Title: Sequences and Structures of Adaptive Variants of *Colias* Phosphoglucose Isomerase.

Chris W. Wheat[1*†] , Ward B. Watt[1], David D. Pollock[1‡], and Patricia M. Schulte[1§]

[1]Department of Biological Sciences, 371 Serra Mall, Stanford University, Stanford, CA 94305-5020, USA, and Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA.

[*]To whom correspondence should be addressed. Work phone (Germany): 49-3641-571415, fax : 49-3641-571402. E-mail: cwheat@ice.mpg.de

[†]Present address: Department of Genetics and Evolution, Max-Planck Institute of Chemical Ecology, Beutenberg Campus, Winzerlaer Str. 10, 07745 Jena, Germany.

[‡]Present address: Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA.

[§]Present address: University of British Columbia, Department of Zoology, University of British Columbia, 6270 University Blvd., Vancouver, B.C. V6T 1Z4, Canada.

Running head: Sequences and structures of PGI variants in *Colias eurytheme*

Key Words: PGI, Colias eurytheme, C. meadii, balancing selection, enzyme evolution

.

Abstract

The enzyme phosphoglucose isomerase, PGI, of *Colias* butterflies (Lepidoptera, Pieridae) displays a widespread allozyme polymorphism. Many studies on the biochemical function, organismal performance, and fitness effects of *Colias* PGI genotypes have given evidence of strong natural selection in the wild to maintain this polymorphism. Here we begin to study the mechanism underlying this adaptive polymorphism at the level of molecular sequence and structure. The common electrophoretically-detectable alleles differ at multiple amino acid positions, and also show some cryptic charge-neutral amino acid variation hidden within the electrophoretic allele classes. Structural modeling shows that all changes are at or near PGI's surface. Several naturally abundant variants that distinguish these alleles are located in a peptide loop running across the monomer interface and connecting the two catalytic centers, where they can potentially alter subunit interaction and catalytic center geometry. There is a large excess of intraspecific variation, both synonymous and nonsynonymous, compared to interspecific fixation; there are no fixed synonymous differences between species, and only two fixed nonsynonymous differences. These fixed differences may be due to positive selection, contributing to the observed interspecies differences in PGI function. However, a sliding window analysis of intraspecific synonymous nucleotide diversity and Tajima's D shows that that the amino acid sites predicted to be foci of balancing selection, based on structural and functional considerations, also coincide with the regions of highest synonymous diversity and these fixed differences. *Colias*' PGI gene, with 1668 bp of cDNA, is divided into 12 exons, spread over ~ 11kb of chromosomal DNA, and intragenic recombination has been active over much of the gene. This case study of persistent polymorphism now offers the integration of the genomic and

molecular-structural bases of natural variation with its consequences for metabolic and organismal performance, thence for fitness, in wild populations.

Introduction

Studies of natural genetic variation at the molecular level often find patterns suggestive of the action of natural selection (Aquadro 2000, Kreitman 2000). In only a few cases have either the functional impacts or the ecological interactions of such variants been studied (Eanes 1999, Feder and Mitchell-Olds 2003, Gillespie 1991). However, by using genetic variants to probe organism-environment interactions, evolutionary biologists can disentangle potential causes of observed patterns and identify processes which maintain variation or lead to its change through time (Dean and Golding 1997, Fields and Somero 1998, Newcomb *et al*. 1997, Watt and Dean 2000). This requires either that molecular variation (e.g. Feder and Mitchell-Olds 2003, Verrelli and Eanes 2001) be studied at functional, organismal, and ecological levels, or that variants first identified at those levels (e.g. Hoekstra and Nachman 2003, Krebs and Feder 1997, Rawson and Burton 2002, Watt 2003) be pursued downward to their molecular bases.

The scales of space and time over which natural molecular variation occurs are of special interest here. Genetic polymorphisms have often been thought to be transient phenomena, displaced eventually by fixation due to selection or neutral drift. Some allozyme polymorphisms appear to support this view, being limited to single species and of apparently recent origin (Eanes 1999). Certain scenarios have been recognized as maintaining longer term variability, whether within populations or on a "global" species-wide basis: frequency-dependent selection in fungal or plant self-incompatibility systems (e.g. Ioerger *et al*. 1990, May et al 1999, Uyenoyama and Takebayashi 2004), or coevolutionary host-disease interactions (e.g Hughes and Nei 1992, Tian et al. 2002, Garrigan and Hedrick 2003). It has seldom been thought that within-population polymorphism affecting metabolism, development, or other aspects of adaptive performance might persist, e.g. through heterozygote advantage, over broad spans of space and time (but see

Gillespie 1991 and Watt 2004). Here, though, we report the study of molecular sequence and mechanism in such a persistent polymorphism, for which there is already evidence of strong selection at biochemical, organismal, and ecological levels in the wild, in multiple species complexes of the butterfly genus *Colias* (Watt 2004).

Four electrophoretically distinct (electromorph, EM) alleles of phosphoglucose isomerase (PGI, E.C. 5.3.1.9), called 2-5 according to their relative mobilities, are common in lowland *Colias* populations (Watt 1977). These alleles form 10 EM genotypes, whose frequency distributions are similar among populations from California to Colorado and across 30 years of study (Watt *et al.* 2003). Those studies' results, all of which were found in direct tests against neutral null hypotheses, can be summarized in terms of the recursive stages of natural selection (Feder and Watt 1992):

1) genotypes ⇒ phenotypes: single copies of each common EM allele were made identical-by-descent (IBD) in our laboratory stock, to preclude the confounding of functional study by mixtures of possibly undetected, electrophoretically cryptic variants (Watt 1977). Enzymes purified from these IBD-alleles' genotypes differ by several-fold in functional properties: 5 of the 6 heterozygotes have superior kinetics (i.e., high V_{max}/K_m ratios), while the homozygotes display a tradeoff of kinetics *vs*. thermal stability; in particular, the three most common genotypes are rank-ordered by V_{max}/K_m values as 3/4 > 3/3 >> 4/4 (Watt 1983).

2) phenotypes ⇒ performances: EM genotypes, differing in V_{max}/K_m, are predicted by metabolic organization theory to differ accordingly in metabolic performance, i.e. in ability to resupply metabolic "fuel"(ATP) to flight muscle *via* glycolysis and later pathways (Watt 1983, Watt and Dean 2000). The observed functional differences thus allowed us to predict genotype-specific differences in organismal performance, i.e., daily breadth of adult flight; these

predictions were sustained in replicated field tests, and in particular the three most common genotypes were rank-ordered exactly as predicted by their V_{max}/K_m values: 3/4 > 3/3 >> 4/4 (Watt *et al*. 1983).

3) performances ⇒ fitnesses: the dependence of all adult fitness components – survival, male mating success, and female fecundity – on flight performance leads directly to prediction of genotypic differences, based on the differences in biochemical function and thence in flight, in those fitness components. These predictions have been tested and sustained in diverse wild populations and taxa over many years, and recently a 10-fold range of relative fitness differences in the wild among the most common EM genotypes has been calculated: 3/3: 3/4 : 4/4 as 1.0 : 1.7 : 0.17 for males, 1.0 : 1.9 : 0.18 for females (Watt 2003).

4) fitnesses ⇒ genotypes: Population-genetic phenomena other than the predicted and observed natural selection – Wahlund effects, segregation distortion, or assortative mating –, which might have plausibly contributed to apparent PGI genotypic differences in fitness components, were all ruled out by direct tests (Watt 1983, Watt *et al*. 1985).

Until now, though, we have known nothing of an important level of phenotypic differentiation within the **genotypes ⇒ phenotypes** stage of the process: the molecular-structural mechanisms underlying the observed differences in function among *Colias*' PGI's EM genotypes. Evidently the IBD EM alleles whose genotypes displayed those functional differences were representative of their EM classes' properties, because all predictions of differences among the EM genotypes' mean flight performances and fitness component values, based on the functional differences, did prove accurate in repeated field tests. While charge-changing amino acids define the EM classes and are likely candidates for having functional effects, it is entirely possible that some parts of the observed genotypic functional differences

resulted from charge-neutral amino acid variation as well. Further, other charge-neutral amino acid changes with functional effect might segregate within the EM classes in the wild, thus contributing to variances found around the successfully predicted mean flight performances or fitnesses (Watt 1992).

Thus, our initial goals for this work were 1) to focus on the range of genetic variation at the PGI locus in a single population; 2) in so doing, assess the range of amino acid variation both among and within *Colias* PGI's EM allele classes; and 3) if possible, develop structurally based working hypotheses of the functional effects of those variants. We have done so by acquiring 19 cDNA sequences, including all four common EM allele classes, and by building a "homology model" (i.e. one based on overall homology with vertebrate PGI sequence) for *Colias* PGI's enzyme structure in which to study the location of *Colias*' polymorphic variants. We report initial results on the pattern of exons and introns in this gene, and on the possible role of intragenic recombination in altering the dynamics of this highly variable genetic system. Finally, we explore the relationship between variable amino acid positions, identified as candidate sites of selection on PGI by their structural placement and by the intermediacy of their frequencies in the wild, and regions in the DNA sequences which depart from neutral expectations in molecular tests of selection.

Materials and Methods

Animals and sequence sampling. All *Colias eurytheme* were taken from alfalfa fields near Tracy, California. Homozygotes of allele copies identical-by-descent (IBD) for each of the four EM allele classes were produced in our laboratory colony of *C. eurytheme*, thus allowing haplotype determination, comprising haplotypes 2A, 3A, 3B, 4A, 5A (Table 1). 14 other

sequences were obtained from *C. eurytheme* individuals, homozygous for the EM 3 or 4 alleles, chosen randomly with respect to other biochemical, genetic, or ecological aspects, from one field sample at Tracy on September 12, 2000.

These two samples, combined, yielded EM allele frequencies representative of wild frequency values. For example, the mean frequencies and standard deviation of EM alleles 2-5 of PGI from 8 population samples collected at Tracy, CA between 1980 and 1999 (average n = 141; Watt *et al*. 2003), were: $p_2 = 0.06 \pm 0.02$, $p_3 = 0.60 \pm 0.04$, $p_4 = 0.29 \pm 0.03$, $p_5 = 0.04 \pm 0.01$. From these frequencies, one would expect among 19 EM allele copies to find 1 2 allele, 11 3 alleles, 6 4 alleles, and 1 5 allele; we found 1, 12, 5, and 1 respectively. Since the sampling was random within each subclass (the actual variant DNA sequences were unknown to us *a priori*), and since the overall sampling within each allele class reflected natural frequencies, our aggregate sample should mimic the properties of a truly random sample of potential genetic variation at the molecular level. In particular, our sampling methodology should not affect the assumptions of the statistical tests we use.

For comparative purposes, alleles from each of the two major PGI EM classes of the closely related species *C. meadii* (sampled in the Snowy Range, Wyoming), were PCR-amplified and sequenced as per *C. eurytheme* below.

