Decoupled differentiation of gene expression and coding sequence among *Drosophila* populations

Michael H. Kohn[1]*, Joshua Shapiro[2] and Chung-I Wu[3]

[1]*Department of Ecology & Evolutionary Biology, Rice University, MS 170, 6100 Main Street, Houston, Texas 77005, U.S.A.*

[2]*Lewis-Sigler Institute for Integrative Genomics & Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, U.S.A.*

[3]*Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL, 60637-1573, U.S.A.*

(Received 5 March 2008, accepted 12 April 2008)

Owing to the relevance to evolutionary theories of genotypic and phenotypic evolution, the correspondence of differentiation among natural populations in complex phenotypic traits and genetic markers has been studied extensively, and generally found to be poor. In contrast, the correspondence of differentiation among natural populations in gene expression, now often considered a genomic era proxy for the phenotype, and genetic markers, remains largely unexplored. Here, an analysis of expression and nucleotide sequence polymorphism of 106 genes in *Drosophila melanogaster* strains of the Cosmopolitan (M) and Zimbabwe, Africa (Z) mating races showed that differentiation of gene expression and of coding sequences, measured as Q_{ST} and G_{ST}, respectively, were uncorrelated and, generally, $Q_{ST} > G_{ST}$. However, an exploratory analysis showed that G_{ST} of the 5 prime sequences of genes was correlated with Q_{ST} calculated from expression data, while G_{ST} of the coding sequences remained uncorrelated with Q_{ST}. This scenario is consistent with the population differentiation at *cis*-regulatory regions that is decoupled from differentiation of the coding regions. However, despite evidence for selection on global levels of gene expression (deduced from $Q_{ST} > G_{ST}$), 5 prime sequence polymorphisms generally were compatible with selective neutrality, suggesting differentiation in *cis*-regulated gene expression for these genes has been promoted by drift or selection too weak or too long ago to be detected, or higher organizational levels underlying the genetic architecture of expression are targets of selection. In all, this raises the question how selection on the expression changes (i.e. the phenotype) can be so obvious yet elusive at the level of the nucleotide sequence. Our contrasts between genetic differentiation of populations in expression and sequences revealed that even when genotype and phenotype can be connected the sources of variation that are the target of selection remain to be identified.

Key words: complex trait, evolution, genetic differentiation, gene expression, microarray

INTRODUCTION

Diversity in phenotypic complex traits is a virtually ubiquitous natural phenomenon, and so is genetic diversity. One goal of molecular population genetic studies is to identify some of the mutations underlying this diversity in phenotypic complex traits, e.g. by means of DNA coding sequence surveys for genes that stand out in their levels of differentiation among populations (Black et al., 2001; Greenberg and Wu, 2006; Merila and Crnokrak, 2001; Purugganan and Gibson, 2003). In addition to such diversity at the level of proteins, diversity in phenotypic complex traits could be due to gene regulatory changes, both *cis*- and *trans*-regulated (Brem et al., 2005; Coffman et al., 2005; Demuth and Wade, 2006; Edwards et al., 2006; Fay et al., 2004; Phillips, 2005; Prud'homme

Edited by Fumio Tajima

* Corresponding author. E-mail: hmkohn@rice.edu

Note: Nucleotide sequence data reported are available in the DDBJ/EMBL/GenBank databases under the accession numbers EU744337-EU744538. Supplementary Tables in this article is at http://wwwsoc.nii.ac.jp/gsj3/sup/83(3)Kohn/

et al., 2006; Tao et al., 2006). Diversity at this level could be quantified on gene expression arrays (Gibson, 2002). Following other such studies of complex traits (Merila and Crnokrak, 2001), i.e. assuming gene expression is a complex trait, e.g. (Crawford and Oleksiak, 2007; Gibson and Weir, 2005; Wang et al., 2006; Wray, 2003), differentiation between populations in gene expression could be summarized by the Q_{ST} statistic. Thus, we are now entering an era in which it is possible to compare and contrast population differentiation at the nucleotide sequence level as measured by F_{ST} or its analogs, and at the levels of gene expression, as measured by Q_{ST}.

One promise held by this era is the ability to close the gap between genotype and phenotype, granted we view gene expression as a valid genomic-era proxy for more classical complex phenotypes studied within this framework. The numeric comparison between Q_{ST} calculated from gene expression data and F_{ST} could reveal specific biological factors affecting such comparisons that were difficult to disentangle during more classical such comparisons, in particular selection (Merila and Crnokrak, 2001). Moreover, such studies may help identify a framework to more broadly search for the link between genotype and phenotype. For example, results may reveal whether selection on gene expression would translate into selection on the regulatory regions of individual genes, and how pervasive selection on transcription is in general (Fay and Wittkopp, 2008; Whitehead and Crawford, 2006a; Wray et al., 2003; Yan and Zhou, 2004). Moreover, results may reveal the relative role of *cis*- versus *trans*-regulatory changes in the evolution of gene expression within and between species (Brem et al., 2002; Brown and Feder, 2005; Doss et al., 2005; Hahn, 2007; Johnson et al., 2005; Ludwig et al., 2000; Osada et al., 2006; Pastinen et al., 2004; Ranz and Machado, 2006; Ronald et al., 2005; Wang et al., 2007; Wayne et al., 2004; Whitehead and Crawford, 2006a; Wittkopp, 2005, 2006; Wittkopp et al., 2004; Wray et al., 2003; Yan and Zhou, 2004). If *cis*-regulatory changes were of notable importance, then, by factoring in their specific effect on the Q_{ST}-F_{ST} relationship, we finally would be able to link the genotype with the phenotype and study some of the unaccounted variance components in more detail.

