Complex genetic interactions underlying expression differences between *Drosophila* races: Analysis of chromosome substitutions

Hurng-Yi Wang[†‡], Yonggui Fu[§], Mary Sara McPeek[¶], Xuemei Lu[§], Sergey Nuzhdin[‖], Anlong Xu[§], Jian Lu[††], Mao-Lien Wu[††], and Chung-I Wu[‡§††]

[†]Graduate Institute of Clinical Medicine, National Taiwan University, 7 Chung-Shan South Road, Taipei 100, Taiwan; [§]State Key Laboratory for Biocontrol, Department of Biochemistry, College of Life Sciences, and International Center for Evolutionary and Genomic Studies, Sun Yat-Sen University, 135 Xingang Xilu, Guangzhou, Guangdong 510275, People's Republic of China; Departments of [¶]Statistics and [††]Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637; and [‖]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Edited by Daniel L. Hartl, Harvard University, Cambridge, MA, and approved February 26, 2008 (received for review December 20, 2007)

Regulation of gene expression is usually separated into *cis* and *trans* components. The separation may become artificial if much of the variation in expression is under multigenic and epistatic (e.g., *cis*-by-*trans*) control. There is hence a need to quantify the relative contribution of *cis*, *trans*, and *cis*-by-*trans* effects on expression divergence at different levels of evolution. To do so across the whole genome, we analyzed the full set of chromosome-substitution lines between the two behavioral races of *Drosophila melanogaster*. Our observations: (*i*) Only ≈3% of the genes with an expression difference are purely *cis* regulated. In fact, relatively few genes are governed by simple genetics because nearly 80% of expression differences are controlled by at least two chromosomes. (*ii*) For 14% of the genes, *cis* regulation does play a role but usually in conjunction with *trans* regulation. This joint action of *cis* and *trans* effects, either additive or epistatic, is referred to as inclusive *cis* effect. (*iii*) The percentage of genes with inclusive *cis* effect increases to 32% among genes that are strongly differentiated between the two races. (*iv*) We observed a nonrandom distribution of *trans*-acting factors, with a substantial deficit on the second chromosome. Between *Drosophila* racial groups, *trans* regulation of expression difference is extensive, and *cis* regulation often evolves in conjunction with *trans* effects.

Knowledge of the genetics of complex traits is fundamental to modern medicine, agriculture, and evolutionary biology. Among all complex traits, gene expression as phenotype may be most amenable to genetic analysis. The first question about expression regulation naturally is whether there is a *cis* component and how strong the *cis* component is. (*cis* regulation refers to the control of expression by the gene itself whereas *trans* regulation refers to the influence of the genetic background.) Many studies have addressed this question at various levels of divergence (1–9). For example, one may measure the expressions of two alleles at the locus of interest in a common genetic background (usually F_1s) (1, 2, 5). Because the collection of *trans*-acting factors in the same cellular environment is assumed to affect the two alleles equally, asymmetric allelic expression implies differences due to *cis*-regulatory divergence. Similarly, expression quantitative trait loci (eQTL) mapping permits inference of *cis* regulation if the eQTL is mapped closely to the expressed gene itself (6–10).

A second question is how strong *cis* regulation is relative to *trans* regulation. In the extreme case where most expression variation is controlled by *cis*–*trans* interactions (e.g., joint actions of *cis* elements and transcription factors), the question would not be very meaningful because *cis* and *trans* components are both indispensable. It is desirable to explicitly model expression regulation to include *cis*, *trans*, and *cis*-by-*trans* control. Many kinds of data allow such explicit modeling. The use of large numbers of recombinant strains for expression analysis is a common approach (6, 7). Such an analysis, however, requires extensive genotyping on many recombinant lines, and each line may receive only limited attention (see *Discussion* for detail). A simpler alternative is to construct a complete set of whole-chromosome substitution lines, which make it possible to measure gene expression repeatedly under the control of the same chromosome combinations.

The genetic architecture, specifically the relative contribution of *cis* and *trans* regulation, may also depend on the extent of divergence between the subjects under comparison. Between two randomly chosen genomes from the same population, it is possible that the 5′-regulatory regions have not differentiated much, and the *cis* component may be small. When two genomes have diverged, both *cis* and *trans* effects would become greater, and their relative magnitude becomes less predictable.

Clearly, the level of divergence should be an important parameter in the analysis of *cis* vs. *trans* effects. However, genetic differences between species generally make their hybrids unhealthy, resulting in difficulties in chromosome replacement. On the other hand, individuals or lines from same population may be too similar to have much *cis* effect. Populations that have shown some degree of phenotypic divergence provide a reasonable balance. The Zimbabwe race of *D. melanogaster* vis-à-vis the cosmopolitan populations are thus well suited to such a purpose. Females from Z lines (for Zimbabwe) do not readily mate with males from M lines (for melanogaster of the cosmopolitan type), whereas the reciprocal crosses experience much weaker or no discrimination (11).

In this study, we used a standard M (Fr) and Z line (Z30). These two lines have been extensively analyzed for their behavioral, physiological, and molecular divergence (11–18). The purpose of this study is to dissect the genetic architecture of transcription regulation. By comparing the expression differences between the two parental (Fr and Z30) and six substitution lines [constructed by Hollocher *et al*. (11)], we were able to determine the contribution of *cis*, *trans*, and *cis*-by-*trans* components to gene expression.