PGI cDNA sequencing. Total RNA was extracted from 30mg of abdomen tissue using RNeasy and QIAshredder kits (Qiagen). Gibco MMLV was used for reverse transcription per the usual protocol. Using degenerate primers based on Drosophila PGI sequence, partial (central) *Colias* cDNA PGI sequence was obtained by RT-PCR (Pollock 1995). The 3' end was obtained by standard RACE (rapid amplification of cDNA ends) protocol. The 5' end was obtained by three rounds of nested PCR following a modified RACE protocol. Terminal deoxynucleotidyl

transferase (TdT) tailing was used to add 15-20 G's to the 5' end of cDNA following manufacturer's protocol; extension time was 10 min at 37˚C followed by inactivation via 10 min at 65˚C (Gibco). 5 µl of the TdT reaction was directly used in the first of three rounds of PCR reactions using a poly C_{15} primer and a gene- specific primer, PGI-A-389 (5' ATA ACT TGC GTG GAG AAC TC 3'; primer nomenclature: S = sense direction, A = antisense, and number = nucleotide position of 3'-primer end relative to the start of the PGI gene). First round PCR product was purified (Qiagen QIAquick kit) and used as template for second round PCR with primer PGI-A-340 (5' TCA GGT GTG ACA TCT TTA CC 3'); this process was repeated for the third round with primer PGI-A-242 (5' TCA CCA GCA AAC ATA GCA TC 3'). Each step of the nested PCR increased specificity. These RT-PCR reactions were done using recombinant Taq polymerase (Gibco) at 1.5 mM $MgCl_2$ (50µl volume; cycle regime 94°C 2 min, then 35 cycles 94°C 30 s, 55°C 30 s, 72°C 1 min).

Initial sequencing found a conserved 30 bp region before the start codon, in which a 5'-end primer was designed. A 3'-end primer was designed, after sequencing by primer walking, in a conserved region beyond the stop codon. These were used to amplify PGI alleles from intact cDNA, using "High Fidelity Platinum Taq" polymerase (Invitrogen): with final concentration 2 mM Mg^{++} in 50 µl volume, with a cycle regime of: 94°C 2 min, then 35 cycles 94°C 30 sec, 55°C 30 sec, 72°C 3 min, with a final extension at 72°C for 10 min. Both strands of purified PCR products were cycle sequenced using Applied Biosystems BigDye 2.0 chemistry, and analyzed with a 377 sequencer. "Internal" primers were designed for use with the "end" primers in routine sequencing; all primers are listed in Table 1. All data were verified by sequencing in both forward and reverse directions.

PGI genomic sequencing. Genomic DNA was extracted from a 5th instar larva, IBD for an EM

4 allele (sequence 4-A, Table 1) using a DNAeasy kit (Qiagen). We amplified a full length genomic copy of PGI from this extract, again using Hi-Fi Platinum Taq with PCR extension times of 15 min/cycle, with a 25 min final extension. This amplicon was cloned into the XL Topo TA cloning vector (Invitrogen). Purified vector with ligated amplicon was then infected with a transposable element (TE) containing a new antibiotic resistance gene and retransformed (GeneJumper, Invitrogen). Resulting new colonies had the TE randomly inserted; 100 of these were screened by PCR mapping for TE insertion position. Those with insertions in PGI were purified and sequenced using forward and reverse primers located in the TE for full coverage.

Structural analysis of *Colias* PGI. Study of *Colias* PGI's structure is facilitated by the high conservation of PGI sequence among animals (Jeffery *et al.* 2001). For example, a 3-allele amino acid sequence from *Colias* (3-A of Table 1) shows 69% sequence identity and 83% Dayhoff similarity to both rabbit and human PGI sequence. We have thus used crystal structures for rabbit and human PGI (Jeffery *et al*. 2001, Arsenieva and Jeffery 2002, Read *et al*. 2001) to build a model for *Colias* PGI (Figure 2) using the program Swiss-Model (Schwede *et al*. 2003). Swiss-Model takes amino acid sequence as input, searches 3-D structural databases, retrieves candidates and aligns their sequences and structures, and based on these alignments calculates and returns a predicted protein structural model for the input amino acid sequence. Structural homologues identified and used as templates were from the RCSB Protein Data Bank (www.pdb.org; Berman *et al*. 2000): 1HOX.pdb (rabbit PGI bound with fructose 6 phosphate), 1G98.pbd (rabbit PGI bound with phosphoarabinonate), and 1IAT.pdb (human PGI bound with sulfate ion) (Jeffery *et al*. 2001, Read *et al*. 2001, Lee *et al*. 2001). There is a 9.01 Å RMS deviation between backbone atoms of the 1HOX.pdb structure and of the Swiss-Model results for *C. eurytheme*.

Solvent-accessible surface, SAS, is a measure of the fraction of each amino acid's surface exposed to surrounding solvent in the native enzyme structure. SAS was calculated with Molmol (Koradi *et al*. 1996) for the PGI structure, using a solvent radius of 1.4.

Statistical analyses. Sequence statistics and sequence-based tests of selection were performed using DnaSP 4.0 and Proseq (Rozas and Rozas 1999, Filatov 2002), unless otherwise noted. Averages of synonymous and nonsynonymous changes were calculated using MEGA2 (Kumar et al. 2001). Incidence of intragenic recombination (Watt 1972) was estimated among the IBD haplotypes (2A, 3A, 3B, 4A, 5A) using the conservative "four gamete" test (Hudson and Kaplan 1985) as implemented in DnaSP 4.0. Tajima's D (Tajima 1989) was calculated using synonymous site data. Significance levels for Tajima's D calculations were determined by 10,000 coalescent simulations in DNAsp.

Analysis of the juxtaposition of extreme values in Tajima's D across exons seven and nine was done as follows. For each exon, 10,000 coalescent simulations using Hudson's MS program (Hudson 2002) were run using the exon size, number of haplotypes sequenced, and number of total mutations as inputs (exon 7 = 126 bp, 19 haplotypes, 8 mutations; exon 9 = 162 bp, 19 haplotypes, 13 mutations). A sliding windows analysis of Tajima's D (window size = 70 bp) was then conducted on the simulations. Significance was determined by counting how many of the 10,000 simulations had windows with more extreme values than our observed ones.

Results

***C. eurytheme* allelic sequences.** A total of 5 IBD individuals were sequenced, along with nine wild caught EM homozygous individuals. Sequences covered the full coding region of 1668 bp, or 556 codons, of PGI cDNA (eight of 19 sequences lacked nine initial bases, while four lacked

the final 249 bases). There were 130 segregating sites, six of which had three variants, for a total of 136 changes; 119 were at synonymous sites (ss) and 17 were at nonsynonymous sites (nss). Theta ($\theta = 0.0224$) and average nucleotide diversity ($\pi_{total} = 0.0199$) were almost 3 fold higher than *Drosophila melanogaster* PGI and nearly five times higher than the average for *D. melanogaster* autosomal genes (37). *Colias* PGI's synonymous site diversity ($\pi_{ss} = 0.0733$) was 30 times higher than its nonsynonymous diversity ($\pi_{nss} = 0.0024$). The mean (± standard deviation) number of nonsynonymous changes among all haplotypes was 2.8 ± 0.85, with changes at five out of 17 nonsynonymous sites altering PGI's charge, always in ways consistent with EM allele mobility (Table 2). The 3 EM class has more synonymous variation than the 4 EM class (28.3 ± 3 *vs*. 22.9 ± 3.1, respectively), but has similar levels of nonsynonymous intra-EM variation (2.2 ± 0.7 *vs.* 2.2 ± 1.0). There are fewer nonsynonymous differences between the 3 EM and 4 EM classes (2.8 ± 1.0) than in all other possible comparisons between EM classes (average = 4.46, Table 3). Finally, of the two multiply sampled EM classes, 3 EM is more diverse than 4 EM ($\pi_3 = 0.021 \pm 0.002$, $\pi_4 = 0.015, \pm 0.004$).

Polarity of amino changes in this section and those that follow has been inferred by the frequency of variants and is designated with a unidirectional arrow, such that a unique amino acid variant, B, in a background of A, would be presented A→B. When polarity is unknown, this is indicated by a bidirectional arrow, A↔B.

One site, 375, has charge-changing variation among all four EM classes: the 2 allele has Gln there, the 5 allele has Glu, and alternative members of the 3 and 4 EM classes have either Asp or Glu. Common variants of the 4 allele (Table 2, 4-B – 4-E) show the charge change Arg→Cys at the nearby residue 369, though we found a unique allele, 4-A, which "mimics" this charge change with a 131 Ala→Glu change instead, retaining 369 Arg. The charge difference in

the 5 EM occurs because although it shares 369 Arg and 131 Ala with the 3 EM alleles, it has both negatively-charged Asp instead of neutral Asn at site 21 and neutral Met instead of positively-charged Lys at site 317.

There are also differences among charge-neutral amino acids in individual sequences at five sites (41, 50, 144, 152, and 458); most of these are variable among the 3 EM alleles, but the neutral variant at site 41 is unique to the 2 EM. An additional four sites (128, 450, 463, and 538) have neutral differences that were sampled more than once. Two of these, sites 463 and 450, are variable only within the 3 EMs, and site 450 is the only site with more than two neutral amino acid variants (Ala, Ser, and Val). The other two sites have neutral variation that crosses EM allele class boundaries: Val at site 538 occurs instead of Ile in one sequence each of the 3 and 4 EM allele classes, while site 128 has widespread segregation of Gly and Ala among different allele classes. Thus, while there is a considerable amount of neutral and charge-changing amino acid variation within and between EM class, several sites clearly distinguish among the EM classes, and their structural location will be discussed below.

***C. meadii* allelic sequences.** There were 35 synonymous and 3 nonsynonymous sites among the 38 sites that differed between single haplotypes of *C. meadii*'s 2 and 3 EM alleles (Figure 1). Theta and nucleotide diversity ($\theta = 0.022$, $\pi_{total} = 0.022$) were roughly similar to *C. eurytheme*. There were two charge-neutral amino acid changes, 67 Thr↔Ala and 235 Ala↔Ser, and one charge-changing amino acid variants, 51 Lys↔Asn. The two species differed by only two fixed nonsynonymous base changes, at positions 1 and 2 of codon 370, changing Gly in *C. meadii* to Ser in *C. eurytheme.* 370 Gly (codon GGG) not only appears to be fixed in a larger sample of *C. meadii*, but also may be ancestral among North American *Colias,* since the more basal *C. palaeno* is also fixed for the same codon at site 370 (based on 12 sampled *C. meadii* haplotypes

and 2 sampled *C. palaeno* haplotypes, C. W. Wheat, unpublished results). Remarkably, we find *no* fixed synonymous differences between species (Figure 1).

Structural context for variants' evaluation. The extensive polymorphism in *Colias* PGI, as in other taxa, is in contrast with the overall sequence conservation of this enzyme across diverse taxa (Jeffery et al. 2000), although all active site residues identified in vertebrates are conserved between them and *Colias*. Not surprisingly, three-dimensional structure also appears to be conserved: *Colias* PGI is predicted to be a dimer with 180º rotational symmetry, as are the known rabbit and human structures (Figure 2). Further, a remarkable feature observed in vertebrate PGIs extends to *Colias*: in the known and predicted native structures, each monomer provides a histidine residue (388 in rabbit, 392 in *Colias*), essential for catalytic function, to the other monomer's catalytic center. This interpenetrance of monomers may predispose PGI to the heterozygote functional advantage seen repeatedly among *Colias* species and other taxa (Gillespie 1991, Watt 2003). The interpenetrance may also explain the reluctance of different homozygous preparations of *Colias* PGI, when mixed, to exchange monomers during thermal cycling or other manipulations which can "reshuffle" subunits of other multimeric proteins (Schulte and Watt, unpublished).