Our goal here was to study the correspondence of Q_{ST} as deduced from gene expression data collected from microarrays with F_{ST} (here Nei's G_{ST}, see below) at the coding regions of genes, and to examine possible explanations for the observed patterns in the *Drosophila* system that we study (Fang et al., 2002; Fay and Wu, 2003; Greenberg et al., 2003; Greenberg and Wu, 2006; Hollocher et al., 1997a, 1997b; Kohn et al., 2004; Osada et al., 2006; Shapiro et al., 2007; Wu and Ting, 2004). The study is part of a long-term effort to identify some of the mutations that underlie the differentiation of *D. melanogaster* strains from Zimbabwe (Z), Africa, from most other African and Cosmopolitan strains (M), in that they preferentially mate with their own kind (Fang et al., 2002; Hollocher et al., 1997a; Takahashi and Ting, 2004). During mating experiments this sexual isolation is manifest as apparent female choice. The genetics of this M and Z differentiation in this phenotype is complex and could be a good model to study complex trait differentiation in general terms and the genetic architecture of gene expression differentiation in particular (Osada et al., 2006; Wang et al., 2008).

We draw upon a published dataset on gene expression differences between M and Z strains of *Drosophila* (Meiklejohn et al., 2003) and a second dataset on coding sequence polymorphism in M and Z strains of *Drosophila* (Shapiro et al., 2007) to compare genomic patterns of differentiation levels in these two measures. In addition, we conduct an exploratory study that considers the expression levels of genes and sequence polymorphisms in their 5 prime regions and coding regions.

METHODS

Nucleotide sequences

Sequence data: Sequence polymorphism and divergence data were taken from Shapiro et al. (2007), whose study included information on 6 Cosmopolitan M strains and 11 African strains from Zimbabwe, as well as *D. simulans*, that were informative for our study (c.f. Shapiro et al. (2007) for accession numbers). The data on coding sequence polymorphism were searched for overlap with the expression data (see below), yielding 106 genes that were common to both surveys and were based on African Z strains and Cosmopolitan M strains. *D. simulans* was used as an outgroup for sequence analyses (Shapiro et al., 2007). In addition, an exploratory analysis of sequence polymorphism and divergence in the 5 prime regions and the coding regions of eight genes (*CG11426, CG7966, Irp-1B, MtnA, trpl, CG16926, CG4757,* and *Cyp6a23*) was done. Briefly, primers amplifying about ~1 kilobasepairs (kb) long segments of 5 prime noncoding sequence and ~1 kb of coding region were designed (Fig. 1). Sequencing was done as described (Shapiro et al., 2007).

Analysis: Sequences were assembled and aligned in PHREDPHRAP (Nickerson et al., 1997) and CLUSTAL (Li, 2003). Genetic differentiation between M and Z strains of *D. melanogaster*, calculated as Nei's G_{ST}, an analog of Wright's F_{ST} for DNA sequences, were estimated from the data with LIBSEQUENCE C++ (Thornton, 2003). Significance of G_{ST} was determined by permutation of sites and Fisher's exact test. Basic population genetic parameters were calculated for the set of 8 genes.

Gene symbols, Flybase identifiers, chromosomal location, accessions, and recombination rates as determined

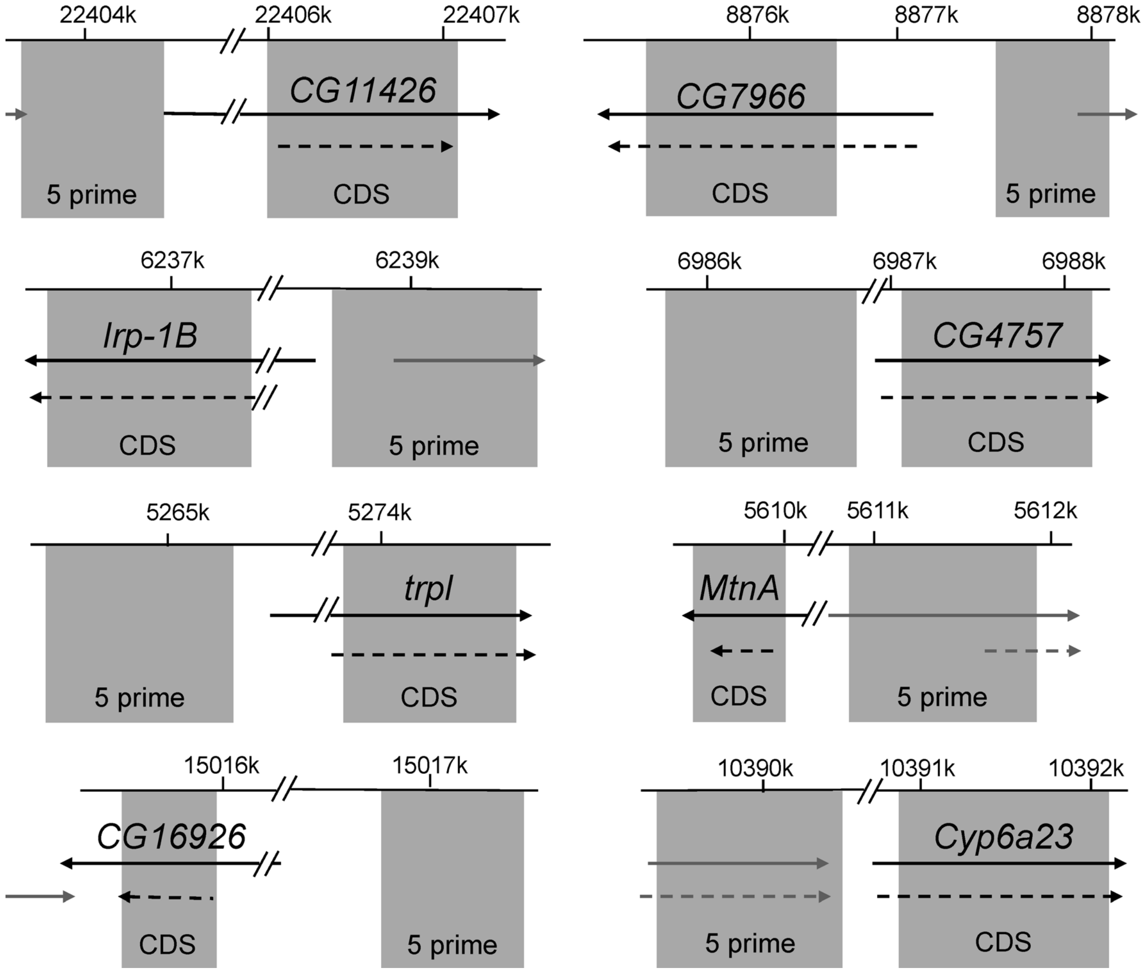

Fig. 1. Overview of sequence data collected for exploratory analysis of the relationship between Q_{ST} and G_{ST} in the 5 prime regions and coding regions of genes. For each of the eight genes the 5 prime- and coding regions (CDS) (grey shaded areas), the length of sequence (~1kb, c.f. top bar for scale), the annotated gene (black arrow pointing from 5' to 3') and the transcript (dashed arrow) are shown. In addition, other annotated genes and their transcripts are depicted correspondingly (grey arrows).