Results

The experimental design is given in Fig. S1. We shall designate the two pure lines (Fr and Z30) and the six whole-chromosome substitution lines as ZZZ (i.e., Z30), MZZ, ZMZ, ZZM, MMZ, MZM, ZMM, and MMM (Fr), respectively. Each letter in this triplet notation designates the origin of the X, II, and III chromo-

Author contributions: H.-Y.W. and Y.F. contributed equally to this work; H.-Y.W., Y.F., and C.-I.W. designed research; H.-Y.W., Y.F., X.L., and M.-L.W. performed research; X.L., A.X., and C.-I.W. contributed new reagents/analytic tools; H.-Y.W., Y.F., M.S.M., S.N., J.L., and C.-I.W. analyzed data; and H.-Y.W., Y.F., M.S.M., and C.-I.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[‡]To whom correspondence may be addressed. E-mail: hurngyi@ntu.edu.tw or ciwu@uchicago.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0711774105/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0711774105

some. The IV chromosome is not monitored and is a mixture of Fr and Z30 when the substitution lines were constructed. Note that the three major chromosomes in each substitution line were extracted as a whole without recombination [see Hollocher *et al.* (11)].

After image analysis, background correction, and normalization, the log-2 expression intensities were analyzed by using a linear mixed model with fixed effect for different fly lines and a random effect for arrays. The residuals from the normalization model were subjected to gene-specific models of the form $r_{ijk} = \mu + A_i + D_j + G_k + \varepsilon_{ijk}$, where G_k is kth genotype (k = 1–8) and will be specifically modified in each section below.

We first identified genes that are differentially expressed between ZZZ and MMM by *t* test. Four kinds of *t* tests were used, and we considered only genes that pass all four *t* tests. The genetic control is then analyzed by studying the expression in the six substitution lines. At a false discovery rate (FDR) of 0.01, 188, 493, 534, and 15 genes on the X, II, III, and IV chromosome, respectively, were found to be differentially expressed between MMM and ZZZ (Dataset S1). These genes are distributed among the four chromosomes much like the rest of the genome ($P = 0.23$ by χ^2 test). The mean expression ratio of the more highly expressed line over the less highly expressed line is 1.69. We did not include the 15 genes on the IV chromosome, because the substitution lines could not be identified by their IV chromosome origin.

Average Effects Across All Genes. The full model for the genetic control of expression differences is given as

$$r_{ijk} = \mu + A_i + D_j + \mathrm{Chr}_{\mathrm{X}} + \mathrm{Chr}_{\mathrm{II}} + \mathrm{Chr}_{\mathrm{III}} + \mathrm{Chr}_{\mathrm{X,II}} + \mathrm{Chr}_{\mathrm{X,III}} + \mathrm{Chr}_{\mathrm{II,III}} + \mathrm{Chr}_{\mathrm{X,II,III}} + \varepsilon_{ijk}, \qquad [1]$$

where A_i and D_j represent effect of spot on the array and dye effect, respectively.

With three chromosomes, the expression of each gene can potentially be explained by seven different terms—the three single-chromosome effects ($\mathrm{Chr}_{\mathrm{X}}$ etc), the 3 interaction terms between any 2 of the 3 chromosomes ($\mathrm{Chr}_{\mathrm{X,II}}$, etc.) and the highest level interaction with all three chromosomes. To apply this model, we took two different approaches. In the first approach, we computed the proportion of variance contributed by each of the seven terms, averaged across all 1,215 genes (n = 188 + 493 + 534, see above). In the second approach, we examined the relative contribution of each term for each individual gene. Although we are mainly interested in the genetic control for each gene, the average patterns across loci, which usually have better statistical support, may provide the corroborative evidence to the gene-by-gene analysis.

The average proportions of variances contributed by the seven terms are given in Table 1. These proportions are presented for genes on each chromosome separately as well as jointly. Note that the four interactive terms collectively account for 32.3% (equal to 27.1% + 5.19%) of the total variance. Because the analysis includes interactions between genes of the same chromosome in the category of single-chromosome (noninteractive) effect, the actual level of genetic interaction underlying expression variation should be >32.3%. (Assuming the number of interacting genes is proportional to chromosome size, we estimate that the within-chromosome interactions might augment the reported 32.3% to nearly 50%.) Given that the contribution of interactive terms to expression variation is nonnegligible, the distinction between *cis* and *trans* regulation may not always be straightforward.

Testing Models of Pure *Cis and Pure *Trans** Control.** In this section, each gene was individually analyzed. We first tested the two simplest models of genetic control. In the pure *cis* model, the expression level is fully controlled by the gene itself with no influence from the genetic background. In the pure *trans* model, the expression is controlled entirely by the genetic background with no input from the expressed gene.

We distinguish between chromosomal *cis* effect and genic *cis* effect. Because the genetic resolution is at the whole-chromosome level, we use the notations, *cis** and *trans** effect, for the whole-chromosome analysis in the statistical models. *Cis** effect is the combined effect of the (true) genic *cis* effect and all *trans* effects of genes on the same chromosome of the expressed locus. *Cis** effect is an overestimate of the true *cis* effect, and *trans** effect does not include same-chromosome *trans* effects. In Table 2, we present the *cis** and *trans** effect; the true *cis* effect is estimated in Table 3.