Amino acid changes in an enzyme's catalytic center can sharply alter substrate specificity or reaction mechanism (Dean and Golding 1997, Newcomb *et al*. 1997). But often, as in lactate dehydrogenase (Fields and Somero 1998, Gerstein and Chothia 1991), quantitative variation of catalytic performance occurs *via* amino acid changes at the enzyme surface rather than in the catalytic center. Despite their distance from the catalytic center, such changes can nonetheless alter catalytic center flexibility and/or geometry, and hence enzyme kinetics (Watt and Dean 2000, Gerstein and Chothia 1991). Furthermore, the integrity of the monomer interface is

essential for PGI's catalysis (Bruch *et al*. 1976), and so amino acid variation at the monomer interface could affect catalysis by altering that interface. We thus expect amino acid changes underlying the observed functional and fitness differences among *Colias* PGI genotypes occur on the enzyme surface, at the monomer interface, *and* in regions affecting catalytic center geometry.

Spatial placement of PGI variant changes. All variable amino acid residues in *C. eurytheme* PGI are partially, and most are substantially, exposed to solvent, with none in the interior core (Figure 3 and Table 4). This observation is similar to those made with other enzymes (e.g., Watt and Dean 2000, Bustamante *et al*. 2000). In contrast, synonymous base changes, which do not change the encoded amino acid residues, show no relation to the solvent accessibility of the amino acids which they encode (Table 4).

Which of these multiple changes among EM allele classes account for the functional variation among genotypes? Consider first PGI's structure and reaction mechanism (cf. Figures 2-4). Starting in one monomer's catalytic center, the peptide chain runs from Glu 361 to the surface, where a loop turn near the monomer interface includes residues 369 and 375. The chain then turns back into the dimer, running across the monomer interface to His 392 in the other monomer's catalytic center. By analogy to mammalian PGI (Jeffery *et al*. 2001, Arsenieva and Jeffery 2002), in each catalytic center His 392 begins isomerization of glucose-6-phosphate, G6P, to fructose-6-phosphate, F6P (or *vice versa*) by protonating the hexose ring oxygen. This yields a *cis*-enediol transition state whose resolution to the aldo- (G6P) or the keto (F6P) isomer is facilitated by Glu 361 and by motion of the C-terminal domain [but see Read *et al*. (2001) for another view of the action of these same residues].

As discussed earlier, charge-changing amino acid variation among *Colias* PGI EM alleles occurs mainly at sites 369 and 375. EM 2 differs from the other EM allele classes 3, 4, and 5 in

having a Gln, rather than Asp or Glu, at 375, removing a negative charge at this position. EM 4 (aside from the one EM charge "mimic" allele, 4-A, see below) differs from the others in having 369 Cys rather than 369 Arg. The change at 369 removes a positive charge, some uncharged polar groups, and much bulk in the form of five backbone C and N atoms and their bound hydrogens. Sites 369 and 375 lie in a surface loop of the peptide chain (Figure 4) that **a)** connects the catalytically active Glu 361 in one active center and the catalytically active His 392 projected into the other active center, and **b)** crosses the monomers' interface. Variants here, changing charge and residue packing, and exerting differential tension on both active centers and the monomer interface, may contribute to both kinetic and stability differences among genotypes. *This 361-to-392 loop is thus a prime candidate for a central focus of functional effects of PGI variation,* and it is notable that the only fixed difference between *C. meadii* and *C. eurytheme* also lies within this loop.

As for other EM class distinctions, EM 5 has two changes with relatively little difference in bulk compared to the others: 21 Asn→Asp and 317 Lys→Met. These are on the outer monomer surface, where varying hydration due to their charge changes could alter protein solubility, overall shape, and thus tension on the active center. EM 2 also has a notable charge-neutral difference, an Ile at site 41 compared to an Asn in other classes, which is a relative decrease in sidechain polarity and an increase in bulk.

Within-class variants often have only small differences in sidechain bulk, e.g., 538 Ile→Val or 128 Ala→Gly. They may also alter function somewhat: e.g., 375 Asp↔Glu could do so by projecting the charged carboxyl group into nearby water relatively less (Asp) or more (Glu). The largest intra-class difference is the uncommon A allele of EM class 4. The functional effect of its 131 Ala→Glu change is of interest for understanding the contribution of such amino

acid variation to our previous measurements of PGI performance, as it essentially “mimics" the other EM 4 alleles, which receive their charge change from a different region of the enzyme at site 369 Arg→Cys.

As noted earlier, the fixed difference between *C. meadii* and *C. eurytheme* at site 370 (Gly in *C. meadii* and Ser in *C. eurytheme*) is close to the intraspecific polymorphic sites 369 and 375 and within the interconnecting surface loop. Along with this fixed difference, there is an overall increase of thermal stability of *C. eurytheme* genotypes compared to those of *C. meadii*, and a reversal in the order of the kinetics vs. stability tradeoff between the EM 2 and EM 3 homozygotes of these species (Watt *et al*. 1996). This underscores further the potential functional importance of natural variation in sites in the interconnecting loop.

Thus, our analysis of the placement of PGI genotypes in the predicted structure agrees with reasonable expectations. They now comprise working hypotheses for future testing in detailed structure-function studies.

Gene structure and intragenic recombination. The evolutionary dynamics that may have generated the observed variation in *Colias* PGI depend partly on the arrangement of the gene on its chromosome. By sequencing genomic DNA for the IBD EM 4A allele, we were able to determine that the 1668 bp PGI cDNA is divided among 12 exons, averaging 139 bp each, and separated by 11 introns that range between 296 and 2445 bp in length, adding up to a total translated region of over 10,000 bp (Figure 1). We also obtained initial evidence for some intron length polymorphism in the wild (C. W. Wheat, unpublished). Residues 369 and 375, candidate foci of functional differences among EM allele classes, are located in exon nine, which is about 8.1 kb and 2.2 kb from the start and stop codons respectively. Based on DNA sequence data from our IBD allele copies, there have been at least 11 intragenic recombination (IGR) events,

all upstream of exon 9 (Figure 1).

Molecular-level tests for natural selection.

The neutral theory of population genetics provides the proper null hypothesis against which patterns of genetic variation in the wild should be tested (e.g. Kreitman 1996). A variety of methods have been developed to do so, providing indirect lines of evidence for selection in some cases. We apply two of these to the *Colias* PGI system.

First, if neutrality holds, the proportions of synonymous and nonsynonymous base changes polymorphic within species *vs.* fixed between species should be similar (McDonald and Kreitman 1991). Setting aside large changes of population size, interspecies directional selection will cause an excess of fixed nonsynonymous differences, but purifying selection within species will lower nonsynonymous variation and balancing selection will yield high nonsynonymous and synonymous variation (McDonald and Kreitman 1991, Sawyer and Hartl 1992). Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts about 13 synonymous fixations alongside the 2 interspecies nonsynonymous fixations, but as noted earlier, *no* fixed synonymous sites are found. These data are significantly heterogeneous (Fisher's exact test, $P = 0.021$; Moriyama and Powell 1996).

The Tajima's D statistic can test for departure of variants from neutral expectations by examining synonymous variation from a population sample: low frequency polymorphic sites lead to negative values, indicating population expansion or directional selection, while intermediate variant frequencies lead to positive values, indicating population bottlenecks or balancing selection (Tajima 1989). Here we used a 70 bp (half the average exon length) sliding window for two separate analyses of synonymous sites, as a compromise between being large

enough to capture potential clouds of variation and yet small enough to detect variation in Tajima's D within an exon. First, we tested our site-specific *a priori* predictions of neutral variation accumulation. Since this was not a *post hoc* search for significance, identified peaks were not subject to a multiple test correction. Second, we examined the distribution of the values of Tajima's D observed across PGI, asking how often such distributions are observed in simulated datasets.

Sliding window analysis across *C. eurytheme* PGI, exon by exon, of Tajima's D finds two positive peaks, with the highest (+1.71, P = 0.04) in the 5' end of exon 9, and the other in the 3' end of exon 7 (+1.59, P = 0.04) (Figure 5, B). (Sliding windows of 35 and 140 bps for the entire cDNA identified the same peaks.) These correspond to the most probable candidates for amino-acid-site foci of balancing selection which could be associated with neutral hitchhiking: sites 369 and 375 in exon 9, and site 317 in exon 7. Site 369 contains the main difference between EM 4 and all other alleles, and site 375, which contains the main difference between EM 2 and other alleles, as well as the main pair of segregating variants shared by EM 3 and EM 4. Site 317 contains one of two unique differences between EM 5 and other alleles. Other possible foci for hitchhiking are sites 21, 41, or even 128, but these are remote from the subunit interface and the latter two, of a smaller degree of physicochemical change. In any case, we find no support from Tajima's D for these other candidate regions. In general agreement with these observations is a sliding window analysis of synonymous nucleotide diversity (π_{ss}), which shows a generally high level of diversity across the broad center of PGI cDNA, maximized over exon 7 (Figure 5, A).

The patterns of Tajima's D within exons 7 and 9 are also of interest. Notice how both these exons change in Tajima's D across their length. Beginning with exon 7, the 5' end has a

Tajima's D of -0.7259 while the 3'end is +1.59. In exon 9, the 5' end is +1.71 while the 3'end is extremely negative (-1.88, $P = 10^{-5}$ without correction). Such juxtapositions of extreme values within these exons occur rarely in coalescent simulations (exon 7 P = 0.02, exon 9 $P = 10^{-4}$, Figure 5), and in both cases the amino acid variants that are candidates for balancing selection are at the ends with highly positive Tajima's D. The gene otherwise shows generally negative values of D (excepting mildly positive values in exons 4 and 8), which are significant as whole exons in exon 1 (-1.861, P = 0.005) and 12 (-2.110, $P < 10^{-5}$).

Discussion

PGI genotypes' molecular structures as sources of their functional differences. Two important themes of genotypic functional difference are shared between the two *Colias* species complexes studied so far in this respect (Watt 1983, Watt *et al.* 1996):

1) widespread heterozygote advantage in kinetics, expressed in high values of V_{max}/K_m and largely in small values of K_m, representing strong substrate affinity;

2) in homozygotes, a consistent trade-off of kinetic performance *vs.* enzyme stability.

How can we begin to account for these in terms of our present structural findings?

The reciprocal interpenetration of monomers within the dimer, embodied in the peptide loop(s) connecting residues 361 and 392, ensures strong interaction of monomers at and across their interface. This interaction may also extend to the catalytic amino acid residues in the two catalytic centers. Amino acid changes in this loop (or elsewhere, e.g. changing surface hydration or residue packing) which can alter its geometry or that of the monomer interface, may easily affect dimer stability, the catalytic center's shape, or the orientations of catalytic functional groups. These structural features give greater specificity to our initial general expectations for location of functionally important amino acid changes (see above). The concentration of most

differences among *C. eurytheme*'s EM alleles into the 361–to-392 loop and monomer interface region, and secondarily into parts of the enzyme surface whose change may affect them, is quite striking.

