# Fitness differences associated with *Pgi* SNP genotypes in the Glanville fritillary butterfly (*Melitaea cinxia*)

L. ORSINI,* C. W. WHEAT,*,† C. R. HAAG,* J. KVIST,*,‡ M. J. FRILANDER‡ & I. HANSKI*

*Metapopulation Research Group, Department of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland
†Department of Biology, Penn State University, University Park, PA, USA
‡Institute of Biotechnology, University of Helsinki, Helsinki, Finland

*Keywords:*

balancing selection;
deleterious allele;
individual fitness;
*Pgi*;
single-nucleotide polymorphism;
viability.

## Abstract

Allozyme variation at the phosphoglucose isomerase (PGI) locus in the Glanville fritillary butterfly (*Melitaea cinxia*) is associated with variation in flight metabolic rate, dispersal rate, fecundity and local population growth rate. To map allozyme to DNA variation and to survey putative functional variation in genomic DNA, we cloned the coding sequence of *Pgi* and identified nonsynonymous variable sites that determine the most common allozyme alleles. We show that these single-nucleotide polymorphisms (SNPs) exhibit significant excess of heterozygotes in field-collected population samples as well as in laboratory crosses. This is in contrast to previous results for the same species in which other allozymes and SNPs were in Hardy–Weinberg equilibrium or exhibited an excess of homozygotes. Our results suggest that viability selection favours *Pgi* heterozygotes. Although this is consistent with direct overdominance at *Pgi*, we cannot exclude the possibility that heterozygote advantage is caused by the presence of one or more deleterious alleles at linked loci.

## Introduction

The gene encoding the enzyme phosphoglucose isomerase (PGI) catalyses the second step in glycolysis. Nearly every species studied for PGI is polymorphic, and genetic variation in this locus has been shown to be correlated with variation in individual performance and fitness in a wide range of taxa (Zera, 1987; Zamer & Hoffmann, 1989; Patarnello & Battaglia, 1992; Katz & Harrison, 1997; Filatov & Charlesworth, 1999; Dahlhoff & Rank, 2000). In *Colias* butterflies, which are among the best-studied species in this respect, PGI variants differ in their biochemical performance, which leads to bioenergetic predictions and observation of differential flight performance and fitness among the PGI variants in the wild (Watt, 1977, 1983, 1992; Watt *et al.*, 1983). These fitness differences among the PGI variants in *Colias* are indicative of balancing selection, consistent with molecular signatures of historical selection (Wheat *et al.*, 2006).

More recently, PGI has been identified as a candidate locus affecting flight metabolic rate and fecundity in a large metapopulation of the Glanville fritillary butterfly (*Melitaea cinxia*) (Haag *et al.*, 2005; Saastamoinen, 2007). The Glanville fritillary occurs as a classic metapopulation in the Åland Islands in south-west Finland, where it has been the object of a large-scale study for 17 years (Hanski, 1999; Nieminen *et al.*, 2004). Local populations are mostly very small and have a high rate of population turnover. Spatial population structure restricts interactions among individuals; hence, demographic stochasticity (Hanski, 1999) and inbreeding depression (Saccheri *et al.*, 1998) are significant causes of population extinction (for a review see Hanski, 1998). Hanski & Saccheri (2006) demonstrated that specific PGI variants correlated with population growth rate in small local populations, challenging the perception that dissimilar performance of individual genotypes leading to fitness differences is irrelevant to population dynamics (Hanski & Saccheri, 2006).

The studies on PGI polymorphism in *Colias* butterflies and in the Glanville fritillary represent important contributions to our understanding of single-gene effects in natural populations, but allozyme electrophoretic data, on which the previous work has been based, have clear

*Correspondence:* Dr Luisa Orsini, Laboratory of Aquatic Ecology and Evolutionary Biology, Katholieke Universiteit Leuven, Ch. Deberiotstraat 32, 3000 Leuven, Belgium.
Tel.: +32 16323707; fax: +32 16320771;
e-mail: luisa.orsini@bio.kuleuven.be

limitations. Studies of allozyme allele frequencies do not contribute directly to a mechanistic understanding of the causes and consequences of molecular variation at the DNA level, pose allele identity problems when populations have a large number of allozyme alleles of similar electrophoretic mobility (as is the case with PGI in the Glanville fritillary) and preclude noninvasive high-throughput sampling necessary for large-scale studies.

Here, we cloned the full-length coding sequence of *Pgi* in the Glanville fritillary and identified nonsynonymous variants that determine the most common allozyme alleles. Targeting three of these sites, we designed single-nucleotide polymorphism (SNP) specific primers, which discriminate among the most common allozyme alleles, and assessed the reliability of these SNP markers in capturing functionally relevant variation by a reanalysis of previous phenotypic and allozyme data. We then used the SNP markers to study *Pgi* genotype frequencies in field-collected population samples as well as in laboratory crosses. Our results suggest that viability selection favouring heterozygotes is maintained by overdominance at *Pgi* or by the effects of loci linked with *Pgi*. A trade-off between negative survival effects of some genotypes balanced by positive effects on fecundity and flight capacity possibly plays a role in favouring heterozygotes.

## Materials and methods

### SNP design and validation

The RNA of one individual of the Glanville fritillary was initially used to clone and sequence the cDNA copy of the *Pgi* gene. The protocol followed is reported in Appendix S1.

To design the SNP panel, we sequenced *Pgi* cDNA from 33 adult butterflies. These butterflies originated from 21 local populations from the Åland Islands in Finland scattered across the entire metapopulation and they represent a subset of a larger sample used in a previous allozyme study (Haag *et al.*, 2005). Total RNA was extracted from adult butterfly tissue using Trizol™ reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. cDNA synthesis used 2 $\mu$g of total RNA using oligo(dT)$_{20}$ primer and Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. For the PCR amplification of the *Pgi* gene, high-fidelity Phusion DNA polymerase (Finnzymes, Espoo, Finland) and two primer sets (mPGI-16 and mPGI-17 or seqPGI-21F and mPGI-4, Appendix S1) were used. Additionally, internal primers were designed to obtain several reads by parallel sequencing of the same region with multiple PCR primer pairs (Appendix S1).