A likelihood ratio test was preformed for the pure *cis**-effect model, which is $r_{ijk} = \mu + A_i + D_j + \mathrm{Chr}_l + \varepsilon_{ijk}$, where l is the chromosome from which the expressed gene is located (l = X, II, or III). The alternative model expands the "Chr" term above into the full model of Eq. **1**. above. In Table 2, it is shown that the *cis**-effect model is accepted over the alternative full model for only a small fraction of genes. At the nominal acceptance rate of 5%, the percentages of genes regulated by a complete *cis** effect on X, II, and III chromosome are 6.9%, 2.4%, and 10.1%, respectively, with an average of 6.5%. Given that the *cis** effect (at only 1.5% above the nominal rate of 5%) is still an overestimate of the genic *cis* effect, it seems safe to conclude that not many genes between the Z and M lines are purely *cis* regulated. As a corollary, 93.5% of the genes must have a *trans* component in their regulation.

We next test the complete *trans**-effect model in which gene expression is controlled exclusively by *trans* chromosomes (e.g., the II and III chromosomes for X-linked genes). The null model of this test is $r_{ijk} = \mu + A_i + D_j + \mathrm{Chr}_m + \mathrm{Chr}_n + \mathrm{Chr}_{m,n} + \varepsilon_{ijk}$, where m and n are the two *trans* chromosomes (m, n = X, II, III). The

Table 1. The proportions of the total genetic variances in expression differences explained by each of the seven terms (the first column) of Eq. 1

	Gene location				
Chromosomal effect	X, % (n = 188)	II, % (n = 493)	III, % (n = 534)	All, % (n = 1,215)	Genic *cis* effect, %
X	30.33 (1.79)	21.14 (0.82)	20.56 (0.81)	22.31 (0.57)	9.49 [8.08] {14.4}
II	6.72 (0.87)	12.96 (0.81)	8.04 (0.52)	9.83 (0.43)	5.27 [4.9] {5.41}
III	32.00 (1.63)	32.69 (0.98)	39.55 (1.06)	35.60 (0.67)	7.05 [6.04] {11.2}
Subtotal				67.74 (0.51)	6.70 [5.95] {8.83}
(X, II)	5.11 (0.59)	5.55 (0.35)	5.32 (0.32)	5.38 (0.22)	
(X, III)	17.39 (1.2)	19.36 (0.74)	17.56 (0.65)	18.26 (0.46)	
(II, III)	3.17 (0.36)	3.36 (0.23)	3.56 (0.24)	3.42 (0.15)	
Subtotal				27.1 (0.47)	
(X, III, III)	5.27 (0.55)	4.94 (0.36)	5.41 (0.40)	5.19 (0.24)	

These proportions are averaged across all genes on X, II, and III and all chromosomes as shown in columns 2–4. Standard errors are given in parentheses. The last column gives the proportions of variances explained by the genic *cis* effect (see *Results*). The genic *cis* effects are separately estimated for two groups of genes—non-race-differentiating and race-differentiating genes (see *Results*). The estimates are shown in brackets and braces for the two groups, respectively

Table 2. Number of genes on each chromosome that are subjected to pure *cis and pure *trans** regulation at the 5% cutoff**

	Gene location			
Cis/*trans*	X (n = 188)	II (n = 493)	III (n = 534)	Total (n = 1,215)
*Cis** effect only (no *trans** effect)	13 (6.9)	12 (2.4)	54 (10.1)	79 (6.5)
Trans effect present	>175 (>93.1)	>481 (>97.6)	>480 (>89.9)	>1,136 (>93.5)
*Trans** effect only (no *cis** effect)	14 (7.4)	183 (37.1)	24 (4.5)	221 (18.2)

* indicates whole-chromosome effect. Data in parentheses are percentages.

alternative is again the full model described above. At $P = 0.05$, the percentages of genes regulated entirely by the *trans** effect are 7.4%, 37.1%, and 4.5% for the X, II, and III chromosome, respectively (Table 2). Because the sums of the first and third row of Table 2 are far <100%, the bulk of genes must be jointly regulated by *cis* and *trans* chromosomes.

A surprising finding in Table 2 is that genes on chromosome II appear to be subjected to weak *cis** control and relatively strong *trans** control. Given the near parity in gene number between the two autosomes, such a chromosome-wide difference in transcription regulation demands an explanation. The answer appears to be a difference in *trans*-regulating capacity for genes differentially expressed between Z and M across chromosomes (see below).

Selecting the Best Model for Each Gene. The criteria for pure *cis** and pure *trans** effect by the likelihood ratio test above may be too stringent to have much explanatory power. An alternative is to search for the best model for each gene by applying the backward selection method with Akaike information criterion (AIC) (19). In this analysis of model selection, each of the seven terms of genetic effect in Eq. **1** is sequentially removed according to AIC, until further removal resulted in worse fit. Different search criteria including Bayesian information criterion (BIC) and likelihood ratio test (LRT) were also used for model selection. Because the results from these methods are in qualitative agreement, we shall present only the AIC model selection in some detail. The results from BIC and LRT are in Table S1 and Table S2.