These observations lead to an hypothesis easily testable by further molecular study of structure and function among *Colias*' PGI genotypes. Following Monod *et al*. (1965), though without any implication of allosteric phenomena for which there is no evidence here, we propose that the rotational symmetry of the PGI dimer, and the close connections of its monomers, may impose rigid constraints on conformations of homodimers whose monomers have, *ipso facto*, identical amino acid sequences. Such constraints may force homozygotes' PGI into discrete alternative conformation states, resulting in either greater stability or greater flexibility which can support tight substrate binding and high V_{max}/K_m ratios – producing the tradeoffs of kinetics *vs.* stability seen in homozygotes of both species complexes. In contrast, in the heterodimers found only in heterozygotes, alternation of key amino acids in the critical loop-interface region, e.g. 369 Arg/369 Cys in *C. eurytheme*'s 3/4 heterozygote, may break some features of these symmetry-based constraints. This could allow heterodimers to take on intermediate configurations, combining near-maximal stability with nonadditivity or actual heterosis in kinetics, as seen empirically in *C. eurytheme*'s 3/4 and *C. meadii*'s 2/3 genotypes.

The sole amino acid difference fixed between species, change at position 370 of Ser in *C. eurytheme* for Gly in *C. meadii* (and *C. palaeno*), is consistent in its location, and in its potential history, with this view. Fixation of the Ser codon, TCG, in the *C. eurytheme* lineage would have occurred most parsimoniously in two steps from ancestral Gly, GGG. Possible single-mutation intermediate stages (each a different transversion, G→T and G→C) are either Ala, GCG, or Trp, TGG. The former would be far less disruptive of structure in either the loop or the surrounding

monomer interface than the latter, and whichever change came first might still have been segregating with its ancestral alternative when the second change occurred. Continued study of the molecular phylogenetics which has already begun in this group (Pollock *et al*. 1998, Wheat and Watt unpubl.), combined with searching for transitional states of the polymorphism in intermediate relatives, may allow us to sort out these alternative possibilities.

Spectrum of effects of amino acid changes. *Colias*' PGI polymorphism now presents a remarkable opportunity to "dissect" natural variants in terms of enzymatic structure-function relationships. This may be feasible either by use of directed mutagenesis, as recently done for interspecies differences in fish LDH sequences (Johns and Somero 2004), or by taking advantage of natural recombinants, extractable from the wild, many of which differ pairwise by only single amino acids. For example, one could isolate in laboratory stock identical-by-descent, IBD, copies of alleles 3E and 4B (Table 2; both of these are frequent in the wild, unpubl.), then measure the functional properties of their homozygotes and their heterozygote with one another. This would evaluate in detail the functional impact of the 369 Arg→Cys change which is the predominant difference between the EM 4 allele class and the others, on a common amino acid sequence "background" including 128 Ala, 375 Asp. Other allele combinations could isolate the same change on other backgrounds, e.g. including 375 Glu, for comparison. The 3EM allele class, which is very diverse in minor within-class changes, offers many such chances to isolate characteristic changes distinguishing each of the other allele classes from one another and the common or "baseline" state of the 3 allele class.

This would apply equally to study of the newly identified level of amino-acid variation within EM classes. For example, comparing genotypes of alleles 4C and 4D in the same manner as above would evaluate the impact of the putatively "minor" change 375 Asp↔Glu on identical

"background". Our original studies of function among EM allele class genotypes (Watt 1983) used IBD representative alleles from each class, to avoid cryptic-allelic confounding as discussed earlier. Since DNA-level analysis was not then possible in our system, we could not identify the sequences of those alleles. Now, we have the potential to discover, by comparing functional parameter values, exactly which intra-class alleles were used in that work. This within-EM class variation may well explain some of the variance found around the mean differences of performance and fitness values, successfully predicted from our functional studies, among EM-class genotypes in the wild (e.g. Watt 1992). Crystallographic determination of "native" *Colias* PGI structures in the future will further illuminate these structure-function questions.

Core-to-surface gradient in constraint on amino acid change. There are two likely sources of the constraint against amino acid variation in enzyme cores. First, the integrity of a protein's hydrophobic core affects stability of its whole tertiary structure. This has led protein chemists to postulate restriction of the flexibility needed for enzyme catalysis to the outer portions of enzymes' structure (e.g. Gerstein and Chothia 1991). Second, a protein must self-assemble as it is made. This assembly has complex dynamics, and any one mature structure may be reachable by only a few folding pathways. Along these, the formation of secondary and tertiary substructures, which can then undergo mutual "docking", will depend on repeatable associations of core residues (Fersht *et al*. 1991).

The relaxation of these constraints, as one passes from protein interiors to protein surfaces, allows room for changes showing a wide range of functional and fitness differences among genotypes (e.g., Watt and Dean 2000). Our sequencing has apparently captured a broad subset of such changes, from large down to neutral, or nearly so, in functional effect. While selection on amino acid variation in enzymes or other proteins may often be "purifying" in

nature (Bustamante *et al*. 2000), it is also true that if only a protein's surface is free to vary, then that surface must be where adaptive, as well as neutral, variation is to be found (Fields and Somero 1998, Watt and Dean 2000).

PGI nucleotide diversity. PGI in *D. melanogaster,* which is unusual among plants and animals in *not* exhibiting PGI allozyme variation, displays apparently recent purifying selection, with 2 nonsynonymous and 2 synonymous polymorphisms, but 1 nonsynonymous and 23 synonymous changes fixed since its divergence from *D. simulans* ~3 million years ago (Moriyama and Powell 1996); *D. simulans*' PGI is more ordinary, with 2 nonsynonymous and 16 synonymous polymorphisms. *Colias*' PGI variation contrasts strikingly with both these patterns in its high levels of synonymous and nonsynonymous polymorphism and the absence of fixed interspecific synonymous changes, despite two fixed nonsynonymous changes.

Indeed, the nucleotide diversity segregating at the PGI locus is among the most variable loci yet studied in insects, and displays perhaps the highest nucleotide diversity observed to date from a well-sampled single population (Table 5). There exists one other intraspecific sampling study in Lepidoptera, on pheromone binding protein in *Ostrinia nubilalis*, which shows much more variation than the average gene in *D. melanogaster*, although synonymous levels are roughly equal to *D. simulans*' autosomal average (Table 5). More genes from more species are needed to assess whether insects with more chromosomes and higher levels of recombination (Lepidoptera average around 30 chromosomes) contain, on average, elevated levels of intrapopulation genetic variation.

In comparison to some other well studied cases of balancing selection, *Colias*' PGI variation seems much older than, e.g., the recently arisen ($\leqslant$ 10 kyr or about 400 generations) balanced polymorphisms, maintained by malarial interaction, in β-hemoglobin and glucose-6-

phosphate dehydrogenase of *Homo sapiens*, which lack neutral hitchhiking variation (Aquadro *et al*. 2001). While we cannot yet numerically estimate the age of the balanced polymorphism(s) in *Colias* PGI, all species, and indeed populations, of *Colias* studied to date exhibit multi-EM-allele PGI polymorphism (>20 species from North America and Eurasia, e.g. Watt 2003).

Structural organization and PGI's DNA-sequence variation. The extensive subdivision of the PGI gene by introns may have impacted its evolutionary dynamics, and may have recorded evidence of the history of those dynamics. For example, the negative Tajima's D values in the 3' half of exon 9, following an area of "hitchhiking" of neutral variants around candidate sites for the focus of balancing selection, may reflect other kinds of events in the downstream 2.2 kb of the gene before the stop codon. The number and size of PGI's introns will have facilitated the action of intragenic recombination, IGR, by expanding the coding region across its chromosomal locus by nearly a factor of 10. IGR may thus have contributed to the evolution of PGI exons by generating diverse new allelic combinations among segregating sites, as long ago proposed (Watt 1972, see above). Further study will clarify how recombination, intron length variation, and neighboring gene evolution interact with selection to shape variation in this chromosomal region.

Molecular tests of selection. The distribution of synonymous variation across the PGI locus is very interesting. Neutral variants should occur randomly along the length of genes, but selection may distort this; e.g., neutral variants closely linked to sites under balancing selection are held longer in populations than would be so under "pure drift" conditions, creating a “cloud” of neutral variation at intermediate frequency (e.g. Kaplan *et al*. 1989). However, neutral hitchhiking is constrained by local recombination (e.g. Asmussen and Clegg 1981). While the high levels of intragenic recombination found here should significantly restrict the extent of neutral hitchhiking to small chromosomal regions (Andolfatto and Nordborg 1998), and indeed

our Tajima's D analysis shows evidence of this around three candidate sites, the extent and amount of synonymous variation found in other regions is high. Also remarkable is the fixation of change at site 370 within *C. eurytheme* since it diverged from *C. meadii*. This occurs in the midst of high levels of synonymous variation still segregating in both species, with none fixed between them. Whether this reflects generally higher genetic variation in *Colias* populations compared to other taxa, perhaps due to population structure, history, or chromosomal dynamics, as well as some combination of these with specific selection on this gene, warrants further study.

This work, and that of Garrigan and Hedrick (2003) on major histocompatibility complex variants, offer important tests of the power of Tajima's D to detect the action of independently demonstrated natural selection within a diploid population. Neutral hitchhiking variation, on which this and other such tests depend, requires substantial time to accumulate (Garrigan and Hedrick 2003). Thus our detection of a selection signature with Tajima's D suggests that selection has acted as strongly on *Colias'* PGI in the past as it does now. In contrast, the cases of human hemoglobin and glucose-6-phosphate dehydrogenase polymorphism, in which fitness differences have also both been measured in the wild and pursued to their molecular origin, show no such signature, probably due to the recent origin of malarial selection pressure maintaining this variation (Harding et al. 1997, Tishkoff et al. 2001, Verrelli et al. 2002). Indeed, in outcrossing species with normal recombination, such as butterflies and humans, evolutionary history and intragenic recombination can potentially erode the footprint of balancing selection to non-detectable levels (Andolfatto and Nordborg 1998, Nordborg and Innan 2003). Thus failure to detect a selection signature with such tests is ambiguous. Furthermore, even when such indirect tests are significant, they do not inform us about either the current level, or the mechanism, of selection (Simonsen et al. 1995, Garrigan and Hedrick 2003).

In our present case, the Tajima's D and the McDonald-Kreitman tests identify the same focal region for selection identified by integrating previous performance and fitness measures with the present structural analysis. Rather than supporting the sole use of such indirect tests, our findings argue for greater emphasis on combining evidence from functional and fitness measurements with molecular analysis of the genes in question. Deeper population sampling across the full 10 kb of the PGI gene, as well as flanking gene regions, is now needed to explore both the effects of local recombination in maintaining hitchhiking variation and the performance of indirect methods in detecting selection or other events within such datasets.