The PCR products were sequenced with ABI 3700 DNA Analyzer (Applied Biosystems, Foster City, CA, USA) using 50 ng of each PCR product either with the PCR primers or with internal primers. The PCR reactions were either column purified with QIAquick PCR Purification Kit (Qiagen Inc., Valencia, CA, USA) or gel purified with QIAquick Gel Extraction Kit (Qiagen Inc.) prior to sequencing. All sequences have been submitted to GenBank. Accession numbers are listed in Table S1.

Sequences of the coding region of *Pgi* were assembled and edited using DNA star software package (Lasergene; DNA STAR Inc., Madison, WI, USA). Sequence alignment was performed visually using the program Bioedit (Hall, 1999). Segregating amino acid variation was studied in pseudo-haplotypes generated from the unphased data by the program DNAsp (Rozas *et al.*, 2003). The phased pseudo-haplotypes were compared with the allozyme genotypes to infer the amino acid changes that give rise to the allozyme alleles. The inference was based on agreement between the electrophoretic mobility of the allozyme alleles (Haag *et al.*, 2005), the amino acid sequences of the pseudo-haplotypes and their predicted *in silico* charge calculated from the translated DNA sequences with the program Protean (Lasergene; DNA STAR Inc.). The PGI protein alleles (enzyme commission number EC 5.3.1.9) were determined by cellulose acetate electrophoresis as reported in (Haag *et al.*, 2005).

The above analysis identified four charge-changing amino acid polymorphisms, which give rise to the previously studied allozyme alleles in the Finnish metapopulation (see Results, below). We designed SNP-specific primers targeting the nucleotide positions corresponding to or in tight linkage with three of these polymorphic sites. The fourth site was not considered because of extensive synonymous polymorphism in the DNA region, which hampered the design of the SNP-specific primer. Thus, the amino acid site targets of SNP markers were AA35, AA111 and AA361 (the amino acid positions are also used to name the SNPs). The first two sites correspond to charge-changing amino acid polymorphisms, whereas the third site is in tight linkage with the charge-changing amino acid site AA372, as demonstrated by a complete analysis of linkage disequilibrium across the *Pgi* gene in the Glanville fritillary butterfly (C.W. Wheat, C.R. Haag, H. Frilander & I. Hanski, unpublished data). Extensive synonymous polymorphism around the site AA372 hampered the design of a SNP-specific primer. The combination of the three SNPs AA35, AA111 and AA361 discriminate among the most common allozyme alleles (see below). The primer sequences designed to target the three SNPs markers are given in Table 1.

Genomic DNA from several data sets was genotyped for the three SNPs. DNA was extracted either from wing clips or larvae using Nucleo spin tissue extraction kit (Mackerey-Nagel, Düren, Germany) with overnight incubation at 56 °C. Alternatively, previously extracted DNA was used (Hanski & Saccheri, 2006). Each PCR reaction (20 $\mu$L final volume) contained 20–30 ng of

**Table 1** PCR primers, annealing temperatures and SNP-specific primers for the three SNPs, AA35, AA111 and AA361.

| PCR region | Primer sequences for genomic amplification | Length (bp) | $T_m$ (°c) | SNP-specific primers | SNP-primer length (bp) | Nucleotide position (bp) | SNP type |
|---|---|---|---|---|---|---|---|
| PCR1 | 16F-CCGTGTACTCGAAAACTTTATTTG<br>209R-GACAAGGCAGCAGCATCTA | 1200 | 65 | AA35-ATATTCYACAACTATTTCAACA | 22 | 105 | A/T |
| PCR2 | 209F-TAGATGCTGCTGCCTTGTC<br>340R-CTTGCCGTTGACTAAGATAGG | 550 | 68 | AA111-CACATAGCACTTCGTAATAGA | 21 | 331 | A/C |
| PCR3 | PGI32-YTTTCTAYTAGATTCGCAGCGT<br>PGI35-GTACCTGGTCCCCTGGTGTA | 200 | 56 | AA361-GTGACGTATTTACCGTTGCT | 20 | 1083 | A/G |

Specific primer length, SNP type and the location on the specific genomic region are also indicated.

genomic DNA, 1 $\mu$м each primer, 200 $\mu$м each dNTPs, 2.5 mм $MgCl_2$, 20 ng BSA and 0.2 U Taq DNA Polymerase (Fermentas Life Sciences, Helsinki, Finland). An initial denaturing step (5 min at 95 °C) was followed by 35–45 cycles of amplification with 1 min at 94 °C, 1 min at the specific annealing temperature and 1.5 min at 72 °C. A final extension step included incubation for 15 min at 72 °C. Occasionally, the amplification of PCR1 (Table 1) was problematic, probably due to poor quality of genomic DNA or to nucleotide variation in the 3′-primer end site. In these cases, a touch-down PCR or less stringent PCR conditions (e.g. lower annealing temperature) were used. PCR products were purified with Exo-SAP-IT (GE Healthcare, GmbH, Germany) at the concentration of 1 $\mu$L/10 $\mu$L PCR reaction. Primer extension reactions employed the SNuPe kit (GE Healthcare, GmbH; Batley & Hayes, 2003), following the manufacturer's instructions. Primer extension reactions were run on a Megabace 1000 (GE Healthcare, GmbH) and the genotypes were called by SNP profiler (GE Healthcare, GmbH).