by using the *D. melanogaster* recombination rate calculator (implemented by Nadia Singh, Peter Arndt and Dmitri Petrov available at http://cgi.stanford.edu/~lipatov/recombination/recombination-rates.txt) are provided in Supplementary Tables 1 and 2. Rates of intragenic recombination were calculated (for the set of 8 genes) following Hudson and Kaplan (1985) as implemented in DNASP (Rozas et al., 2003).

Gene expression

Printed cDNA arrays: Published gene expression data for the whole fly collected by Meiklejohn et al. (2003) were retrieved from the Gene Expression Omnibus (GEO, accession series GSE539; http://www.ncbi.nlm.nih.gov/geo). These were obtained from four Cosmopolitan M strains (Canton-S, Oregon R, Hikone R, St. Louis) and from four African Z strains (Z53, Z30, Z29, Z2). Those genes that were common to the sequencing survey by Shapiro et al. (2007) were retained for analysis. The relative transcript abundance for each gene was expressed relative to the strain with the lowest transcript abundance (set to be 1), and these were adopted as reported from Meiklejohn et al. (2003).

Quantitative Real Time-PCR (qRT-PCR): For an exploratory analysis the eight genes *CG11426, CG7966, Irp-1B, MtnA, trpl, CG16926, CG4757,* and C*ypa23* expression data were collected with quantitative real time-PCR (qRT-PCR). These genes were drawn from a set of 364 genes that emerged as significantly different in expression between females of one M strain and one Z strain (Z30) (Osada et al., 2006) GO accessions GSM29579 and GSM29581 [France; M], and GSM29584 and GSM29585 [Z30; Z].

Briefly, RNA was harvested from the female head tissues of 4 M isofemale strains French, LA20, LA47, LA66 (LA designating the origin from Zambia, Africa) and the 4 Z strains ZS30, ZH18, ZH21, ZH12 from Sengwa and

Table 1. Sequence differentiation and polymorphisms underlying exploratory analysis

Gene name	G_{ST}	$P(G_{ST})$	M	M	M	M	M	M	Z	Z	Z	Z	Z	Z
			θ_W	θ_π	$\theta_{H'}$	D	F	H'	θ_W	θ_π	$\theta_{H'}$	D	F	H'
5 prime region														
CG11426	0.30	0.008	0.002	0.002	0.003	–0.17	0.44	–0.74	0.001	0.001	0.000	–0.05	–0.16	0.73
CG16926	0.00	1.000	0.014	0.013	0.013	–0.57	–0.62	0.70	0.021	0.021	0.018	0.04	0.44	1.65
CG4757	0.05	0.236	0.005	0.005	0.003	–0.40	–0.78	1.44	0.012	0.012	0.007	–0.06	–0.49	1.55
CG7966	0.27	0.010	0.006	0.004	0.001	–1.24	–1.83	1.40	0.006	0.005	0.008	–0.78	–0.39	–0.85
Cyp6a23	0.11	0.104	0.001	0.001	0.002	–0.75	–0.56	0.14	0.006	0.006	0.009	0.29	0.76	–0.45
Irp–1B	0.37	0.012	0.002	0.002	0.004	–0.83	0.13	–1.40	0.002	0.002	0.001	–0.41	–0.89	0.78
MtnA	0.10	0.185	0.002	0.001	0.002	–1.30	–1.04	–0.12	0.002	0.002	0.002	0.97	0.63	–0.23
trpl	0.06	0.071	0.008	0.008	0.012	–0.13	0.80	–0.75	0.027	0.023	0.070	–0.92	1.77	–4.38*
CDS														
CG11426	0.13	0.066	0.003	0.003	0.003	–0.19	0.54	–0.26	0.004	0.003	0.003	–0.72	–0.93	0.26
CG16926	0.04	0.138	0.012	0.011	0.010	–0.65	0.10	0.66	0.018	0.018	0.016	–0.23	–0.24	0.44
CG4757	0.17	0.028	0.005	0.005	0.003	–0.45	–0.78	1.27	0.007	0.007	0.007	0.24	0.63	0.16
CG7966	0.03	0.182	0.003	0.003	0.001	–1.16	–1.97	1.52	0.015	0.014	0.008	–0.42	–1.87	1.37
Cyp6a23	0.10	0.034	0.006	0.007	0.006	1.22	1.49	0.73	0.005	0.006	0.006	0.48	0.46	0.02
Irp-1B	0.41	0.009	0.002	0.001	0.000	–1.34	–0.58	1.60	0.001	0.001	0.000	–0.17	1.45	1.90
MtnA	0.10	0.178	0.002	0.002	0.001	0.31	–0.16	0.37	0.001	0.001	0.000	–0.93	–1.13	0.55
trpl	0.12	0.028	0.004	0.004	0.003	–0.20	–0.74	0.68	0.011	0.011	0.009	–0.03	–0.09	0.37

Provided are gene name, genetic differentiation between M and Z strains (G_{ST}). The polymorphism estimates for 4Nu θ_W, θ_π, $\theta_{H'}$ and the corresponding tests for deviations from neutral expectations based on tests by Tajima's D, Fu and Li's F, and Fay and Wu's H' are provided. Significance of each at $\alpha = 0.05$ (*) after corrections for multiple testing were made was $\alpha = 0.0016$.