In our analysis, 77 genes could not be unambiguously assigned to a specific model and were removed from further analysis. Among the 1,138 genes analyzed, the best model is a single-chromosome effect model for 21.4% of the genes (Table 3). In these cases, only one of the three terms, $\mathrm{Chr}_{\mathrm{X}}$, $\mathrm{Chr}_{\mathrm{II}}$, or $\mathrm{Chr}_{\mathrm{III}}$, is needed in the selected model. A more complex two-chromosome model is as follows: All models that include the $\mathrm{Chr}_{\mathrm{X,II}}$ term or contain both the $\mathrm{Chr}_{\mathrm{X}}$ and $\mathrm{Chr}_{\mathrm{II}}$ terms are grouped into the X+II two-chromosome category, provided that all other terms containing the chromosome III effect are not needed. A parallel procedure is used for X+III and for II+III, respectively. Among these categories, 28.6% of the genes can be explained by the control of two different chromosomes (that act either independently or jointly). All three major chromosomes are necessary to explain the expression differences for 50% of the genes. Nearly 80% of the genes are regulated by two or three different chromosomes for their expression.

Note that Tables 1 and 3 present very different information. For example, (X,II) in Table 1 is for the interactive term only, whereas "X and II" in Table 3 indicates the presence of effects from both X and II, which include both interactive and additive terms. In Table 1, the average contribution of a single chromosome ranges from 6.7% to nearly 40%. It is therefore not unreasonable to see only 21.4% of genes whose expression variation is entirely attributed to the effect of one single chromosome.

Multichromosome control can be either additive or epistatic. In Table 3, 326 genes are controlled by two chromosomes. Among them, 241 (73.3%) have the interactive term, $\mathrm{Chr}_{\mathrm{X,II}}$, $\mathrm{Chr}_{\mathrm{X,III}}$, or $\mathrm{Chr}_{\mathrm{II,III}}$ of Eq. **1**. Of 569 genes under the control of all three chromosomes, 560 (98.3%) of them have at least one interactive term. In total, 70.3% [(241 + 560)/1,138] of the genes in Table 3 are influenced by epistatic interactions. Because we cannot account for epistatic interactions between genes on the same chromosome, the percentage 70.3% could be a conservative estimate. The common occurrence of epistatic interactions underlying gene expression again suggests the limitation of a strict *cis* vs. *trans* dichotomy in describing the genetic control of gene expression.

In comparison with the results of Table 3, Table 1 shows that the average contribution of the epistatic component to the total variance of expression across all loci is ≈32%, as mentioned above. Thus, whereas the average epistatic effect is 32%, for nearly 30% (100% − 70.3%) of the genes, the epistatic effect is statistically indistinguishable from 0%. The distribution of the epistatic effect among genes appears fairly broad.

Estimating the Pure *Cis* Effect at the Genic Level. The proportion of genes controlled solely by the *cis** chromosome is estimated to be 9.6% (Table 3 and column 1 of Table 4). Because the chromosomal *cis** effect includes both the genic *cis* effect and intrachromosomal *trans* effect, the estimate of 9.6% is most likely an overestimate of the genic *cis* effect. (Without *, *cis* denotes genic *cis* effect.) The

Table 3. Chromosomal effect on differently expressed genes between MMM and ZZZ

Chromosomal	Gene location			Subtotal	*Trans*-acting effect
effect	X (n = 170)	II (n = 466)	III (n = 502)	(n = 1,138)	per 2,000 target genes
X	24 [24]	22 [9.7]	34 [12.7]	243 (21.4)	11.2
II	5 [5]	18 [8.0]	5 [1.9]		3.4
III	18 [18]	50 [22.1]	67 [25.1]		20.1
X and II	11	23	22	326 (28.6)	8.2
X and III	38	89	92		39.4
II and III	5	16	30		5.0
All three chromosomes	69	248	252	569 (50.0)	

The numbers of such genes, classified by the chromosomal effect and gene location, are given. The numbers in brackets are adjusted for the size of the chromosome where the regulated genes are located. They correspond to the numbers of differentially regulated genes per 2,000 genes. The numbers of genes on the microarrays from the X, II, and III chromosome are, respectively, 1,992 (rounded off to 2,000), 4,509, and 5,324. The data in parentheses are percentages.

Table 4. The number of genes under *cis* control

Chromosome	*Cis** effect only, *n* (%)	Genic *cis* effect only, *n* (%)	Inclusive *cis** effect, *n* (%)	Inclusive genic *cis* effect, *n* (%)
X (*n* = 170)	24 (14.1)	12.8 (7.5)	142 (83.5)	19.5 (11.4)
II (*n* = 466)	18 (3.9)	10.3 (2.2)	305 (65.4)	77.0 (16.5)
III (*n* = 502)	67 (13.3)	13.5 (2.7)	441 (87.8)	61.1 (12.2)
Total (*n* = 1,138)	109 (9.6)	36.6 (3.2)	888 (78.0)	157.7 (13.9)

"*Cis** effect only" is for the genes regulated only by the *cis* chromosome. "Inclusive *cis** effect" indicates the involvement of the *cis* chromosome in gene regulation, either singularly (*cis** effect only) or jointly. The genic *cis* effect is estimated by subtracting the estimated number of genes under intrachromosomal *trans* control (see *Materials and Methods*) from the number under *cis** control. Genes controlled only by *cis* effect (Genic *cis* effect) were estimated only from a single chromosome. "Inclusive genic *cis* effect" was estimated from all the terms that involves the *cis* chromosome.

question is how many genes, if any, would show a genic *cis* effect if the overestimation is removed. In this section, we will estimate the proportion of genes that are controlled solely by the *cis* element. This proportion is of course no more than 9.6%. (In the next section, we will estimate the proportion of genes that show a genic *cis* effect, accompanied by some *trans* effects in the background.)