We assessed the reliability of these SNP markers by reanalysing two data sets in which there is correlation between PGI allozyme variation with flight metabolism, fecundity and population growth. In the data set of Haag *et al.* (2005) ($N = 25$), we tested for correlation between SNP genotype and flight metabolic rate at the individual level. In the data set of Hanski & Saccheri (2006) ($N = 261$ populations), we studied the most isolated populations ($N = 43$) to test for correlation between SNP genotype and local population growth. In the allozyme study, there was significant correlation between PGI allelic composition and local population growth rate (Hanski & Saccheri, 2006).

### SNP genotype frequencies in population samples and in laboratory crosses

We assessed SNP genotype frequencies and tested for Hardy-Weinberg (HW) equilibrium in three population samples from the field and in laboratory crosses:

**1** A random sample of adult butterflies ($N = 335$) from a network of 55 local populations within an area of $5 \times 4$ $km^2$ in the Åland Islands in Finland, collected in 2006 (A. Mattila, unpublished data).

**2** A random sample of larvae ($N = 224$) from 23 local populations in the same network of local populations as in sample 1, collected in 2006 (A. Mattila, unpublished data). This sample includes from one to four larvae per larval group (mostly full-sibs) depending on the size of the larval groups.

**3** Progeny of seven crosses between individuals from the Åland Islands ($N = 62$). All parents were heterozygous at SNP AA111. Under autosomal inheritance and in the absence of selection, we expected to observe Mendelian genotype frequencies (25% of each homozygote and 50% heterozygotes) among the offspring.

The analysis of larvae and adults allowed us to assess which life stages were affected by deviations from the HW proportions. The progeny from known parental genotypes was used to exclude nonrandom mating as a possible explanation for deviations from the HW proportions.

## Results

### Sequence analysis

The coding region of *Pgi* in the Glanville fritillary is 1671 bp (557 codons) long, which is one codon longer than the 3′-terminal region of the coding region in *Colias euritheme* (Wheat *et al.*, 2006). The intron–exon boundaries in the Glanville fritillary are identical to those in *C. eurithem*e and *Bombyx mori*, and the per cent sequence identities for the coding region compared with GenBank blast hits of *C. euritheme, B. mori, Drosophila melanogaster* and *Anopheles gambiae* are 78%, 76%, 71% and 71%, respectively, for nucleotide sequence, and 89%, 88%, 76% and 74%, respectively, for amino acid sequences.

We obtained full-length sequences of the *Pgi* coding region for most individuals ($N = 29$) that were sequenced ($N = 33$). A few sequences lacked from 10 to 35 bp at the 5′-end (marked with an asterisk in Table S1). Overall, there were 46 synonymous and 13 nonsynonymous polymorphisms. Here, we focus on nonsynonymous variation because of its potential functional significance. A complete summary of both synonymous and nonsynonymous variation is presented in Fig. S1.

## Mapping of allozyme alleles to sequences

Differences in the electrophoretic mobility of allozyme alleles on cellulose acetate gels are primarily due to charge variation, with minor effects of structural modifications affecting mobility (Johnson, 1977; Barbadilla *et al.*, 1996). Table 2 summarizes the amino acid variation in the entire data set.

The total number of distinct pseudo-haplotypes inferred from the unphased data was 16. The predicted net charge was used for the inference of the pseudo-haplotypes (Table S2). Of the 66 individual pseudo-haplotypes, 40 had the same predicted net charge of 1.91 (Table 2), although they belong to two different charge classes. These classes consistently differ at six amino acid sites, including three charge-changing sites. The net charge of the two groups is the same because the different charge changes cancel each other out. Thirty-six of the 40 pseudo-haplotypes with predicted charge 1.91 fitted the interpretation that the rarer one ($N = 15$) represents allozyme allele F and the more common one ($N = 25$) represents allozyme allele D (Haag *et al.*, 2005; Hanski & Saccheri, 2006). The remaining mismatches are probably explained by allozyme genotyping errors, given the very similar electrophoretic mobilities of the D and F alleles.

The predicted net charge allowed us to classify the remaining individual pseudo-haplotypes with respect to the allozyme alleles (Table S2), with one exception. Nine pseudo-haplotypes had a unique profile at the charge-changing amino acid sites, different from those of alleles D and F but scored by allozyme electrophoresis as either D or F with approximately equal frequency (Table S1). The predicted net charge of this new allele, called O (Tables 2 and S2), was different from D and F, yet our data suggest that its electrophoretic mobility is intermediate between these two alleles.

Overall, excluding the nine allele O haplotypes, 46 of the 57 (81%) pseudo-haplotypes yielded a perfect match between the allozyme genotype and the amino acid sequence. The remaining cases where mismatches were observed included alleles with very similar electromobilites and were likely to be due to allozyme genotyping errors.

## Comparing SNP genotypes and sequences

A perfect match between the three SNPs and the respective sites in the cDNA sequence was observed in all but three cases (Table S1). In these cases, the mismatch concerned only one of the three SNPs (AA361), and may be due to low quality of the SNP genotyping reactions (samples PP44_04 and SS22_04, Table S1). Only in one case (sample 197_06), the SNP genotype clearly indicated that codon AA361 was heterozygous, whereas it was homozygous according to the sequence.