Harare, Zimbabwe, Africa. RNA extractions and qRT-PCRs using the SyberGreen chemistry were conducted as described (Osada et al., 2006) (c.f. Supplementary Table 3). Cycle thresholds (CT), i.e. expression levels, for a control gene (actin 57B; CG10067) were adjusted such that the CT difference between strains was set as zero. CT-values for each gene were then adjusted using the multiplication factor used to adjust actin. Normalized CT-values were used to compute $\log_2$-ratios and relative expression levels for each gene. These were expressed relative to the strain with the lowest transcript abundance (set to be 1) following Meiklejohn et al. (2003) (Supplementary Table 3).

Initially, we replicated the results obtained by Osada et al. (2006), who assayed expression from the M strain from France and Z strain ZS30 on microarrays. Here, first, expression from the M strain from France and Z strain ZS30 was assayed with qRT-PCR. Then, expression from the M strain from France and the Z strains ZS30 and ZS53 was assayed a second time with qRT-PCR based on different RNA extracts (c.f. Supplementary Table 3). Subsequently, a separate set of concurrently ran qRT-PCR reactions were used to assay expression in the strains Fr, LA20, LA47, LA66, ZS30, ZH18, ZH21, ZH12, and these were used for the computation of Q_{ST}. Note, while the three different runs based on different RNA extracts were compared with respect to the up or down regulation of genes, they were not compared in magnitude (c.f. Supplementary Table 3).

Analysis: Q_{ST}, a measure for differentiation among populations in complex traits, was computed as an estimator of differentiation in gene expression as follows.

The within population variance σ_w^2 was computed as:

$$\frac{1}{a(n-1)}\sum_{i=1}^{i=a}\sum_{j=1}^{j=n}(Y_{ij}-\overline{Y}_i)^2$$

The between population variance σ_b^2 was computed as:

$$\frac{n}{(a-1)}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{\overline{Y}})^2$$

Where a is the number of populations, n the number of expression measurements (i.e. samples) per population, Y_{ij} is the relative expression level of gene j in population i, $\overline{Y}_i$ is the average expression level in population i, and $\overline{\overline{Y}}$ is the grand mean (Sokal and Rohlf, 1995).

The variance estimators where used to calculate, analogous to Wright's (1951) F_{ST}, a per gene Q_{ST}:

$$\frac{\sigma_b^2}{(\sigma_b^2+2\sigma_w^2)}$$

The average Q_{ST} across genes was calculated as:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_b^{\ 2}}{(\sigma_b^{\ 2}+2\sigma_w^{\ 2})}$$

RESULTS AND DISCUSSION

Genomic inferences The comparison between differentiation of M and Z populations in coding sequence, expressed as G_{ST}, with gene expression differentiation expressed as Q_{ST}, showed that these two measures were uncorrelated (R^2~0.2%; Mantel's test n.s., Fig. 2). Genes with significant G_{ST} values ($\alpha = 0.05$, n = 45) only had slightly higher average Q_{ST} than genes with non-significant G_{ST} (n = 60) (Q_{ST} = 0.35 versus 0.27, Wilcoxon Rank Sums test, Chi-square approximation, n.s.). This non-correlation is consistent with at least four biological interpretations, among them: no differentiation, non-additivity and/or epistasis of gene expression, and selection. A fourth interpretation relies on the assumption that gene expression changes should be governed by genetic polymorphisms located either in the *cis*-regulatory region of genes or in *trans*-factors. The non-correlation that we observe would then be consistent with the evolutionary decoupling of the *cis*-regulatory regions and coding region of genes as well as with the overwhelming importance of *trans*-regulation of genes.

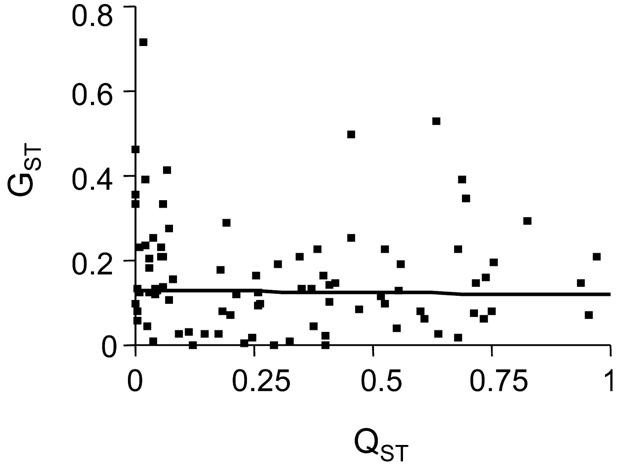

Fig. 2. Inferring the relationship of Q_{ST} with G_{ST} for a set of 106 genes.

First, when there is no differentiation in expression or sequence no correlation is expected (Fay et al., 2004), but this was not observed; 45 of 106 genes (42.5%) showed significant G_{ST} values (e.g. at $\alpha = 0.05$), indicating that there were genes in the data for which coding sequence differentiation between M and Z was substantial (Fig. 2). Similarly, 70 of the 106 (66.0%) genes had Q_{ST} values ~10% or larger, 28 of the 106 genes (26.4%) had Q_{ST} values ~50% or larger, suggesting that differentiation of M and Z strains in gene expression was considerable (Fig. 2). However, after applying the stringent Bonferroni correction for 106 tests at $\alpha = 0.05$ (i.e. $\alpha = 0.0005$) none of the G_{ST} values remained significant and, moreover, stringent analyses of the expression differences as done by Meiklejohn et al. (2003) indicated that none of the gene expression differences were fixed between M and Z. Thus, our study was concerned with the more common type of differentiation where natural populations differ in allele frequencies, not fixed differences.