To estimate the genic *cis* effect, we need to evaluate the intrachromosomal *trans* effect and subtract this effect from the whole-chromosome *cis** effect. We hence compared the *trans*-acting effects of each chromosome on the other two chromosomes and on itself. By doing so, we would know whether intra- and interchromosomal *trans* effects are comparable. In Table 1, the *trans* effect of X on the II chromosome accounts for 21.14% of the variance. The corresponding percentage for the III chromosome is 20.56%. The two numbers are close with an average of 20.84%, suggesting that the *trans*-regulating strength of X is similar across target chromosomes. For chromosomes II and III, the two interchromosomal *trans* effects are also very close, $\approx$7.7% and 32.5%, respectively. Thus, each chromosome's *trans* effects on the other two chromosomes are rather consistent.

We cannot directly evaluate the intra- vis-à-vis interchromosomal *trans* effect but indirect evidence can be obtained as follows. *Trans*-acting factors may affect the expression of a gene by interacting with either its 5′ or 3′ end. On the 5′ end, transcription factors (TFs) interact with their binding sites to regulate the level of expression. On the 3′ end, miRNAs interact with their target sites posttranscriptionally, often resulting in the degradation of transcripts (20). We searched for transcription factor-binding sites (TFBS) that have been experimentally verified. Table S3 shows that TFBSs appear evenly distributed among *trans* and *cis* chromosomes. Furthermore, using the TARGETSCAN algorithm (21), we compiled and separated microRNA targets that are located on the same chromosome as the microRNAs from other predicted targets. The chromosomal distributions of the microRNAs and their targets also appear random with respect to chromosomal origin (see Fig. S2). In short, for the two major classes of *trans*-acting factors (TFs and microRNAs), the assumption that intra- and interchromosomal *trans* effects are comparable seems reasonable.

Although the targets appear to be independent of the chromosomal location of the *trans*-acting factors, these factors themselves may not be randomly distributed because there is a large difference in the average *trans* effect of each chromosome. The difference is particularly notable between the two autosomes. Although the II and III chromosomes are roughly of the same size, the proportion of variance contributed by the latter is >3-fold as high as that contributed by the former. In contrast, the X chromosome contributes as much as 2/3 of that of chromosome III, whereas it is slightly greater than half the size of the latter.

A most interesting feature of Table 1 is the difference between *cis** and *trans** effects. For X, the *cis** effect is 30.33%, whereas the two *trans** effects are 21.14% and 20.56%, respectively. The difference between the *cis** effect and the average *trans** effect, at 9.49%, is highly significant ($P < 10^{-8}$). In the last column of Table 1, we refer to this difference as "genic *cis* effect". The patterns are also true for the two autosomes with the genic effects are 5.27% and 7.05%, respectively ($P < 10^{-6}$). We estimate the genome-wide genic *cis* effect at $\approx$6.7%. Again, the contribution of pure *cis* effect to expression variation is quite modest, leaving substantial room for *trans* effects and for *cis*-by-*trans* interactions.

The similar procedure may now be applied to estimating the number of genes that are under pure *cis* regulation (or those that have a *cis* component in the next section). In Table 3, when the number of genes with expression difference on each chromosome is adjusted for the size of chromosome, the *cis* effect is also detectable. The effect of X on X is seen in 24 genes (per $\approx$2,000 genes), but in only 9.7 or 12.7 genes for the effect of X on II or X on III, respectively. The average of the two latter numbers is considered the *trans*-acting effect of X. In the last column of Table 3, the *trans*-acting effects per 2,000 target genes are shown to be 11.2, 3.4, and 20.1 for X, II, and III, respectively. Each chromosome's *trans* effect appears comparable on different target chromosomes, but there is a large difference in the average *trans* effect of each chromosome. Adjusted for its size, chromosome II is deficient in its *trans*-acting effect in comparison with the other two.

The number of differentially expressed genes on each chromosome that are under *cis** effect is reproduced in column 1 of Table 4. The estimated number of genes with functional *cis*-regulatory polymorphisms is thus the difference between the number in column 1 of Table 4 and the estimated number of genes under same-chromosome *trans* control (see *Materials and Methods*). For example, the number for the X chromosome is 12.8 (24 − 11.2), or 7.5%, as shown in column 2 of Table 4. For II and III, the proportions are 2.2% and 2.7%, respectively. Among all genes that are differentially expressed between the Z and M genomes, only 3.2% are purely *cis* regulated (genic *cis* effect only).