In summary, across the three SNPs for 33 individuals, there was a disagreement in three out of possible 3 × 33 = 99 cases between the SNP genotype and the sequencing result, giving a compound error rate of 3% (homozygous and heterozygous sites). This good correspondence validates the SNP markers and reinforces the

**Table 2** Segregating amino acid variation in the Finnish metapopulation of the Glanville fritillary butterfly inferred on the pseudo-haplotypes generated from the unphased data.

| Pseudo- | | Amino acid sites | | | | | | | | | | | | | Charge | Observed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| haplotype | *N* | 30 | 35 | 49 | 64 | 111 | 155 | 173 | 174 | 241 | 308 | 343 | 372 | 375 | at pH 7 | mobility |
| E-1 | 2 | P | H | M | I | K | D | E | A | A | N | P | H | Q | 3.08 | < 0 |
| E-2 | 1 | • | • | • | • | • | • | • | G | • | • | Q | • | • | 3.08 | < 0 |
| A-1 | 4 | L | Q | T | V | • | • | • | • | S | • | • | • | • | 2.91 | 0 |
| A-2 | 1 | L | Q | T | V | • | • | • | • | S | S | • | • | • | 2.91 | 0 |
| B | 1 | L | Q | T | V | • | • | • | • | S | • | • | D | R | 2.75 | 5 |
| C | 3 | L | Q | T | V | • | • | • | • | S | • | Q | • | • | 2.91 | 12 |
| D-1 | 19 | • | • | • | • | • | • | • | • | • | • | • | D | • | 1.91 | 100 |
| D-2 | 4 | • | • | • | • | • | • | • | • | S | • | • | D | • | 1.91 | 100 |
| D-3 | 1 | • | • | • | • | • | E | G | G | • | • | • | D | • | 1.91 | 100 |
| D-4 | 1 | • | • | • | • | • | • | • | G | • | • | • | D | • | 1.91 | 100 |
| F-1 | 15 | • | Q | T | V | Q | • | • | • | S | S | • | • | • | 1.91 | 112 |
| H | 2 | • | Q | T | V | Q | • | • | • | S | • | • | D | R | 1.75 | 125 |
| O-1 | 6 | L | Q | T | V | • | • | • | • | S | • | • | D | • | 1.75 | ? |
| O-2 | 3 | L | Q | T | V | • | • | • | • | • | • | • | D | • | 1.75 | ? |
| G-1 | 2 | • | Q | T | V | Q | • | • | • | S | S | • | D | • | 0.75 | 210 |
| G-2 | 1 | • | Q | T | V | Q | • | • | • | • | • | • | D | • | 0.75 | 210 |

Segregating amino acid variants are listed by their codon position and shaded for positive (grey) or negative (black) charge. Residues identical to the first haplotype are coded by a dot. The first letter of the pseudo-haplotype in the first column indicates the inferred allozyme allele. *N* indicates the frequency of the pseudo-haplotypes in the data set. Observed electrophoretic mobilities of the different allozyme alleles are from Haag *et al.* (2005). Allozyme allele O was only detected by sequencing (see text), and thus its electrophoretic mobility is unknown.

**Table 3** Mapping of the three polymorphic sites target of SNP markers design to the allozyme alleles.

| Allozyme allele | SNP allele | AA35 | AA111 | AA361 |
|---|---|---|---|---|
| A, C | *a, c* | CA**A** | **A**AA | **A**AG |
| D | *d* | CA**T** | **A**AA | **G**AG |
| E | *e* | CA**T** | **A**AA | **A**AG |
| F | *f* | CA**A** | **C**AG | **A**AG |
| B, O | *b, o* | CA**A** | **A**AA | **G**AG |
| H, G | *h, g* | CA**A** | **C**AG | **G**AG |

The haplotypes obtained combining the SNPs distinguished among the most common allozyme alleles. The nucleotide sequence of each targeted codon is shown. The nucleotide site containing the polymorphism and target of the SNP genotyping is shown in bold.

interpretation that most mismatches between allozyme genotypes and sequence data were due to erroneous interpretation of the allozyme genotypes. Finally, we compared the multilocus SNP genotypes with the inferred allozyme alleles. The three SNPs can identify most of the allozyme alleles. The mapping of allozyme alleles to the three SNPs is given in Table 3.

### Correlation between SNP genotypes and phenotypic variation

Using allozyme data, Hanski & Saccheri (2006) showed that the frequency of the allozyme allele F explains a large and significant proportion of the variation in the growth rate of local populations. We repeated the association analysis between molecular markers and phenotypic traits using the SNPs designed in the present study. Both the combination of AA111 and AA361, which identifies the allozyme allele F (Table 3), and AA111 alone, explain variation in population growth rate essentially as well as the allozyme allele F (Table 4, Fig. 1a).

Haag *et al.* (2005) showed that flight metabolic rate is significantly associated with the allozyme allele F. Flight metabolic rate is explained equally well by AA111 and AA361 separately and even better by the combination of the two SNPs (Table 4, Fig. 1b). The combination of the two SNPs did not give exactly the same result as the F allele because three of the 25 individuals were scored differently by the two methods.

It is important to note that many individuals scored as FF homozygotes in the allozyme studies appear to have been heterozygotes between F and some other allele, usually D, which is the most common allele. The allozyme FF homozygotes are often difficult to distinguish from FD heterozygotes, and possibly from some genotypes including allele O (see above), because of their nearly identical electrophoretic mobilities (Haag *et al.*, 2005). The SNP results thus indicate a greater frequency of heterozygotes than recorded in the allozyme studies and corroborate the previous finding that variation in flight metabolic rate is associated with genetic variation at *Pgi*.

**Table 4** Statistical tests of the effects of the allozyme allele F and the three SNPs on population growth rate and flight metabolic rate.

| | Population growth* | | | | Flight metabolism† | |
|---|---|---|---|---|---|---|
| Marker | $F_{3,39}$ | $r^2$ | $P_{\text{marker}}$ | $P_{\text{int}}$ | $F_{2,22}$ | $P$ |
| Allozyme F | 6.19 | 0.27 | 0.0008 | 0.0028 | 4.60 | 0.025 |
| SNP AA35 | 2.41 | 0.10 | 0.013 | 0.017 | 2.38 | 0.478 |
| SNP AA111 | 5.18 | 0.23 | 0.0006 | 0.0011 | 3.83 | 0.048 |
| SNP AA361 | 1.35 | 0.02 | 0.197 | 0.328 | 4.30 | 0.032 |
| SNP *f* | 4.44 | 0.20 | 0.0021 | 0.0044 | 10.88 | 0.0005 |

SNP *f* is a combination of SNPs AA111 and AA361 distinguishing allozyme allele F from all other allozyme alleles (see Table 3). Note that in the model for population growth each data point is a local population and in the model for flight metabolism each data point is an individual.