Second, an assumption underlying the expected 1:1 relationship of Q_{st} with F_{ST} is that the effects of the genetic loci underlying the phenotype are additive. Here, we need to assume that the effects of gene expression are additive and no epistasis occurs. However, violation of this assumption is expected to result in the reduction and underestimation of Q_{ST} (Goudet and Buchl, 2006; Goudet and Martin, 2007), resulting in potentially false $Q_{ST} < F_{ST}$ relationships. However, here $Q_{ST} > F_{ST}$ was observed, and thus, other biological phenomena or violations of assumptions likely are of greater relevance. However, this topic remains disputed (Lopez Fanjul and Toro, 2007; Lopez-Fanjul et al., 2003) and more studies on additivity and epistasis of gene expression in natural populations are needed.

Third, contrasts between Q_{ST} and G_{ST} provide insights into the importance of selection as a cause for differentiation. Under neutrality as well as other assumptions (Merila and Crnokrak, 2001) there should be a 1:1 correspondence between Q_{ST} and G_{ST}. When genetic markers are neutral violation to this correspondence is consistent with divergent selection ($Q_{ST} > G_{ST}$) or stabilizing selection ($Q_{ST} < G_{ST}$) on the phenotype (Merila and Crnokrak, 2001). Here, Q_{ST} generally exceeded G_{ST} for genes with non-significant G_{ST} as well as for genes with significant G_{ST} (Wilcoxon Sign-Rank tests, $p < 0.0001$ and $p = 0.011$, respectively). Note that G_{ST} values < 0 were set to be zero for this comparison. This pattern is compatible with divergent selection promoting the differentiation in gene expression between M and Z. However, it is also conceivable that balancing selection on the nucleotide sequences, which would lead to a reduction of G_{ST}, resulted in the overall pattern $Q_{ST} > G_{ST}$. However, an increasing number of reports on (recurrent) divergent selective sweeps between African and Non-African populations of *D. melanogaster* suggest that the risk of overestimations of G_{ST} is higher than the risk of underestimations. When genes that potentially are subject to divergent selection were excluded, i.e. those with a significant G_{ST}, the slope in Fig. 2 became (non-significantly) negative. In other words, differentiation between M and Z strains in gene expression tended to be more pronounced in genes with lower levels of DNA sequence differentiation. The observed pattern $Q_{ST} > G_{ST}$ indicative of divergent selection reminds of the same pattern observed for numerous other comparisons between differentiation in genetic markers and classical phenotypic traits (Merila and Crnokrak, 2001), suggesting global levels of gene expression (inferred as Q_{ST} across numerous genes) behaves similar to other complex traits.

Fourth, consider the discussion surrounding the relative importance of *cis*- versus *trans*-regulation of gene expression (Doss et al., 2005; Goto et al., 2005; Hahn, 2007; Kulkarni and Arnosti, 2005; Osada et al., 2006; Prud'homme et al., 2006; Ranz and Machado, 2006; Ronald et al., 2005; Wittkopp, 2005; Wittkopp et al., 2004). A correlation between the differentiation in the coding region of genes and the 5 prime regions of genes would be expected only if it is assumed that the expression changes are caused in large part by mutations in the *cis*-regulatory regions of genes. In contrast, if most gene expression differences among natural populations were due to *trans*-factors then there would be no correlation between differentiation in the expression of a gene and its coding sequence expected. Our observed non-correlation between expression and coding sequence differentiation is consistent with *trans*-regulation of gene expression differences. However, *cis*-regulation can also account for the result as long as it is assumed that recombination has decoupled coding and non-coding regions of genes. We will address this issue further by means of an exploratory analysis in the following section.

Exploratory inferences of *cis* versus *trans* regulation

Because of the possibility to link gene expression with genetic differentiation at the molecular level contrasts between gene expression and genetic differentiation potentially reveal specific factors affecting the Q_{ST}-G_{ST} relationship. This should be an advantage compared to studies examining this relationship based on Q_{ST} computed from more classical phenotypes. The potential reward of such an analysis has been documented, for example, during a high-profile single-gene study in maize (Wang et al., 2001), which reported that whereas domestication selection has elicited evolutionary changes in the promoter of the *tb2* gene that are associated with the domesticated phenotype the coding region of the gene remained unaffected (c.f. also Hubbard et al. (2002), Weber et al. (2007), Wright et al. (2005)). Divergent selection on gene expression may elicit a similar response in natural populations as in domesticated species, but this remains to be investigated more broadly.

Consider that if *cis*-regulation was a common mode of gene regulation (note that this does not require the non-importance of *trans*-regulation) then there should be a positive relationship of G_{ST} in the 5 prime regions of genes with Q_{ST}. For this to be detected by our study it needs to be assumed that *cis*-regulatory elements are located in the ~1 kb –spanning sequenced region upstream of the genes (Fig. 1) or are in strong linkage disequilibrium with any such functional sites. In contrast, simple models of *trans*-regulation would not predict a correlation between G_{ST} in the 5 prime regions and Q_{ST}. Distinguishing between these two scenarios requires polymorphism data on the 5 prime regions of genes, coding regions of genes, and gene expression data. Unfortunately such a combination of systematically collected genomic data is only beginning to accumulate for the *Drosophila* strains under study. Thus, we tested these predictions during an exploratory study based on 5 prime and coding sequence data for eight genes (Fig. 1) and qRT-PCR gene expression measurements (Supplementary Table 2).

First, the small dataset mirrored the genomic pattern (see above), where a non-significant relationship of G_{ST} in the coding regions with Q_{ST} was observed ($R^2 < 3.9\%$; n.s., Fig. 3). However, the relationship of G_{ST} in the 5 prime regions with Q_{ST} was significant ($R^2 < 55.6\%$; $p = 0.03$, Fig. 3). False positive association of Q_{ST} and G_{ST} for the 5 prime regions due to the small number of Q_{ST} and G_{ST} values, and false negative dismissal of such a correlation for the coding regions, were excluded as source of error (Fig. 4). This pattern is consistent with *cis*-regulation if intragenic recombination has decoupled the 5 prime regions from the coding regions. Testing for intragenic (fine-scale) recombination rates indicated that all but one gene (MtnA) had at least one recombination event (Supplementary Table 4). However, broad-scale rates of recombination were of little explanatory power for differentiation in G_{ST} and Q_{ST} at the genomic scale and the exploratory study, with less than ~3% of the variance of the data explained by it (not shown). Low levels of fine-scale intragenic recombination may be sufficient to decouple *cis*-regulatory sites from the coding region.