Integrating multiple approaches to study of adaptive evolution. This work connects formal molecular-evolutionary study, emphasizing statistical variation patterns, with the mechanistic study of the processes shaping those patterns. We have emphasized the natural connection between these perspectives through the evolutionary recursion (above, Feder and Watt 1992). Each element of this recursion offers in turn new insights and predictive abilities, integrating genomic and protein structural variation with its biochemical and organismal performance, and measurable fitness consequences in well-studied populations.

The pervasiveness of PGI EM polymorphism in a broad range of prokaryote, plant, and animal taxa as noted earlier, offers many potential tests of the generality of our results. Using indirect molecular tests of selection, some other studies have also rejected neutral patterns of PGI nucleotide variation in favor of balanced polymorphism (Katz and Harrison 1997, Filatov and Charlesworth 1999). A correlation has recently been found in beetles among PGI EM alleles, their functional properties, and HSP 70 expression in what appears to be selection on the thermal stability of PGI and its impacts on chaperone maintenance of cytosol protein integrity (Dahlhoff and Rank 2000, Rank and Dahlhoff 2002). Each of these studies, in its own way, extends the

general question we have begun to address here: why does strong selection maintain PGI's persistent variability across such divergent species? Evolutionary insight into the common features of these cases, building on elementary ideas already in view (Feder and Mitchell-Olds 2003, Powers *et al*. 1991, Watt 2000), will illuminate this question further.

Acknowledgements

Parts of this work were submitted by C.W. Wheat and D. D. Pollock to Stanford University in partial fulfillment of requirements for their Ph.D. degrees. We thank S. Ramos-Onsins for his help in coalescence analysis and C. Boggs, M. Feder, R. Hudson, T. Mitchell-Olds, D. Petrov, R. Simoni, G. Somero, J. Stamberger, and C.Yanofsky for stimulating discussions and/or for helpful commentary on this manuscript. We have been partly supported by the Ford Motor Co. Fund through the Center for Evolutionary Studies at Stanford, by the U.S. Dept. of Energy (grant ER-61667 to W.B. Watt), by the U.S. National Science Foundation (graduate fellowship to C.W. Wheat and grant IBN 01-17754 to W.B. Watt), and the Max Planck Gesellschaft. Our findings are our own and represent no official policy of any government agency or corporate entity.

LITERATURE CITED

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

Aquadro CF (2000) Limits to knowledge in population genetics: the problems of inferences of selection and evolutionary history. Evol Biol 32: 135-149.

Aquadro CF, Dumont VB, Reed FA (2001) Genome-wide variation in the human and fruitfly: a comparison. Curr Op Genet Devel 11: 627-634.

Arsenieva D, Jeffery CJ (2002) Conformational changes in phosphoglucose isomerase induced by ligand binding. J Mol Biol 323: 77-84.

Asmussen MA, Clegg MT (1981) Dynamics of the linkage disequilibrium function under models of gene-frequency hitchhiking. Genetics 99: 337-356.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28, 235-242.

Bruch P, Schnackerz KD, Gracy W (1976) Matrix-bound phosphoglucose isomerase; Formation and properties of monomers and hybrids. Eur J Biochem 68: 153-158.

Bustamante CD, Townsend JP, Hartl DL (2000) Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica.* Mol Biol Evol 17: 301-308.

Dahlhoff EP, Rank NE (2000) Functional and physiological consequences of genetic variation at phosphoglucose isomerase: heat shock protein expression is related to enzyme genotype in a montane beetle. Proc Nat'l Acad Sci USA 97: 10056-10061.

Dean AM, Golding GB (1997) Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc Nat'l Acad Sci USA 94: 3104-3109.

Eanes WF (1999) Analysis of selection on enzyme polymorphisms. Ann Rev Ecol Syst 30: 301-

326.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nature Rev Genet 4: 649-655.

Feder ME, Watt WB (1992) Functional biology of adaptation. In: Berry RJ, Crawford TJ, Hewitt GM, editors. Genes in ecology. Blackwell Sci Pub, Oxford, UK. pp. 365-392.

Fersht AR, Bycroft M, Horovitz A, Kellis JT jr, Matouschek A, et al. (1991) Pathway and stability of protein folding. Phil Trans Roy Soc London B 332: 171-176.

Fields PA, Somero GN (1998) Hot spots in cold adaptation: localized increases in conformational flexibility in lactate dehydrogenase A_4 orthologs of Antarctic notothenoid fishes. Proc Nat Acad Sci USA 95: 11476-11481.

Filatov DA (2002) ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets. Molec Ecol Notes 2: 621-624.

Filatov DA, Charlesworth D (1999) DNA polymorphism, haplotype structure, and balancing selection in the Leavenworthia PgiC locus. Genetics 153, 1423-1434

Fincham JRS (1966) Genetic complementation. Benjamin, NY.

Garrigan D, Hedrick PW (2003) Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. Evolution 57:1707-1722.

Gerstein M, Chothia C (1991) Analysis of protein loop closure: two types of hinges produce one motion in lactate dehydrogenase. J Mol Biol 220: 133-149.

Gillespie JH (1991) The causes of molecular evolution. Oxford Univ Press, NY. 336 p.

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. American Journal of Human Genetics 60:772-789.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hoekstra HE, Nachman MW (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice. Molec Ecol 12: 1185-1194.

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.

Hughes AL, Nei M (1992) Maintenance of MHC polymorphism. Nature 355: 402-403.

Ioerger TR, Clark AG, Kao T-H (1990) Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc Nat'l Acad Sci USA 87: 9732-9735.

Jeffery CJ, Hardre R, Salmon L (2001) Crystal structure of rabbit phosphglucose isomerase complexed with 5-phospho-D-arabinoate identifies the role of Glu357 in catalyis. Biochemistry 40: 1560-1566.

Johns GC, Somero GN (2004) Evolutionary convergence in adaptation of proteins to temperature: A_4-lactate dehydrogenases of Pacific damselfishes (*Chromis* spp). Mol Biol Evol 21: 314-320.

Kaplan NL, Hudson RR, Langley CH (1989) The 'hitchhiking effect' revisited. Genetics 123: 887-899.

Katz LA, Harrison RG (1997) Balancing selection on electrophoretic variation of phosphoglucose isomerase in two species of field cricket: *Gryllus veletis* and *G. pennsylvanicus*. *Genetics* 147: 609-621.

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Molec Graphics 14: 51-55.

Krebs RA, Feder ME (1997) Natural variation in the expression of the heat-shock protein hsp70

in a population of *Drosophila melanogaster* and its correlation with tolerance of ecologically relevant thermal stress. Evolution 50: 173-179.

Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. Bioessays 18:678-683.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genom Hum Genet 1: 539-559.

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetic analysis software. *Bioinformatics* 17:1244-1245.

Lee JH, Chang KZ, Patel V, Jeffrey CJ (2001) Crystal structure of rabbit phosphoglucose isomerase complexed with its substrate fructose-6-phosphate. Biochemistry 40: 7799-7805.

May G, Shaw F, Badrane H, Vekemans X (1999) The signature of balancing selection: Fungal mating compatibility gene evolution. Proc Nat'l Acad Sci USA 96: 9172-9177.

Mcdonald JH, Kreitman M (1991) Adaptive protein evolution at the AdH locus in *Drosophila*. Nature 351: 652-654.

Monod J, Wyman J, Changeux J-P (1965) On the nature of allosteric transitions: a plausible model. J. Mol. Biol. 12: 88-118.

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA polymorphism in *Drosophila*. Mol Biol Evol 13: 261-277.

Newcomb RD, Campbell PM, Ollis DL, Cheah E, Russell RJ, et al. (1997) A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. Proc Nat'l Acad Sci USA 95: 11476-11481.

Nordborg M, Innan H (2003) The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics 163:1201-1213.

Pollock DD (1995) Molecular evolutionary dynamics and pierid butterflies. PhD thesis, Stanford University (University Microfilms, Ann Arbor, MI).

Pollock DD, Watt WB, Rashbrook VK, Iyengar EV (1998) Molecular phylogeny for *Colias* butterflies and their relatives (Lepidoptera, Pieridae). Ann. Ent. Soc. Am. 91: 524-531.

Powers DA, Lauerman T, Crawford D, Dimichele L (1991) Genetic mechanisms for adapting to a changing environment. Annu Rev Genet 25: 629-659.

Rank NE, Dahlhoff EP (2002) Allele frequency shifts in response to climate change and physiological consequences of allozyme variation in a montane insect. Evolution 56: 2278-2289.

Rawson PD, Burton RS (2002) Functional coadaptation between cytochrome *c* and cytochrome *c* oxidase within allopatric populations of a marine copepod. Proc Nat'l Acad Sci USA 99: 12955-12958.

Read J, Pearce J, LI X, Muirhead H, Chirgwin J, Davies C (2000) The crystal structure of human phosphoglucose isomerase at 1.6 A resolution: implications for catalytic mechanism, cytokine activity, and haemolytic anemia. J Mol Biol 309: 447-463.

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

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161-1175.

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31, 3381-3385.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

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

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M (2002) Signature of balancing selection in *Arabidopsis*. Proc Nat'l Acad Sci USA 99: 11525-11530.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, et al (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455-462.

Uyenoyama MK, Takebayashi N (2004) Genus-specific diversification of mating types. In: Singh RS, Uyenoyama MK, editors. The evolution of population biology. Cambridge Univ. Press, Cambridge, UK. pp. 254-271.

Verrelli BC, Eanes WF (2001) Clinal variation for amino acid polymorphisms at the Pgm locus in *Drosophila melanogaster*. Genetics 157: 1649-1663.

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, et al. (2002) Evidence for balancing selection from nucleotide sequence analysis of human G6PD. American Journal of Human Genetics 71:1112-1128.

Watt WB (1972) Intragenic recombination as a source of population genetic variability. Amer Nat 106: 737-753.

Watt WB (1977) Adaptation at specific loci. I. Natural selection on phosphoglucose isomerase of *Colias* butterflies: biochemical and population aspects. Genetics 87: 177-194.

Watt WB (1983) Adaptation at specific loci II. Demographic and biochemical elements in the maintenance of the *Colias* PGI polymorphism. Genetics 103: 691-724.

Watt WB (1992) Eggs, enzymes, and evolution - natural genetic variants change insect fecundity. Proc Nat'l Acad Sci USA 89: 10608-10612.

Watt WB (2000) Avoiding paradigm-based limits to knowledge of evolution. Evol Biol 32: 73-96.

Watt WB (2003) Mechanistic studies of butterfly adaptations. In: Boggs CL, Watt WB, Ehrlich PR, editors. Butterflies: ecology and evolution taking flight. Univ of Chicago Press, Chicago, IL. pp. 319-352.

Watt WB (2004) Adaptation, constraint, and neutrality: mechanistic case studies with butterflies and their general implications. In: Singh RS, Uyenoyama MK, editors. The evolution of population biology. Cambridge Univ. Press, Cambridge, UK. pp. 275-296.

Watt WB, Carter PA, Blower SM (1985) Adaptation at specific loci. IV. Differential mating success among glycolytic allozyme genotypes of *Colias* butterflies. Genetics 109: 157-175.