*Data from Hanski & Saccheri (2006), including 261 individuals from 43 local populations (isolated populations in fig. 2, panel A, in Hanski & Saccheri, 2006). The dependent variable is the residual from the regression of population growth on spatially correlated changes in population size and local population size. The explanatory variables are the molecular marker and habitat patch area. $P_{\text{marker}}$ and $P_{\text{int}}$ are the $P$-values for the effects of the molecular marker and interaction with habitat patch area. Fig. 1a in this paper shows the result for SNP *f*.

†Data from Haag *et al.* (2005). The dependent variable is peak flight metabolic rate, which is explained by body mass and the molecular marker without interaction. The effect of body mass was significant in all cases (results not shown). Fig. 1b in this paper shows the result for SNP *f*.

### SNP genotype frequencies in the metapopulation and in laboratory crosses

The population samples exhibited strong and highly significant deviations from the HW equilibrium at AA111 and AA361 due to an excess of heterozygotes (Table 5). One of the homozygous classes was much more abundant than the other one in each SNP (Table 5). The multilocus SNP genotype corresponding to the allozyme genotype FF (AA in AA35, CC in AA111 and AA in AA361) was the rarest one.

For the laboratory crosses, we present the results for the SNP AA111, which is the one most strongly associated with phenotypic variation (see above). Additional reason for giving the results for this SNP is that all parents were known to be heterozygous at AA111, but not uniformly so at the other SNPs. Moreover, the three SNPs are nonindependent, and the phase in the parents is unknown.

Among the 62 offspring from seven crosses, which were sampled and genotyped at the larval stage, there was a large and significant excess of heterozygotes (Table 5). The expected numbers, given heterozygous parents, are AA = 15.5, AC = 31 and CC = 15.5. The observed numbers were AA = 15, AC = 45 and CC = 2 ($P < 0.001$). Again, the homozygote CC, corresponding to the allozyme genotype FF, was rarer than AA.

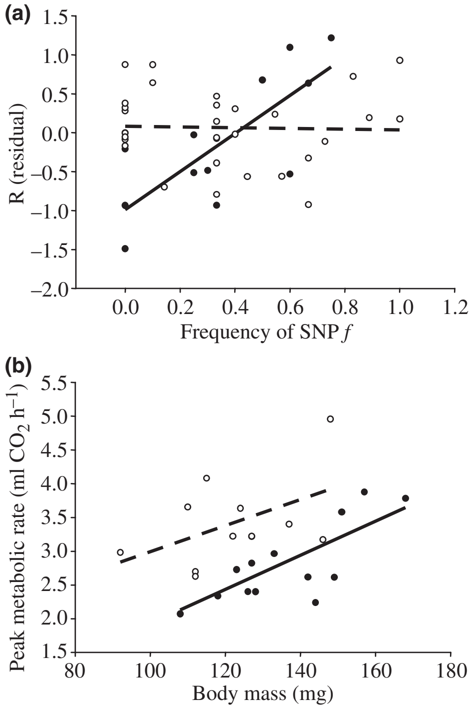

**Fig. 1** Population growth rate (a) and flight metabolic rate (b) explained by SNP *f* (combination of AA111 and AA361 as explained in Table 3). In (a) the dependent variable is the residual from the regression of population growth rate on spatially correlated changes in population size and local population size, and the explanatory variable is the frequency of individuals with the SNP *f* genotype in the local population. The two symbols distinguish small (< 0.05 ha, close symbols, continuous line) and large habitat patches (> 0.05 ha open symbols, broken line). In (b) the two symbols distinguish individuals with the SNP *f* genotype (open symbols, broken line) from individuals with different genotype (close symbols, continuous line). For statistics see Table 4.

Assuming that these deviations from Mendelian ratios were caused by viability selection at *Pgi* (see below for a discussion of this assumption), we can employ a simple model of selection at one locus (e.g. Hartl & Clark, 2007). Denoting the relative fitness of genotypes AA, AC and CC as 1, $1 - hs$ and $1 - s$, we obtain an estimate of $h = -0.58$, indicating heterozygote advantage. This means that among the three genotypes, the heterozygotes had the highest fitness, and thus heterozygote excess cannot be explained just by C being a partially recessive deleterious allele. If we now employ a standard model of one-locus two-allele overdominant selection (e.g. Hartl & Clark, 2007), denoting the relative fitnesses of the homozygotes relative to the heterozygote as $1 - s_{\mathrm{AA}}$ and $1 - s_{\mathrm{CC}}$, we obtain the estimates $s_{\mathrm{AA}} = 0.33$ and $s_{\mathrm{CC}} = 0.91$. If the polymorphism at SNP AA111 was maintained by overdominance, the expected equilibrium frequency of the C allele would be $q = s_{\mathrm{AA}}/(s_{\mathrm{AA}} + s_{\mathrm{CC}}) =$ 27%. The observed frequency (after selection) of C in samples 1 and 2 (Table 5) was 28% in both cases.

## Discussion

Patterns of molecular variation at *Pgi* are inconsistent with patterns expected for neutrally evolving loci. In two random population samples, we observed systematic deviations from the HW equilibrium, with a significant excess of heterozygotes. Laboratory crosses produced offspring with a similar excess of heterozygotes. In the latter case, because we know the parents' genotypes, we can exclude nonrandom mating as an explanation for the observed heterozygote excess.