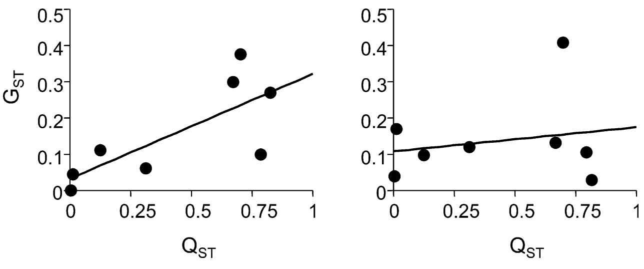

Fig. 3. Inferring the relationship between Q_{ST} and G_{ST} for the 5 prime regions (left panel) and G_{ST} for the coding regions (right panel) of eight genes.

Second, as for the large dataset, for the small dataset we found that, during Wilcoxon Signed Rank tests of matched pairs, $Q_{ST} > G_{ST}$ for the 5 prime regions ($p = 0.04$), $Q_{ST} > G_{ST}$ for the coding regions ($p = 0.04$), while G_{ST} for the 5 prime region and G_{ST} for the coding region did not differ (averages 0.157 versus 0.136, $p = 0.95$). Thus, assuming neutrality of the sequence data this observation

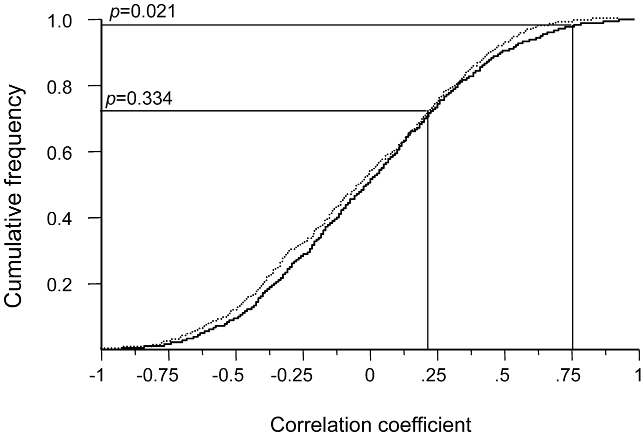

Fig. 4. Evaluation of the relationship between Q_{ST} and G_{ST} obtained during exploratory analysis. Using 1000 rounds of random sampling with replacement generated two datasets. The first dataset was generated from the random samples of Q_{ST} results and G_{ST} values for the 5 prime regions and the second dataset consisted of the corresponding random samples of G_{ST} values for the coding regions. Randomly paired Q_{ST} and G_{ST} values where used to estimate the distribution of correlation coefficients between Q_{ST} and G_{ST} for the 5 prime region (solid line) and the coding region (dashed line) to be used to estimate the probability associated with the observed correlations (0.75 for the 5 prime region and 0.23 for the coding region). The associated probabilities of obtaining these observed correlations by chance were estimated as 0.021 and 0.334, respectively.

is consistent with divergent selection on the expression differences. In light of this suggested relevance of selection, can we reject selective neutrality for the 5 prime regions of genes? After correcting for multiple testing (n = 32 per statistic) for significance at $\alpha = 0.05$ ($\alpha = 0.0016$) we found little evidence for selection on either the 5 prime regions or the coding regions when using Tajima's D, Fu and Li's F, or Fay and Wu's H tests (Table 1). One possible exception was the 5 prime region of the *trpl* gene where Fay and Wu's H was significant ($p < 0.001$) in the Z population. Neither Tajima's D or Fu and Li's F were significant for the gene however. Tajima's D was significant for CG7966 prior to corrections for multiple testing, but not thereafter. Overall, these results suggested that selection at the sequence level generally appeared to be absent, or that the signatures of selection were too spurious to be detected. Thus, genetic drift could have been the predominant process underlying the differentiation in the 5 prime sequences and coding sequences between M and Z for the genes studied.

CONCLUSION

An analysis of genomic patterns recapitulated the classical result $Q_{ST} > G_{ST}$. This suggested that gene expression behaves similar to other phenotypic traits and that divergent selection at some organizational level of the architecture of gene expression is plausible. An exploratory analysis of the eight genes for which the 5 prime sequences, in addition to the coding sequences, were available, revealed that selection was either absent or too elusive to be detected at the level of the nucleotide sequences. The role of the 5 prime sequences in promoting differentiation in expression was inferred from the observed correspondence between Q_{ST} and G_{ST}. This failure to reject selective neutrality at the DNA sequence level of the 5 prime regions raises the interesting question how selection appears to be so obvious a factor driving population differentiation at the level of gene expression (c.f. also Fay and Wittkopp (2008), Hahn (2007), Holloway et al. (2007), Lemos et al. (2005), Meiklejohn et al. (2003), Metta et al. (2006), Ranz and Machado (2006), Townsend et al. (2003), Whitehead and Crawford (2006a; 2006b), Wray (2003), Wray et al. (2003)), as it has been for other complex traits in other studies, yet so elusive at the sequence level. A few studies that were able to document the role of 5 prime sequences in governing expression and phenotypes were able to document selection at the sequence level, however, suggesting both selection and drift can promote the differentiation among natural populations in expression and sequence (Edwards et al., 2006; Fang et al., 2002; Weber et al., 2007).

Overall, our observed pattern was similar to the high-profile single-gene study in maize (Wang et al., 2001), which reported that whereas domestication selection has elicited evolutionary changes in the promoter of the *tb2* gene the coding region of the gene remained unaffected (c.f. also Hubbard et al. (2002), Weber et al. (2007), Wright et al. (2005)). However, in contrast to the study on domestication selection in maize, where artificial selection might have been intense, differentiation at the DNA sequence level in our set of genes and study system was compatible with genetic drift. Drift is an important process promoting genetic differentiation in natural populations, and it is plausible that drift frequently governs non-coding as well as coding sequence differentiation. In addition, conceivably, the difference between studies is further explained by the fact that neither of the genes we have examined, albeit differentiated in expression between M and Z, might directly affect the M and Z phenotype.