Estimating the Inclusive *Cis* Effect at the Genic Level. In the Z-M system, the proportion of differently expressed genes that are purely *cis* regulated, at 3.2%, is almost negligible. A more useful estimate may be the proportion of genes with a detectable *cis* effect regardless of the presence or absence of other effects. We suggest the term "inclusive *cis** effect," which is defined as follows. The inclusive *cis** effect for any X-linked gene is the summation of terms that include X. In other words, the effect is defined as including any combination of the following four terms in the selected model: $\mathrm{Chr}_{\mathrm{X}} + \mathrm{Chr}_{\mathrm{X,II}} + \mathrm{Chr}_{\mathrm{X,III}} + \mathrm{Chr}_{\mathrm{X,II,III}}$. We define "inclusive *cis* (genic) effect" similarly; i.e., the summation of all genetic effects that involves the gene itself.

We first estimate the number of X-linked genes with an inclusive *cis** effect, which is 142 (24 + 11 + 38 + 69 from Table 3), shown in the row of X in Table 4. These are cases where the X chromosome plays a role (singly or jointly with other chromosomes) in the expression of X-linked genes. The same procedure is used for other chromosomes. On average, 78% of the expression differences have an inclusive *cis** effect, and chromosome II indeed has the lowest percentage of such genes (column 3 of Table 4). Like the estimation of genic *cis* effect shown in column 2 of Table 4, we extrapolate the interchromosomal *trans* effect (which is observable) to estimating the intrachromosomal *trans* effect. We use X-linked genes as an example. We need to estimate the *cis* effect of x, x+II, x+III, and x+II+III, where the lowercase x denotes individual X-linked genes (rather than the whole X chromosome). The *cis* effect of x, in terms of the number of genes affected, has been shown to be 12.8. The *cis* effects of x+II, x+III, and x+II+III are estimated to be 2.76, 0, and 3.9, respectively (see *Materials and Methods*). The inclusive *cis* effect for X-linked genes is therefore 19.5 (12.8 + 2.76 + 0 + 3.9), or

11.4% of the affected genes. Across all chromosomes, 13.9% of genes have an inclusive *cis* effect (column 4).

In summary, only 3.2% of differentially expressed genes are purely *cis* regulated but, when inclusive *cis* effect is considered, 13.9% are *cis* regulated. This comparison means that (*i*) *trans* effect dominates the regulation of gene expression at this level of divergence, and (*ii*) when there is *cis* effect, *cis* regulation often occurs in a genetic background where *trans* regulation is also in operation. The conclusion is consistent with the estimate of the average genic *cis* effect across all genes, which is 6.7% of the total variation, as shown in Table 1.

Cis Effect in Race-Differentiating vs. Non-Race Differentiating Genes.

We now address the issue whether the relative strength of *cis* regulation increases as the level of divergence increases. Among the 1,138 genes of Table 3, 238 are considered the more divergent group by the following criteria. In a separate study using multiple M and Z lines, 421 genes were labeled "race-differentiating" (Y.F. and H.-Y.W., unpublished work). For each of these 421 genes, the mean expression of the 24 M lines is significantly different from that of the 15 Z lines at FDR <0.01. Among the 421 genes, 238 are also represented in the set of 1,138 genes analyzed in Table 3 and are referred to as "race-differentiating" genes. The remaining 900 genes are "non-race-differentiating" because their expressions are different both between the Fr (the MMM line) and Z30 (ZZZ) lines and among different M and different Z lines.

We computed the average genic *cis* effects separately for the two groups of genes, as shown in Table 1. When all genes are considered, genic *cis* effect accounts for 6.7% of the expression variation. This proportion increases to 8.83% for race-differentiating genes and decreases to 5.95% for non-race-differentiating genes. The difference is 2.88% ($P < 0.05$, two-tailed t test).

In Table S4, we again analyzed each gene individually. For race-differentiating genes, the percentages of genes with an inclusive *cis* effect are 45.1%, 32.3%, and 29.6% for X, II, and III, respectively; the average is 32.0%. For "non-race-differentiating" genes, the percentages of genes with a significant inclusive *cis* effect are 14.3%, 15.3%, and 10.9% for X, II, and III, respectively; the average is 13.2%. The proportion of genes that have a *cis*-regulatory component is >2.5 times higher among "race-differentiating" genes than among the set that are not race-differentiating. Moreover, the magnitude of changes for race-differentiating genes is generally higher than for the non-race-differentiating ones. The mean fold differences are 2.32 for the former and 1.59 for the latter.

Discussion

By analyzing the complete set of chromosome substitution lines between the M and Z races, we have been able to quantify the relative contribution of *cis* and *trans* regulation as well as the interaction between the two components. The limitation of this study is that it could only estimate the proportion of *cis*-regulated genes without knowing their identity. A number of previous studies (6, 7, 22), especially those of eQTL analysis, have addressed similar questions. In this study, a full set of whole-chromosome substitution lines, in all possible combinations and kept as permanent stocks, permit repeatable analysis to obtain quantitative estimates of *cis*, *trans*, and *cis*-by-*trans* effects. Given the reproducibility, we were able to tease apart the within-chromosome *trans* effects and true genic *cis* effects (see *SI Materials and Methods*). Previous studies using random F2 segregants (9, 10) or F1 hybrids (2, 23) did not systematically quantify *cis* vs. *trans* effects and rarely addressed the magnitude of the interactive *cis*-by-*trans* terms.