Molecular variation at *Pgi* is in contrast to the variation at other allozyme, SNP and microsatellite loci studied in the same metapopulation of the Glanville fritillary (Hanski & Saccheri, 2006; Orsini *et al.*, 2008). These other loci were either in HW equilibrium or showed a deficit of heterozygotes due to spatial population structure (Orsini *et al.*, 2008). The different behaviour of *Pgi* when compared with the other loci provides a control for confounding factors, such as population history, inbreeding or spatial population structure as possible causes of departure from the HW equilibrium. All such factors would be expected to affect all loci in the same manner.

The present results suggest that there could be overdominance, selective advantage of heterozygotes, at *Pgi* in the metapopulation of the Glanville fritillary. Overdominance is a form of balancing selection that can affect both populations (maintenance of genetic variation) and individuals (heterozygote superiority) (Pamilo & Palsson, 1998). Overdominance could explain the significantly higher genetic variation at PGI than in other allozyme loci as well as the higher molecular variation at *Pgi* in comparison with other central metabolic genes (Hanski & Saccheri, 2006; C.W. Wheat, C.R. Haag, H. Frilander & I. Hanski, unpublished data).

One of the classic examples of overdominance is *Pgi* in *Colias* butterflies. The explanation suggested by Watt (1977, 1983) and Watt *et al.* (1983, 2003) for the superior performance of heterozygous individuals is based on the kinetic properties and thermal stability of the different PGI alleles. In the case of the Glanville fritillary, the flight metabolic rate of AC heterozygotes at AA111 is higher than that of AA homozygotes (Niitepõld *et al.*, in press), which could be due to kinetic superiority of heterozygotes over homozygotes. Similarly, there are interactions between the *Pgi* genotype (AA111) and body temperature at flight and fecundity (Haag *et al.*, 2005; Saastamoinen, 2007; Saastamoinen & Hanski, 2008) that are consistent with Watt's enzyme kinetic hypothesis. However, one of the main differences between the *Colias* butterflies and the Glanville fritillary is that fitness differences in *Colias* are only observed in adults, whereas selection against homozygotes in the Glanville fritillary is evident already in early larval development.

Although the results presented here would suggest that variation in the *Pgi* itself is associated with phenotypic variation, other studies have shown that traits that

**Table 5** The observed genotypic frequencies in the three SNPs in population samples from Finland.

| | | | Genotype numbers | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | *N* | Locus | Hom1 | Het | Hom2 | $H_{obs}$ | $H_{exp}$ | $\chi^2$ |
| Random sample of adults* | 335 | AA35 | 103 | 208 | 24 | 0.621 | 0.472 | 33.22**** |
| | | AA111 | 153 | 174 | 8 | 0.519 | 0.406 | 25.94**** |
| | | AA361 | 13 | 200 | 122 | 0.597 | 0.447 | 37.69**** |
| Random sample of larvae† | 224 | AA35 | 67 | 148 | 9 | 0.661 | 0.466 | 38.84**** |
| | | AA111 | 105 | 111 | 8 | 0.496 | 0.406 | 10.82** |
| | | AA361 | 1 | 147 | 76 | 0.656 | 0.444 | 51.23**** |
| Progeny of the crosses‡ | 62 | AA35 | na | na | na | | | |
| | | AA111 | 15 | 45 | 2 | 0.726 | 0.5 | 18.10*** |
| | | AA361 | na | na | na | | | |

Sample size (*N*), genotype frequencies, observed ($H_{obs}$) and expected ($H_{exp}$) heterozygosities and chi-squared values of the tests for departure from the Hardy–Weinberg equilibrium (random samples) or departure from Mendelian expectations (crosses) are shown. Tests are based on one (population samples) or two (crosses) degrees of freedom. Hom1 and Hom2 indicate homozygotes 1 and 2 (arbitrarily defined as Hom1 = AA at all three loci), Het indicates the number of heterozygotes. In the progeny of the crosses, SNP genotypes at AA35 and AA361 were not assessed (na), because parents of different crosses had different genotypes at these loci, whereas they were all heterozygous at AA111. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

*Adult butterflies caught in 55 habitat patches in a network of local populations within an area of $5 \times 4$ km$^2$.

†One to four larvae from each family in 23 local populations in a network within an area of $5 \times 4$ km$^2$.

‡Progeny of seven laboratory crosses. The parents were all heterozygote at the SNP AA111.

appear to follow simple patterns of Mendelian inheritance may in fact have more complex genetic architectures (e.g. Mackay & Fry, 1996; Clark & Wang, 1997; Matioli & Templeton, 1999), including epistatic interactions (Grimsley *et al.*, 1998; Richman *et al.*, 2003; Kroymann & Mitchell-Olds, 2005; Shiina *et al.*, 2006) and linkage of mutations with antagonistic fitness effects (Gratten *et al.*, 2008). These results suggest the following alternative explanation for the observed heterozygote excess in the metapopulation of the Glanville fritillary.

It is possible that selection affects the entire chromosomal region in which *Pgi* is located, because linkage disequilibrium may extend over several loci. In this case, selection at linked loci may contribute to an apparent selection at *Pgi*, including apparent overdominance. In particular, deleterious alleles at two or more loci in 'repulsion disequilibrium' (i.e. in linkage disequilibrium with deleterious alleles occurring on the opposite homologous chromosomes; Frydenberg, 1963; Ohta, 1971) can lead to an apparent overdominance at linked loci if the deleterious alleles are recessive or partially recessive. Such linkage disequilibrium has the consequence that homozygosity of either chromosome leads to homozygosity of at least one (partially) recessive deleterious allele, and hence the overall effect is an apparent fitness advantage of the heterozygotes. This process is also known as associative overdominance (Frydenberg, 1963; Ohta, 1971). Associative overdominance is expected to be stronger under inbreeding, in small populations, in the presence of tight linkage and under weak selection (Ohta, 1971; Palsson & Pamilo, 1999). Several characteristics of the Glanville fritillary metapopulation, such as small size of local populations, spatial population structure and colonization history, may increase the strength of inbreeding and linkage disequilibrium in comparison with large outbreeding populations (Hanski & Saccheri, 2006).