Watt WB, Cassin RC, Swan MS (1983) Adaptation at specific loci III. Field behavior and survivorship differences among *Colias* PGI genotypes are predictable from in vitro biochemistry. Genetics 103: 725-739.

Watt WB, Dean AM (2000) Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu Rev Genet 34: 593-622.

Watt WB, Donohue K, Carter PA (1996) Adaptation at specific loci. VI. Divergence vs. parallelism of polymorphic allozymes in molecular function and fitness-component effects among *Colias* species (Lepidoptera, Pieridae). Mol Biol Evol 13: 699-709.

Watt WB, Wheat CW, Meyer EH, Martin J-F (2003) Adaptation at specific loci. VII. Natural selection, dispersal, and the diversity of molecular-functional variation patterns among butterfly species complexes (*Colias*: Lepidoptera, Pieridae). Molec Ecol 12: 1265-1275.

Willett CS, Harrison RG (1999) Insights into genome differentiation: pheromone-binding protein variation and population history in the European corn borer (*Ostrinia nubilalis*). Genetics 153: 1743-1751.

Table 1. Primers for cDNA PCR amplification and sequencing. Primer characteristics: S = sense direction, A = antisense. All primers listed in 5'→3' direction. Degenerate base symbols: K = G/T , R = A/G, S = C/G , W = A/T ,Y = C/T.

PGI-S-5'end: CTG CTT CAA ATC ACG TAC G

PGI-S-408: GAG TTC TCC ACG CAA GTT AT

PGI-S-906: AAC TTT ATG GAC AAC CAC TT

PGI-S-1297: CAG ACT GAG GCK YTK ATG

PGI-A-3'end: SCT AYT GGT CTA WAA TTC

PGI-A-1157: TGA TGC ACG AGC TGG TAR AA

PGI-A-883: GGT TGT CCA TAA AGT TGG CG

PGI-A-389: ATA ACT TGC GTG GAG AAC TC

Table 2. Amino acid segregating sites in the PGI gene from a representative population sample of *Colias eurytheme* with two haplotypes from *C. meadii* for comparison. "EM class" = electromorph class. Net charge change and amino acid change are shown relative to the 2-A EM allele, with charge changing amino acids in bold (* = same as in allele 2-A). The bottom two rows are from *C. meadii* PGI; all others are from *C. eurytheme*.

EM	Allele	Net	21	41	50	51	67	110	128	131	144	152	235	317	369	370	375	450	458	463	538
2	A	0	Asn	Ile	Pro	Asn	Thr	Arg	Gly	Ala	Ala	Pro	Ala	Lys	Arg	Ser	Gln	Ala	Ala	Ala	Ile
3	A	-1	*	Asn	Leu	*	*	*	Ala	*	*	*	*	*	*	*	**Asp**	*	*	*	*
3	B	-1	*	Asn	*	*	*	*	Ala	*	*	*	*	*	*	*	**Asp**	Ser	*	*	*
3	C/D	-1	*	Asn	*	*	*	*	*	*	*	*	*	*	*	*	**Asp**	*	*	*	*
3	E/F	-1	*	Asn	*	*	*	*	Ala	*	*	*	*	*	*	*	**Asp**	*	*	*	*
3	G/H	-1	*	Asn	*	*	*	*	Gly / Ala	*	*	*	*	*	*	*	**Asp / Glu**	Ala / Val	*	*	*
3	I/J	-1	*	Asn	*	*	*	*	Ala	*	*	Pro / Ser	*	*	*	*	**Asp / Glu**	*	Ala / Val	Thr / Ala	*
3	K/L	-1	*	Asn	*	*	*	*	Ala	*	Ala / Val	*	*	*	*	*	**Asp**	*	*	Thr / Ala	Val
4	A	-2	*	Asn	*	*	*	*	Ala	**Glu**	*	*	*	*	*	*	**Glu**	*	*	*	Val
4	B	-2	*	Asn	*	*	*	*	Ala	*	*	*	*	*	**Cys**	*	**Asp**	*	*	*	*
4	C	-2	*	Asn	*	*	*	*	Ala	*	*	*	*	*	**Cys**	*	**Glu**	*	*	*	*
4	D	-2	*	Asn	*	*	*	*	Ala	*	*	*	*	*	**Cys**	*	**Asp**	*	*	*	*
4	E	-2	*	Asn	*	*	*	His	Ala	*	*	*	*	*	**Cys**	*	**Asp**	*	*	*	*
5	A	-3	**Asp**	Asn	*	*	*	*	*	*	*	*	*	**Met**	*	*	**Glu**	*	*	*	*
2	A	0	*	*	*	**Lys**	*	*	*	*	*	*	*	*	*	Gly	**Glu**	*	*	*	*
3	A	-1	*	*	*	*	Ala	*	*	*	*	*	Ser	*	*	Gly	**Glu**	*	*	*	*

Table 3. Number of synonymous and nonsynonymous differences among EM classes of *C. eurytheme* PGI. Standard errors determined by 500 bootstrap replicates using MEGA2 (Kumar *et al*. 2001).

EM allele comparison	Nonsynonymous differences	Synonymous differences
2 *vs*. 3	4.4 ± 2.1	35.8 ± 3.9
2 *vs*. 4	5.0 ± 2.2	41.4 ± 5.0
2 *vs*. 5	4.0 ± 1.9	34.0 ± 5.4
3 *vs*. 4	2.8 ± 1.0	30.7 ± 3.2
3 *vs*. 5	4.4 ± 1.8	30.6 ± 3.8
4 *vs*. 5	5.0 ± 2.1	29.2 ± 4.1

Table 4. Solvent exposure of amino acids at *Colias* PGI's codon positions. Distributions of amino acid change and nucleotide diversity π for non-synonymous site (*nss*) and synonymous site (*ss*) nucleotide changes among alleles of *Colias* PGI in relation to percent solvent accessibility surface (% SAS) of the structurally aligned amino acid sites (SAS calculated using Molmol; see supplemental material). Amino acid residues binned by rank SAS in 1/6 intervals. Counts of *nss* and *ss* changes were square-root-transformed and regressed on mean SAS for each of the 6 bins, with the following results: for *nss* changes, b = 0.053 , $F_{1,5}$ = 8.46 , P = 0.043; for *ss* changes, b = -0.007, $F_{1,5}$ = 3.09 , P = 0.153. Further, a contingency table test for heterogeneity of *nss* and *ss* changes in the low half and high half of the SAS range was highly significant (G_1 = 13.25 with Yates' correction , P < 0.001). Nonsynonymous changes are thus associated with high exposure to solvent water, while synonymous changes are not so associated.

Bin	Residues	SAS	*ss*	*nss*	π_{ss}	π_{nss}	π_{nss}/π_{ss}
1	92	0-0.5 %	21	0	0.071	0	0
2	92	0.5-2.7 %	21	0	0.065	0	0
3	92	2.7-7.1 %	21	1	0.094	0.001	0.015
4	92	7.2-15.5 %	20	0	0.081	0	0
5	92	15.6-29.3 %	17	9	0.06	0.006	0.104
6	92	29.3-60 %	19	7	0.075	0.006	0.081

Table 5. Comparison of nucleotide diversity among relevant taxa. Data presented for *Colias eurytheme* PGI and relevant loci from *Drosophila melanogaster* and *D. simulans*, as well as the European Corn Borer moth *Ostrinia nubilalis* (Moriyama and Powell 1996, Willett and Harrison 1999). All nucleotide diversity values (π) are multiplied by 1000.

Species	Gene	Haplotypes	π (total)	π (synonymous)
Colias eurytheme	Pgi	19	19.90	73.30
Drosophila melanogaster	Pgi	11	0.78	1.62
	Adh	15	8.11	28.51
	Est-6	13	7.21	22.05
	autosomal	127	4.43	13.40
Drosophila simulans	Pgi	6	3.95	15.53
	Adh	5	6.77	27.24
	Est-6	4	22.24	79.65
	autosomal	108	9.61	37.68
Ostrinia nubilalis	Pheromone binding protein	16	17.3	35.5

Figure 1. Genetic variation, cDNA, and genomic structure of *Colias* phosphoglucose isomerase. A) Start and stop positions, in cDNA and in genomic DNA, for exons of *C. eurytheme*'s PGI gene. B) Scaled diagram of the Colias PGI gene's genomic structure, showing exons (boxes) separated by introns (dark lines). Arrows mark the midpoints of detected intragenic recombination sites (see text). C) Segregating nucleotide variation in exons. *C. eurytheme* alleles are listed by their EM allele numbers followed by their haplotype designator letters; two *C. meadii* haplotypes are listed by their (independent) EM allele numbers and "Cme". Synonymous polymorphic positions are printed in ordinary type, while non-synonymous polymorphic positions are printed in bold italic type for maximum contrast, with identical bases to the *C. eurytheme* EM 2 allele represented as dots. Two base changes, at positions 1108 and 1109, which are fixed between *C. eurytheme* and *C. meadii* are printed in bold type. Asterisks above columns mark charge-changing nonsynonymous variants. Degenerate base symbols: K = G/T, M = A/C, R = A/G, S = C/G , W = A/T ,Y = C/T. The scaled nuclear PGI genetic structure was drawn using "Gene Structure Draw" by V. Veeramachaneni (http://warta.bio.psu. edu/cgi-bin/Tools/StrDraw.pl).

Figure 2. Ribbon diagram of *Colias* PGI enzyme "homology model". A top view of the native dimer structure is shown, with the two interwoven 56 kD monomers colored green and yellow. Bound substrate, fructose 6 phosphate (F6P), in each active site is displayed in spacefill CPK coloring. Notice His 392's reciprocal participation in the other monomer's active sites, highlighted by showing spacefilled His392 residues in their monomer colors (e.g., His392 from green monomer makes contact with F6P in the yellow monomer).

Figure 3. Placement of segregating amino acid sites in *C. eurytheme* PGI enzyme. A space-filling "homology model" shows monomers in green and yellow, with segregating amino acid

sites highlighted in white. The two enzyme images are rotated 90˚ on the y axis. All segregating amino acid sites within our sample are visible since they occur at the enzyme surface.

Figure 4. PGI monomer showing segregating sites in surface loop region connecting residues in both catalytic centers. View of monomer interface surface with only one monomer shown (green). The 31 amino acids connecting catalytic center residues Glu 361 and His 392 are shown in yellow and substrate within active center is spaced-fill-colored blue. Sites 369 and 375 lie in a surface loop of the peptide chain that **a)** connects the catalytically active Glu 361 in one active center and the catalytically active His 392 projected into the other active center, and **b)** crosses the monomers' interface region. The only fixed genetic difference, two nonsynonymous change within the codon 370, between *C. meadii* and *C. eurytheme* also lies within this loop region (not shown).

