerase of *Colias* butterflies: biochemical and population aspects. Genetics 87:177–194.
28. Whitt SR, Wilson LR, Tenaillon MI, Gaut BS, Buckler ES. 2002. Genetic diversity and selection in the maize starch pathway. Proc Natl Acad Sci USA 99:12959–12962.
29. Mendoza L, Thieffry D, Alvarez-Buylla ER. 1999. Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. Bioinformatics 15:593–606.
30. Johnson NA, Porter AH. 2001. Toward a new synthesis: population genetics and evolutionary developmental biology. Genetica 112–113: 45–58.
31. Jeong H, Mason SP, Barabasi AL, Oltavi ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.
32. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. 2001. The protein-protein interaction map of *Helicobacter pylori*. Nature 409:211–215.
33. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. The large-scale organization of metabolic networks. Nature 407:651–654.
34. Stuart JM, Segal E, Koller D, Kim SK. 2003. Gene-Coexpression network for global discovery of conserved genetic modules. Science 302:249–254.
35. Wagner A. 2003. How the global structure of protein interacion networks evolves. Proc R Soc Lond B 270:457–466.
36. Barabasi AL, Albert R. 1999. Emergence of scaling in random networks. Science 286:509–512.
37. Eisenberg E, Levanon EY. 2003. Preferential attachment in the protein network evolution. Phys Rev Lett 91:138701.
38. Qin H, Lu HH, Wu WB, Li WH. 2003. Evolution of the yeast protein interaction network. Proc Natl Acad Sci USA 100:12820–12824.
39. Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1.
40. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary Rate in the protein interaction network. Science 296:750–752.
41. Pal C, Papp B, Hurst LD. 2001. Genomic function (communication arising): Rate of evolution and gene dispensability. Nature 411:1046–1049.
42. Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68.
43. Wuchty S, Oltavi ZN, Barabasi AL. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35:176–179.
44. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple building blocks of complex networks. Science 298:824–827.
45. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298:799–804.
46. Wagner A. 2000. Robustness against mutations in genetic networks of yeast. Nat Genet 24:355–361.
47. Smith V, Chou KN, Lashkari D, Botstein D, Brown PO. 1997. Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science 274:2069–2074.
48. Kim SK. 2001. Functional genomics: The worm scores a knockout. Curr Biol 11:R85–R87.
49. Kitami T, Nadeau JH. 2002. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. Nat Genet 32:191–194.
50. Hansen TF, Wagner GP. 2001. Modelling genetic architecture: A multilinear theory of gene interaction. Theor Pop Biol 59:61–86.
51. Gibson G, Wagner GP. 2000. Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22:372–380.
52. Omholt SW, Plahte E, Oyehaug L, Xiang K. 2000. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. Genetics 155:969–980.