In principle, excess heterozygosity might arise due to the presence of only one partially recessive deleterious allele at or near *Pgi* (Wallace, 1958; Lewontin & Cockerham, 1959; Wilder & Hammer, 2004). However, the negative estimate of the dominance coefficient in our results does not fit this scenario. If heterozygote excess was caused by only one (partially) deleterious allele, the estimate of the dominance coefficient would be expected to be $0 < h < 0.5$, and thus one of the homozygotes should have the highest estimated fitness of the three genotypes (Fu & Ritland, 1994). One caveat, however, is that as our data do not allow us to distinguish between selection at *Pgi* and selection at linked loci; the inferred selection against the two homozygotes at AA111 should be interpreted cautiously. Nonetheless, these results highlight three points.

First, the selection coefficients calculated based on the crosses indicate that heterozygotes at AA111 (or in the chromosomal region marked by AA111) have a clear advantage over both homozygotes. Second, selection against one of the homozygotes is stronger than against the other homozygote. And third, assuming that selection operates at *Pgi* itself, the results imply an equilibrium frequency of 0.27 of the allele associated with the less fit homozygote. This latter value corresponds closely to the estimated allele frequencies in random population samples (although we do not have an estimate of allele frequencies before selection). The calculation of the expected equilibrium allele frequency assumes that selection operates only through differential survival. Alternatively, as suggested by previous results, the homozygote that appears to have a less severely reduced juvenile survival in our data has reduced flight capacity and fecundity (Haag *et al.*, 2005; Saastamoinen, 2007; Saastamoinen & Hanski, 2008). It is therefore possible that a balance between these effects on different fitness components contributes to the observed net effects of selection at *Pgi*.

In conclusion, the evidence presented here suggests that molecular variation at *Pgi* is associated with individual performance and population dynamics in the Glanville fritillary butterfly. Our results suggest that natural selection at *Pgi* and/or at closely linked genes results in a net fitness advantage of heterozygotes, although we do not have conclusive evidence that this would be due to direct overdominance at *Pgi*. Although some issues remain open, we have progressed by identifying SNPs that are associated with major phenotypic variation in a well-studied butterfly metapopulation.

## Acknowledgments

We thank Joost Raeymaekers and Pekka Pamilo for their comments on the manuscript and Toshka Nyman and Marja-Leena Peltonen for technical assistance. This study has been funded by the Academy of Finland (grant numbers 38604 and 44887, Finnish Centre of Excellence Programme, 2000–2005). L. Orsini acknowledges a grant from the Finnish Cultural Foundation and I. Hanski acknowledges support from the Academy of Finland (grant numbers 38604 and 44887, Finnish Centre of Excellence Programme, 2000–2005) and the US National Science Foundation (grant IBN-0412651).

## References

Barbadilla, A., King, L.M. & Lewontin, R.C. 1996. What does electrophoretic variation tell us about protein variation? *Mol. Biol. Evol.* **13**: 427–432.

Batley, J. & Hayes, P.K. 2003. Development of high throughput single nucleotide polymorphism genotyping for the analysis of *Nodularia* (Cyanobacteria) population genetics. *J. Phycol.* **39**: 248–252.

Clark, A.G. & Wang, L. 1997. Epistasis in measured genotypes: *Drosophila* P-element insertions. *Genetics* **147**: 157–163.

Dahlhoff, E.P. & Rank, N.E. 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. Natl Acad. Sci. USA* **97**: 10056–10061.

Filatov, D.A. & Charlesworth, D. 1999. DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. *Genetics* **153**: 1423–1434.

Frydenberg, O. 1963. Population studies of a lethal mutant in *Drosophila melanogaster.* I. Behaviour in populations with discrete generations. *Hereditas* **50**: 89–116.

Fu, Y.B. & Ritland, K. 1994. Evidence for the partial dominance of viability genes contributing to inbreeding depression in *Mimulus guttatus*. *Genetics* **136**: 323–331.

Gratten, J., Wilson, A.J., McRae, A.F., Beraldi, D., Visscher, P.M., Pemberton, J.M. & Slate, J. 2008. A localized negative genetic correlation constrains microevolution of coat color in wild sheep. *Science* **319**: 318–320.

Grimsley, C., Mather, K.A. & Ober, C. 1998. HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. *Mol. Biol. Evol.* **15**: 1581–1588.

Haag, C.R., Saastamoinen, M., Marden, J. & Hanski, I. 2005. A candidate locus for variation in dipersal rate in a butterfly metapopulation. *Proc. R. Soc. B* **272**: 2449–2456.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **41**: 95–98.

Hanski, I. 1998. Metapopulation dynamics. *Nature* **396**: 41–49.

Hanski, I. 1999. *Metapopulation Ecology*. Oxford University Press, New York.

Hanski, I. & Saccheri, I. 2006. Molecular-level variation affects population growth in a butterfly metapopulation. *PLoS Biol.* **4**: e129.

Hartl, D.L. & Clark, A.G. 2007. *Principles of Population Genetics*, 4th edn. Sinauer and Associates, Sunderland, MA.

Johnson, G.B. 1977. Assessing electrophoretic similarity – problem of hidden heterogeneity. *Annu. Rev. Ecol. Syst.* **8**: 309–328.

Katz, L.A. & Harrison, R.G. 1997. Balancing selection on electrophoretic variation of phosphoglucose isomerase in two species of field cricket: *Gryllus veletis* and *G. offnsylvanicus*. *Genetics* **147**: 609–621.

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

Lewontin, R.C. & Cockerham, C.C. 1959. The goodness-of-fit test for detecting natural selection in random mating populations. *Evolution* **13**: 561–564.