Figure 5. A sliding window analysis of genetic variation across the PGI gene in *Colias eurytheme*. All data presented derived from analysis of synonymous sites, with A) nucleotide diversity, π, and B) Tajima's D, across all exons of *C. eurytheme* PGI. Step size was 25 bp and window length was 70 bp, which is half the average exon length. A generally high level of diversity across the broad center of PGI cDNA is observed, which reaches a maximum over exon 7. Sliding window analysis of Tajima's D finds two positive peaks, with the highest (+1.71, P = 0.04) in the 5' end of exon 9, and the other in the 3' end of exon 7 (+1.59, P = 0.04) (Sliding windows of 35 and 140 bps for the entire cDNA identified the same peaks.) These correspond to the most probable candidates for amino-acid-site foci of balancing selection which could be associated with neutral hitchhiking: sites 369 and 375 in exon 9, and site 317 in exon 7. Notice how both these exons change in Tajima's D across their length. Beginning with exon 7, the 5'

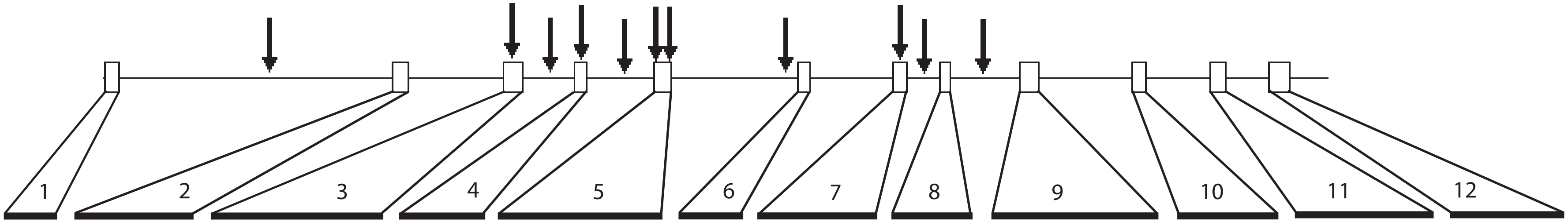
end has a Tajima's D of -0.7259 while the 3'end is +1.59. In exon 9, the 5' end is +1.71 while the 3'end is extremely negative (-1.88, $P = 10^{-5}$ without correction). Such juxtapositions of extreme values within these exons occur rarely in coalescent simulations (exon 7 P = 0.02, exon 9 $P = 10^{-4}$). The gene otherwise shows generally negative values of D (excepting mildly positive values in exons 4 and 8), which are significant as whole exons in exon 1 (-1.861, P = 0.005) and 12 (-2.110, $P < 10^{-5}$).

Figure 1. Genetic variation, cDNA, and Genomic Structure of *Colias* Phosphoglucose Isomerase

A)

		exon 1	exon 2	exon 3	exon 4	exon 5	exon 6	exon 7	exon 8	exon 9	exon 10	exon 11	exon 12	
cDNA	Start	1	130	274	442	553	715	829	955	1048	1210	1342	1483	
	stop	129	273	441	552	714	828	954	1047	1209	1341	1482	1671	
	5' UTR													3' UTR
genomic	1	35	2610	3597	4239	4944	6224	7079	7502	8214	9212	9918	10435	10624
	34	163	2753	3764	4349	5105	6337	7204	7594	8375	9343	10058	10623	10990

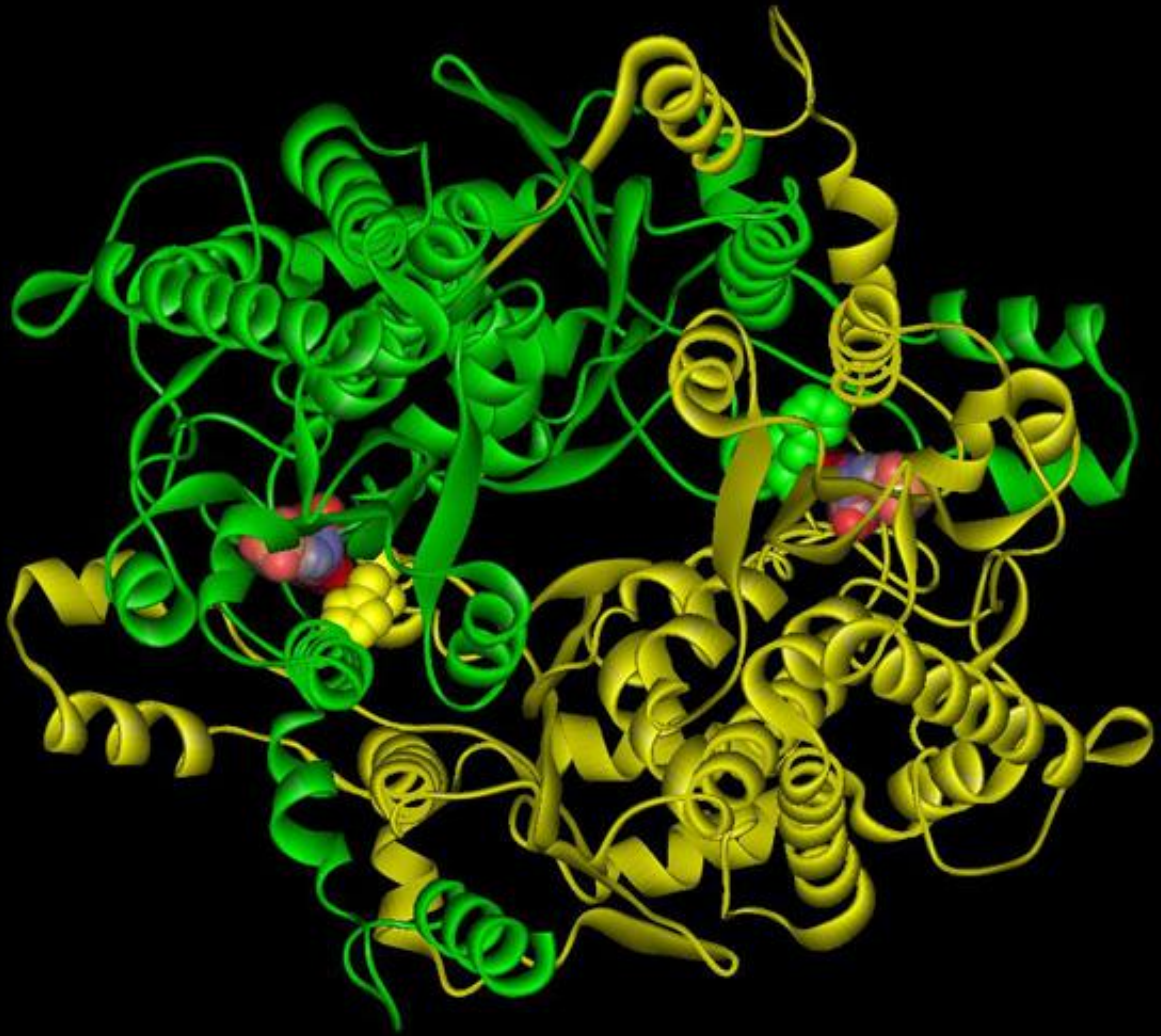
B)