Studies that compare genetic differentiation and differentiation in expression could be instrumental when attempting to understand the relationship of genotypic with phenotypic differentiation in natural populations, to close the gap between the genotype-phenotype relationship, and to identify the targets of selection. We suspect that this will require a better future understanding of the genetic architecture of gene expression and the genotype-phenotype relationship (Gibson, 2002; Wittkopp, 2007; Yan and Zhou, 2004).

For help we thank Mao-Lien Wu, Anthony Greenberg, Xinmin

Li, Qun (Jean) Shi, Jian (Jamie) Zhou, Jaejung Kim (University of Chicago), Wei Huang and the Chinese High-Tech Program (863). We thank two anonymous reviewers for their comments on the manuscript. The project has been supported by NIH grants to C-I Wu.

REFERENCES

Black, W. C., Baer, C. F., Antolin, M. F., and DuTeau, N. M. (2001) Population genomics: genome-wide sampling of insect populations. Ann. Rev. Entomol. **46**, 441–469.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. Science **296**, 752–755.

Brem, R. B., Storey, J. D., Whittle, J., and Kruglyak, L. (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature **436**, 701–703.

Brown, R. P., and Feder, M. E. (2005) Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. BMC Genomics **6**, 110.

Coffman, C. J., Wayne, M. L., Nuzhdin, S. V., Higgins, L. A., and McIntyre, L. M. (2005) Identification of co-regulated transcripts affecting male body size in *Drosophila*. Genome Biol. **6**, R53.

Crawford, D. L., and Oleksiak, M. F. (2007) The biological importance of measuring individual variation. J. Exp. Biol. **210**, 1613–1621.

Demuth, J. P., and Wade, M. J. (2006) Experimental methods for measuring gene interactions. Annu. Rev. of Ecol. Evol. Syst. **37**, 289–316.

Doss, S., Schadt, E. E., Drake, T. A., and Lusis, A. J. (2005) *Cis*-acting expression quantitative trait loci in mice. Genome Res. **15**, 681–691.

Edwards, A. C., Rollmann, S. M., Morgan, T. J., and Mackay, T. F. C. (2006) Quantitative genomics of aggressive behavior in *Drosophila melanogaster*. Plos Genetics **2**, 1386–1395.

Fang, S., Takahashi, A., and Wu, C. I. (2002) A mutation in the promoter of *desaturase 2* is correlated with sexual isolation between *Drosophila* behavioral races. Genetics **162**, 781–784.

Fay, J. C., and Wu, C. I. (2003) Sequence divergence, functional constraint, and selection in protein evolution. Annu. Rev. Genom. Hum. Genet. **4**, 213–235.

Fay, J. C., and Wittkopp, P. J. (2008) Evaluating the role of natural selection in the evolution of gene regulation. Heredity **100**, 191–199.

Fay, J. C., McCullough, H. L., Sniegowski, P. D., and Eisen, M. B. (2004) Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*. Genome Biol. **5**, R26.

Gibson, G. (2002) Microarrays in ecology and evolution: a preview. Mol. Ecol. **11**, 17–24.

Gibson, G., and Weir, B. (2005) The quantitative genetics of transcription. Trends Genet. **21**, 616–623.

Goto, H., Szmidt, A. E., Yamazaki, T., and Inomata, N. (2005) Effect of nucleotide polymorphism in *cis*-regulatory and coding regions on amylase activity and fitness in *Drosophila melanogaster*. Heredity **95**, 369–376.

Goudet, J., and Buchl, L. (2006) The effects of dominance, regular inbreeding and sampling design on Q_{ST}, an estimator of population differentiation for quantitative traits. Genetics **172**, 1337–1347.

Goudet, J., and Martin, G. (2007) Under neutrality, $Q_{ST} \le F_{ST}$ when there is dominance in an Island model. Genetics **176**, 1371–1374.

Greenberg, A. J., and Wu, C. I. (2006) Molecular genetics of natural populations. Mol. Biol. Evol. **23**, 883–886.

Greenberg, A. J., Moran, J. R., Coyne, J. A., and Wu, C. I. (2003) Ecological adaptation during incipient speciation revealed by precise gene replacement. Science **302**, 1754–1757.

Hahn, M. W. (2007) Detecting natural selection on *cis*-regulatory DNA. Genetica **129**, 7–18.

Hollocher, H., Ting, C. T., Pollack, F., and Wu, C. I. (1997a) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Variation in mating preference and correlation between sexes. Evolution **51**, 1175–1181.

Hollocher, H., Ting, C. T., Wu, M. L., and Wu, C. I. (1997b) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Extensive genetic divergence without reinforcement. Genetics **147**, 1191–1201.

Holloway, A. K., Lawniczak, M. K. N., Mezey, J. G., Begun, D. J., and Jones, C. D. (2007) Adaptive gene expression divergence inferred from population genomics. Plos Genetics **3**, 2007–2013.

Hubbard, L., McSteen, P., Doebley, J., and Hake, S. (2002) Expression patterns and mutant phenotype of *teosinte branched1* correlate with growth suppression in maize and teosinte. Genetics **162**, 1927–1935.

Hudson, R. R., and Kaplan, N. L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. Genetics **111**, 147–164.

Johnson, A. D., Wang, D. X., and Sadee, W. (2005) Polymorphisms affecting gene regulation and mRNA processing: Broad implications for pharmacogenetics. Pharmacol. Therapeut. **106**, 19–38.

Kohn, M. H., Fang, S., and Wu, C. I. (2004) Inference of positive and negative selection on the 5' regulatory regions of *Drosophila* genes. Mol. Biol. Evol. **21**, 374–383.