At the level of divergence analyzed in this study (between racial groups of the same species), the genetic control of expression is already quite complex, and the contribution of pure *cis* effect is very modest. Multigenic control, of which *cis*-by-*trans* interaction is a most interesting example, appears to be the rule. In the context of complex genetics, the inclusive *cis* effect may be a useful concept. A gene is considered as having an inclusive *cis* effect if the *cis* factor contributes to the expression variation, regardless of whether it exerts the effect singly or jointly with *trans* factors.

Seemingly incongruent results from different studies of expression regulation may be resolvable if the distinction between *cis* and inclusive *cis* effects is made. Osada *et al.* (13) used the same set of lines as ours and found the difference in gene expression to be correlated with the presence of the *cis* chromosome (see figure 1 in ref. 13). They concluded that *cis** effect (using the new notation) is common at this level of divergence. There may appear to be a discrepancy between our conclusion and theirs because strict *cis** effect is detected in only 9.6% of all differentially expressed genes in Table 4. However, because Osada *et al.* measured the inclusive effect, the discrepancy disappears if we note that 78% of the genes in Table 4 have an inclusive *cis** effect.

Likewise, the analysis of expression in the F1 progeny [both between and within species (2, 23)] may be primarily about the inclusive *cis* effect. This seems to be a plausible suggestion because both *cis* and *trans* elements from both parents are present in the F_1 progeny. Again, the extent of *cis* regulation in these reports appears quite substantial. In comparison, our results (Table S4) suggest that 32% of the race-differentiating genes have an inclusive *cis* effect. Because the level of *cis* regulation in this study appears lower than the studies of Wittkopp *et al.* (2, 23), which focused on the more strongly divergent genes, the differences might suggest a possible correlation between expression divergence and the strength of inclusive *cis* effect. As the *cis*-regulatory sites become more and more divergent, the majority of differently expressed genes can be expected to have a measurable *cis* effect. This has been reported (23), and our results (Table S4) also support such an interpretation. Other studies have suggested various degrees of *cis* regulation (1–3, 6, 8, 23, 24). For example, in the analysis of genome-wide variation in human gene expression, Morley *et al.* (6) concluded that 19% of 142 genes have only a *cis*-acting transcriptional regulator. By integrating transcriptional profiling and linkage analysis, Hubner *et al.* (7) have shown that 35–40% of the eQTLs between two mouse strains were regulated *in cis*. In yeast the proportion of *cis* eQTL from different studies ranges from 10% to 25% (1, 9, 25). The differences may, to some degree, depend on the level of divergence analyzed. (The distinction, or lack thereof, between pure *cis* and inclusive *cis* effects also contributes to the variation in the reported estimates.)

As the level of divergence increases from, say, within-species variation to between-species divergence, the strength of inclusive *cis* effect is expected to increase. Nevertheless, the demarcation between *cis* and *trans* effects may become less clear-cut as the interactions between *cis* and *trans* factors become more prevalent. We have shown in Table 1 that the contribution of interactive terms to expression variation between populations is >30% and can be as large as 50%. In the analysis of interspecific F_1 hybrids, Landry *et al.* (5) have shown that 13 of the 23 misexpressed genes exhibited *cis*-by-*trans* interaction. Furthermore, *cis* and *trans* factors can interact in opposition to each other. We have observed substantial counterbalancing effects in gene expression between Z and M races (H.-Y.W., unpublished results). Although strong epistasis between genic actions seems complex, it is hardly surprising. Indeed, numerous introgression studies have suggested that strong epistatic interactions underlie fitness-associated characters between races and species [see Wu and Palopoli (26); Wu and Ting (27) for reviews].

A surprising finding in this study is the much weaker *trans*-acting effect of chromosome II, relative to that of the X chromosome or chromosome III. Because X and autosomes have many different evolutionary characteristics (28–31), the X–II difference may have many explanations. However, the strong contrast between the two autosomes indeed demands an explanation. A suggestion may be that the divergence in the transcription factors (TFs) or *trans*-acting components on chromosome II is less than that of TFs on chromosome III, because the Z and M races separated. Hollocher *et al.* (11) and Ting *et al.* (14) have reported similar chromosome-wide differences. Between the Z and M races, the contribution to behavioral differentiation is substantially larger for chromosome III than for chromosome II.

It has been known that many phenotypic variations between species are subjected to complex genetic control (26, 27). An important reason for studying the control of gene expression is that it may ultimately inform about the genetics of these "classical" phenotypes, including behavior, disease resistance, reproductive success, and so on. The regulation of expression divergence is indeed complex even at the incipient stage of speciation between the Z and M races. Nevertheless, we can often detect inclusive *cis* effect, which can then be a link between expression phenotypes and (part of) the genetic circuitry.

Materials and Methods

Fly Strains. One isofemale M line (Fr, hereafter referred as MMM) and one isofemale Z line (Z30, hereafter referred as ZZZ), both of which are commonly used as standards to study mating behaviors (32), and six chromosome substitution lines MZZ, ZMZ, ZZM, MMZ, MZM, and ZZM (11), were used. For chromosome substitution lines, M indicates that a homozygous X, second, or third chromosome is derived from the MMM (Fr) line, and Z indicates the corresponding homozygous chromosomes derived from the ZZZ (Z30) line. The origin of fourth chromosome is ignored in this setting.