Mackay, T.F. & Fry, J.D. 1996. Polygenic mutation in *Drosophila melanogaster*: genetic interactions between selection lines and candidate quantitative trait loci. *Genetics* **144**: 671–688.

Matioli, S.R. & Templeton, A.R. 1999. Coadapted gene complexes for morphological traits in *Drosophila mercatorum*. Two-loci interactions. *Heredity* **83** (Pt 1): 54–61.

Nieminen, M., Siljander, M. & Hanski, I. 2004. Structure and dynamics of *Melitaea cinxia* metapopulations. In: *On the Wings of the Checkerspots: A Model System for Population Biology* (P.R. Ehrlich & I. Hanki, eds), pp. 63–91. Oxford University Press, New York.

Niitepõld, K., Smith, A.D., Osborne, J.L., Reynolds, D.R., Carreck, N.L., Martin, A.P., Marden, J.H., Ovaskainen, O. & Hanski, I. 2009. Flight metabolic rate and *Pgi* genotype influence butterfly dispersal rate in the field. *Ecology*, in press.

Ohta, T. 1971. Associative overdominance caused by linked detrimental mutations. *Genet. Res.* **18**: 277–286.

Orsini, L., Corander, J., Alasentie, A. & Hanski, I. 2008. Genetic spatial structure in a butterfly metapopulation correlates better with past than present demographic structure. *Mol. Ecol.* **17**: 2629–2642.

Palsson, S. & Pamilo, P. 1999. The effects of deleterious mutations on linked, neutral variation in small populations. *Genetics* **153**: 475–483.

Pamilo, P. & Palsson, S. 1998. Associative overdominance, heterozygosity and fitness. *Heredity* **81** (Pt 4): 381–389.

Patarnello, T. & Battaglia, B. 1992. Glucosephosphate isomerase and fitness: effects of temperature on genotype dependent mortality and enzyme activity in 2 species of the genus *Gammarus* (Crustacea, Amphipoda). *Evolution* **46**: 1568–1573.

Richman, A.D., Herrera, L.G., Nash, D. & Schierup, M.H. 2003. Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. *Genet. Res.* **82**: 89–99.

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

Saastamoinen, M. 2007. Life-history, genotypic, and environmental correlates of clutch size in the Glanville fritillary butterfly. *Ecol. Entomol.* **32**: 235–242.

Saastamoinen, M. & Hanski, I. 2008. Genotypic and environmental effects on flight activity and oviposition in the Glanville fritillary butterfly. *Am. Nat.* **171**: 701–712.

Saccheri, I.J., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W. & Hanski, I. 1998. Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**: 491–494.

Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M., Anzai, T., Kulski, J.K., Kikkawa, E., Naruse, T., Kimura, N., Yanagiya, K., Watanabe, A., Hosomichi, K., Kohara, S., Iwamoto, C., Umehara, Y., Meyer, A., Wanner, V., Sano, K., Macquin, C., Ikeo, K., Tokunaga, K., Gojobori, T., Inoko, H. & Bahram, S. 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**: 1555–1570.

Wallace, B. 1958. The comparison of observed and calculated zygotic distributions. *Evolution* **12**: 113–115.

Watt, W.B. 1977. Adaptation at specific loci. I. Natural selection on phosphoglucose isomerase of Colias butterflies: biochemical and population aspects. *Genetics* **87**: 177–194.

Watt, W.B. 1983. Adaptation at specific loci. II. Demographic and biochemical elements in the maintenance of the Colias Pgi polymorphism. *Genetics* **103**: 691–724.

Watt, W.B. 1992. Eggs, enzymes, and evolution: natural genetic variants change insect fecundity. *Proc. Natl Acad. Sci. USA* **89**: 10608–10612.

Watt, W.B., Cassin, R.C. & Swan, M.S. 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, W.B., Wheat, C.W., Meyer, E.H. & 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). *Mol. Ecol.* **12**: 1265–1275.

Wheat, C.W., Watt, W.B., Pollock, D.D. & Schulte, P.M. 2006. From DNA to fitness differences: sequences and structures of adaptive variants of Colias phosphoglucose isomerase (PGI). *Mol. Biol. Evol.* **23**: 499–512.

Wilder, J.A. & Hammer, M.F. 2004. European ACP1*C allele has recessive deleterious effects on early life viability. *Hum. Biol.* **76**: 817–835.

Zamer, W.E. & Hoffmann, R.J. 1989. Allozymes of glucose-6-phosphate isomerase differentially modulate pentose-shunt metabolism in the sea anemone Metridium senile. *Proc. Natl Acad. Sci. USA* **86**: 2737–2741.

Zera, A.J. 1987. Temperature-dependent kinetic variation among phosphoglucose isomerase allozymes from the wing-polymorphic water strider, *Limnoporus canaliculatus*. *Mol. Biol. Evol.* **4**: 266–285.

## Supporting information

**Figure S1** Synonymous and nonsynonymous segregating sites across 12 exons at *Pgi*. Each row is an individual from the Finnish metapopulation, with segregating variation indicated by nucleotide positions at the top of the column. Dots represent nucleotides that are identical to the first entry in the column. Nonsynonymous polymorphisms are identified by an asterisk (*). The three charge-changing amino acids used in the SNP design (Table 1) are highlighted in bold face. Vertical lines delimit exon boundaries.

**Table S1** List of samples sequenced at the *Pgi*. Sequences results at the three target SNPs, SNP genotypes, allozyme genotypes and accession numbers to GenBank are shown.

**Table S2** List of pseudo-haplotypes as inferred from the unphased data using the program DNAsp (Rozas *et al.*, 2003). For each pseudo-haplotype, the allozyme allele based on the allozyme electrophoresis, the inferred allozyme group and the net charge are shown. The inferred haplotype group is based on the information in Table 2.

**Appendix S1** RNA isolation, cDNA synthesis and sequencing of *Pgi* cDNA.