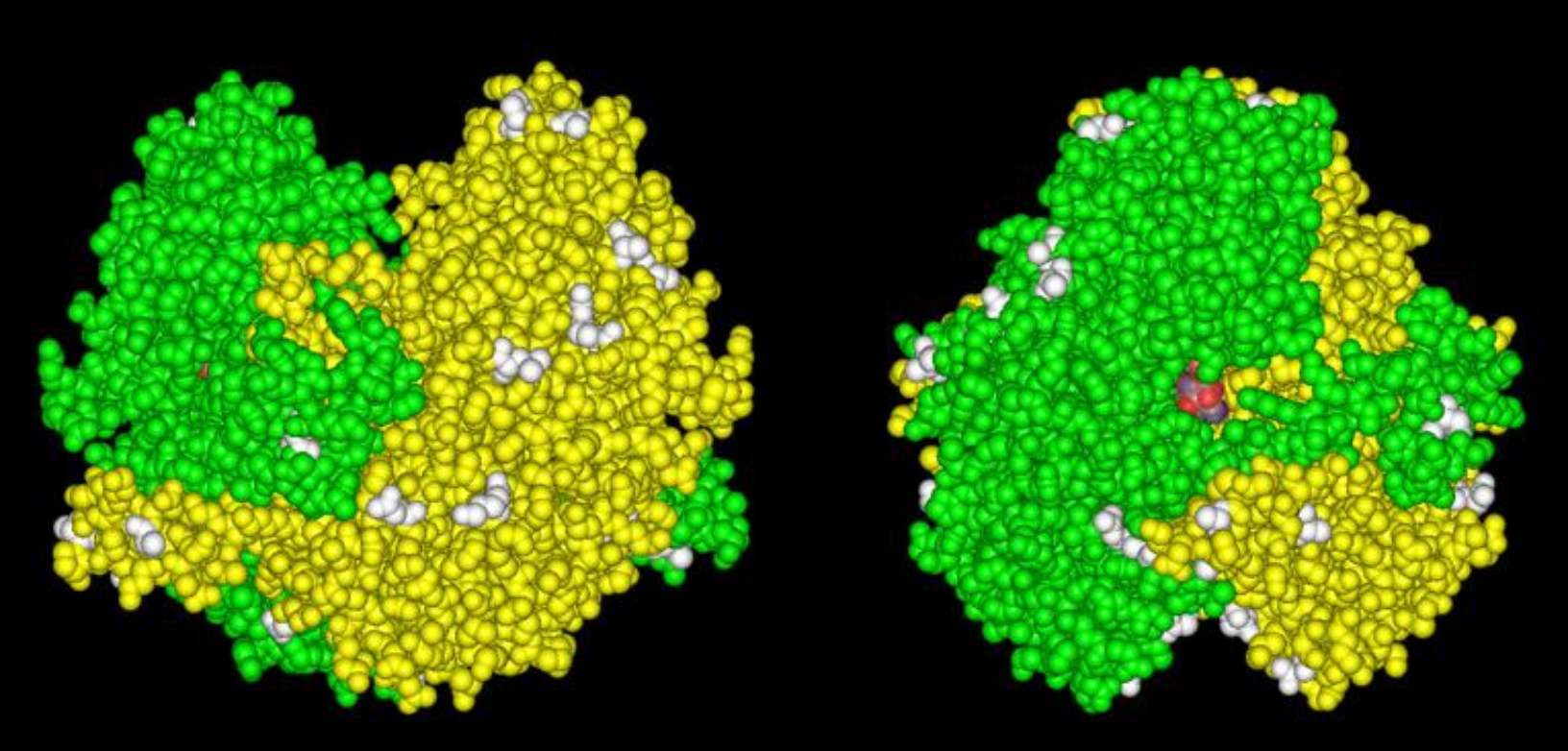


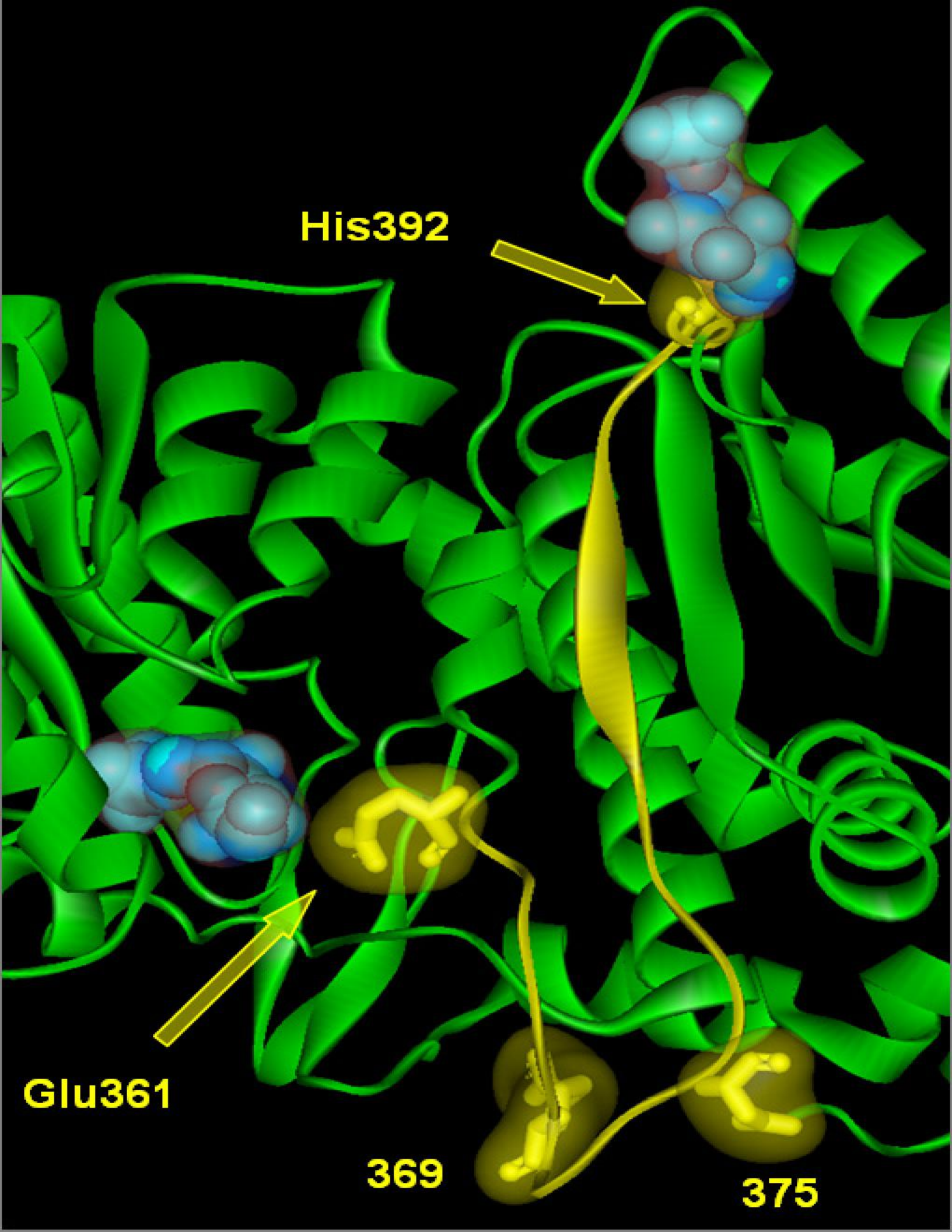
C)

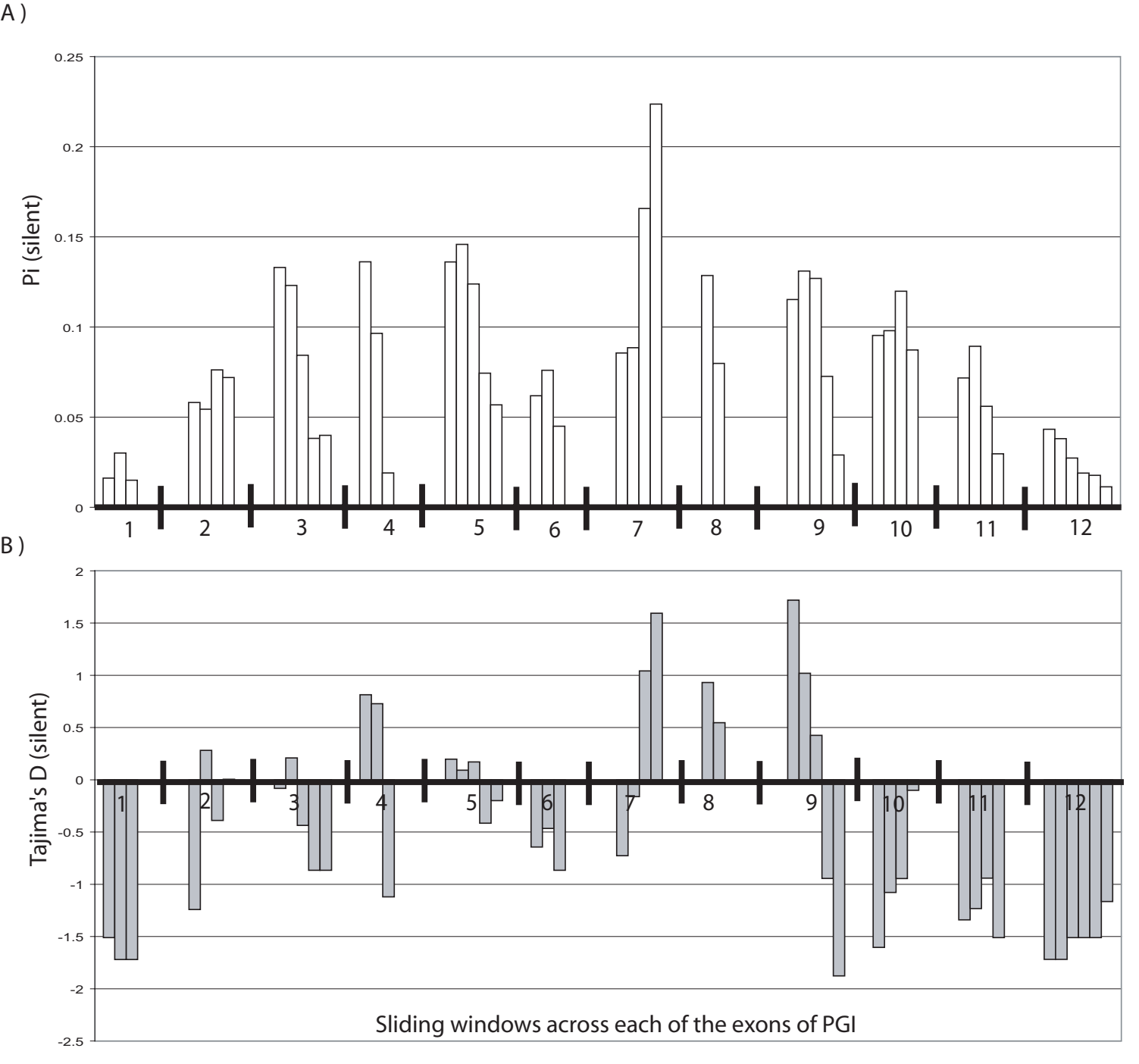
Segregating Sites by Exon

```
                                                                                                              1111 111111111111111111 11111111111 111111111111111 111111111111
                1 1111111112222 223333333333333444 444444455 555555566666666677 7777788 8888888899999 999990000 000111111111111122 22222222233 333333333344444 455555555666
           356992 3344558990123 890112234478899233 455578802 577888902233568901 2345600 3456678233345 567790024 579000002233466700 11355666902 444777888911245 803566668014
Alleles    611392 2819033897154 246281932521302912 403446975 806358434739167631 9849514 7324732103950 792595837 948145893547014306 58014069386 489347079836268 590103691221
             *        *                     *                                                       *                  *  **
2A         GAAACT TTTCATACATTAC CAGCCCGGAATCGCCACG CCACCTTGG CCTCACGAATCCCCGCGG ATGCGTT CCTTACGCCTGGA GTTGCCTAC GGTGTCTCCGGGCCCGGC CAACGGCGTCC CGCCTCAGTTCGCTG TGTACCGAGTAC
3A         .....A ...TG....CC.. ..........C..C.... .....A... .TC..GA.GC.GA....A ....ACC AG...T.T..... .A....A.. A..CC...GC........ .........TG .............CA CAC.........
3B         ...G.A ....G........ ..........C..C.... ......GCA .........GC.GA....A ..T.AC. A......T.CAT. TA.A..A.T ...CC...GC..A..... ........CTG .T..GA.......CA CAC.........
3C/D       .....A ....G.....CG. ........A.C....... ....TA... ..C..G..GC.GA....A ....ACC AG.....T..... TAKR..W.. R.C.....GC........ ........CTR .............CA CAC....R.Y..
3E/F       .....A ....G........ ..A..T.A.C..C.... ......... .TC..G..GC.GA....A ...TACC A...G..T..AT. TA.A..A.. R..C....GC........ .....C.T.TG G.....G......CA CAC.........
3G/H       K...YA ....G.R...C.. ...Y...R.C..S.... ..R.YWG.. YYC..S..RY.SA....R ...YRCC AG.....T..... TA.AY.W.Y ..YSY...GSC..Y.... .M.....CTG M.Y.KMR..K...-- ------------
3I/J       .R...A ....S.....Y.. M...YY...C.YCS.... Y.GYYAG.. .YY.RS..GC.GA....R .YKYACY AS..G.R..YRK. .W.R.SA.. R..S....GS..M..... .....K.RYTG ...YKM.RY..A.YA ------------
3K/L       .....A ....G.....C.Y .R..YY...C.YC...Y. .MR.YAGS. .YYY.S.RRY..A.R..A ..K.ACC ASY.......R.. KWK...W.Y ..YSY...GCC.M.SR.. Y.W.RK..YTM ....KM.RY..RYYA CACW......G.
4A         .....A .GC......CC.. ..A....A.C..C.A... T....A... T.....A.GC.GAT...A GC..ACC AG........... TA.A..A.T ...CC...G.C...T... ......A..T. .............CA CAC.......GT
4B/C       .....A .KY.G.....Y.. ..R..T.AMCWYC..M.. ..G.TAG.. .TC..GR.R...AT...A ....ACC A...G....CAT. KA.A.GW.. ...CYT..GS........ ..WY.....TR ............R.CA CAC.........
4D/E       .....A .K..G....YY.. ..R..TRACCW.C..M.. ..G.YAG.. .TC..GR.RY..AY...A ....ACC AS..G..Y.CAT. KA.A.GW.. ..YSYT..GC........ ..WY.....TG ............A.CA CAC.........
5A         ..G..A ........CC.. ........A.C....... ..G.TAG.. T.......GCTGAT...A ....ACC A..C...TTCA.T ......... ..C.....G..A...... .........TG .............CA CAC.........
2Cme       .....A CCC..G...CC.. ........A.C..C....T ..G..AG.. T.......GC.GA....A ..TT.CC .G..G....CAT. .A....... ...C..GGG.......... .....C..CTG G.............. .AC.....A...
3Cme       .....A ....G..TG.C.. .G.....A.C..C.... .....A... TT......GC.GAT.TT. ..TTA.. AG..G........ .A.....T. .C.C..GGG.C.....AT .....C...TG G.........T...A C.C.TAA.....
```






His392
Glu361
369
375


A)
Pi (silent)
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
10
11
12
B)
Tajima's D (silent)
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Sliding windows across each of the exons of PGI