Kulkarni, M. M., and Arnosti, D. N. (2005) *Cis*-regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. Mol. Cell. Biol. **25**, 3411–3420.

Lemos, B., Meiklejohn, C. D., Caceres, M., and Hartl, D. L. (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories. Evolution **59**, 126–137.

Li, K. B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics **19**, 1585–1586.

Lopez Fanjul, C., and Toro, M. A. (2007) The effect of dominance on the use of the QST – FST contrast to detect natural selection on quantitative traits. Genetics **176**, 725–727.

Lopez-Fanjul, C., Fernandez, A., and Toro, M. A. (2003) The effect of neutral nonadditive gene action on the quantitative index of population divergence. Genetics **164**, 1627–1633

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**, 564–567.

Meiklejohn, C. D., Parsch, J., Ranz, J. M., and Hartl, D. L. (2003) Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. USA **100**, 9894–9899.

Merila, J., and Crnokrak, P. (2001) Comparison of genetic differentiation at marker loci and quantitative traits. J. Evol. Biol. **14**, 892–903.

Metta, M., Gudavalli, R., Gibert, J. M., and Schlotterer, C. (2006) No accelerated rate of protein evolution in male-biased *Drosophila pseudoobscura* genes. Genetics **174**, 411–420.

Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997) PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. **25**, 2745–2751.

Osada, N., Kohn, M. H., and Wu, C. I. (2006) Genomic inferences of the *cis*-regulatory nucleotide polymorphisms underlying gene expression differences between *Drosophila melanogaster* mating races. Mol. Biol. Evol. **23**, 1585–1591.

Pastinen, T., Sladek, R., Gurd, S., et al. (2004) A survey of genetic and epigenetic variation affecting human gene expression. Physiol. Genomics **16**, 184–193.

Phillips, P. C. (2005) Testing hypotheses regarding the genetics of adaptation. Genetica **123**, 15–24.

Prud'homme, B., Gompel, N., Rokas, A., et al. (2006) Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. Nature **440**, 1050–1053.

Purugganan, M., and Gibson, G. (2003) Merging ecology, molecular evolution, and functional genetics. Mol. Ecol. **12**, 1109–1112.

Ranz, J. M., and Machado, C. A. (2006) Uncovering evolutionary patterns of gene expression using microarrays. Trends Ecol. Evol. **21**, 29–37.

Ronald, J., Brem, R. B., Whittle, J., and Kruglyak, L. (2005) Local regulatory variation in *Saccharomyces cerevisiae*. Plos Genetics **1**, 213–222.

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

Shapiro, J. A., Huang, W., Zhang, C. H., et al. (2007) Adaptive genic evolution in the *Drosophila* genomes. Proc. Natl. Acad. Sci. USA **104**, 2271–2276.

Sokal, R. R., and Rohlf, F. J. (1995) Biometry: the principles and practice of statistics in biological research, 3rd edn. Freeman W. H. and Co, New York.

Takahashi, A., and Ting, C. T. (2004) Genetic basis of sexual isolation in *Drosophila melanogaster*. Genetica **120**, 273–284.

Tao, H., Cox, D. R., and Frazer, K. A. (2006) Allele-specific KRT1 expression is a complex trait. Plos Genetics **2**, 848–858.

Thornton, K. (2003) libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19**, 2325–2327.

Townsend, J. P., Cavalieri, D., and Hartl, D. L. (2003) Population genetic variation in genome-wide gene expression. Mol. Biol. Evol. **20**, 955–963.

Wang, D. Y., Sung, H. M., Wang, T. Y., et al. (2007) Expression evolution in yeast genes of single-input modules is mainly due to changes in *trans*-acting factors. Genome Res. **17**, 1161–1169.

Wang, H. Y., Fu, Y., McPeek, M. S., Lu, X., Nuzhdin, S., Xu, A., Lu, J., Wu, M. L., and Wu, C. I. (2008) Complex genetic interactions underlying expression differences between *Drosophila* races: Analysis of chromosome substitutions. Proc. Natl. Acad. Sci. USA **105**, 6362–6367.

Wang, J. L., Tian, L., Lee, H. S., et al. (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172**, 507–517.

Wang, R. L., Stec, A., Hey, J., Lukens, L., and Doebley, J. (2001) The limits of selection during maize domestication (vol 398, pg 236, 1999). Nature **410**, 718–718.

Wayne, M. L., Pan, Y. J., Nuzhdin, S. V., and McIntyre, L. M. (2004) Additivity and *trans*-acting effects on gene expression in male *Drosophila simulans*. Genetics **168**, 1413–1420.

Weber, A., Clark, R. M., Vaughn, L., et al. (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). Genetics **177**, 2349–2359.

Whitehead, A., and Crawford, D. L. (2006a) Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. USA **103**, 5425–5430.

Whitehead, A., and Crawford, D. L. (2006b) Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. **15**, 1197–1211.

Wittkopp, P. J. (2005) Genomic sources of regulatory variation in *cis* and in *trans*. Cell. Mol. Life Sci. **62**, 1779–1783.

Wittkopp, P. J. (2006) Evolution of *cis*-regulatory sequence and function in Diptera. Heredity **97**, 139–147.

Wittkopp, P. J. (2007) Variable gene expression in eukaryotes: a network perspective. J. Exp. Biol. **210**, 1567–1575.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004) Evolutionary changes in *cis* and *trans* gene regulation. Nature **430**, 85–88.

Wray, G. A. (2003) Transcriptional regulation and the evolution of development. Intl. J. Dev. Biol. **47**, 675–684.

Wray, G. A., Hahn, M. W., Abouheif, E., et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**, 1377–1419.

Wright, S. I., Bi, I. V., Schroeder, S. G., et al. (2005) The effects of artificial selection of the maize genome. Science **308**, 1310–1314.

Wu, C. I., and Ting, C. T. (2004) Genes and speciation. Nat. Rev. Genet. **5**, 114–122.

Yan, H., and Zhou, W. (2004) Allelic variations in gene expression. Curr. Opin. Onc. **16**, 39–43.