Microarray Hybridization. We designed the hybridization scheme to efficiently estimate differences between MMM and ZZZ lines and chromosome effects (Fig. S1). Each of six chromosome substitution lines was measured four times, and two parental lines were measured 14 times. One hundred to 150 3- to 5-day-old male flies were starved for 1 hour before sedation on ice. Heads were dislodged from the bodies by using sieves after flies were snap-frozen in liquid nitrogen TRIzolR reagent (GIBCO–BRL) was used for RNA extraction. After precipitation, the RNA was then resuspended and purified with phenol and chloroform. cDNA syntheses were carried out by Array 900 expression array detection kit (Genisphere) with manufacture's protocol. Arrays were from DGRC (the Drosophila Genomics Resource Center) of Indiana University (Bloomington, IN). We hybridized and washed the samples according to the protocol from DGRC (33).

Array Analyses. Twenty-six arrays were scanned by using a GenePix Axon scanner, and data were extracted by using GenePix 6 to give Cy3 and Cy5 intensities. The flowing analyses were conducted in the R computing environment (www.r-project.org). We first transformed the data using spatial-intensity joint loess. This was done with the R/maanova package (34). The transformed $\log_2$ intensities for all 15,552 spot measures y_{ijk} were subjected to a normalization model of the form $y_{ij} = \mu + A_i + D_j + AD_{ij} + \varepsilon_{ij}$, where μ is the sample mean, A_i is the effect of the *i*th array (i = 1–26), D_j is the effect of *j*th dye (Cy3 or Cy5), AD_{ij} is the interaction involving dye and array, and $\varepsilon_{\mathrm{ij}}$ is the stochastic error. In this model, array (A_i) was modeled as random effect and was assumed to have a normal distribution with a mean of zero [$N(0,\sigma^2_A)$].

Differentially Expressed Genes Between ZZZ and MMM. The residuals from the normalization model were subjected to gene-specific models of the form $r_{ijk} = g + A_i + D_j + G_k + \varepsilon_{ijk}$, where g is the average intensity associated with a particular gene; G_k is kth genotype (k = 1–8). Array (A_i) here and thereafter represents as spot effects on the arrays and is also treated as random effect. A *t* test was performed to identify differentially expressed genes between MMM and ZZZ. Four kinds of *t* tests namely, standard *t* test; global *t* test, using an estimate of error variance that is pooled across all genes; regularized *t* test, combing information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator for a gene-specific *t* test; and *F*s test proposed by Cui *et al.* (35) were conducted to access the number of differentially expressed genes between MMM and ZZZ. We also carried out permutation analysis by randomizing the residuals (within each gene) from the fitted null model 1,000 times and recomputing the *t* statistics. The result did not differ from those obtained by using standard statistical tables. All above tests were done with the R/maanova package.

The details of model selection, *cis/trans* estimations, hypothesis testing, and variance component estimation are in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank X. Li, E. Touma, and O. Cruz for technical support; J. Moran, T. Greenberg, S. Fang, C.-T. Ting, S.-C. Tsaur, J. Shapiro, and K. Kumar for comments and discussions during the study. H.-Y.W. thanks James C.-K. Shen for his inspiration and encouragement. This work was supported by a National Institutes of Health grant (to C.-I.W.) and National Science Council Taiwan Grants 96-2311-B-002-009-MY3 and 96-2311-B-002-004- (to H.-Y.W.).

1. Wang D, *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.
2. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature* 430:85–88.
3. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. *Nat Genet* 32:432–437.
4. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science* 297:1143.
5. Landry CR, *et al.* (2005). Compensatory *cis–trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* 171:1813–1822.
6. Morley M, *et al.* (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743–747.
7. Hubner N, *et al.* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* 37:243–253.
8. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436:701–703.
9. Yvert G, *et al.* (2003). *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* 35:57–64.
10. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–755.
11. Hollocher H, Ting CT, Wu ML, Wu CI (1997) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Extensive genetic divergence without reinforcement. *Genetics* 147:1191–1201.
12. Hollocher H, Ting CT, Pollack F, Wu CI (1997) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Variation in mating preference and correlation between sexes. *Evolution (Lawrence, Kans)* 51:1175–1181.
13. Osada N, Kohn MH, Wu CI (2006) Genomic Inferences of the *cis*-regulatory nucleotide polymorphisms underlying gene expression differences between *Drosophila melanogaster* mating races. *Mol Biol Evol* 23:1585–1591.
14. Ting CT, Takahashi A, Wu CI (2001) Incipient speciation by sexual isolation in *Drosophila*: Concurrent evolution at multiple loci. *Proc Natl Acad Sci USA* 98:6709–6713.
15. Alipaz JA, Fang S, Osada N, Wu CI (2005) Evolution of sexual isolation during secondary contact: Genotypic versus phenotypic changes in laboratory populations. *Am Nat* 165:420–428.
16. Alipaz JA, Karr TL, Wu CI (2005) Evolution of sexual isolation in laboratory populations: Fitness differences between mating types and the associated hybrid incompatibilities. *Am Nat* 165:429–438.
17. Greenberg AJ, Moran JR, Fang S, Wu CI (2006) Adaptive loss of an old duplicated gene during incipient speciation. *Mol Biol Evol* 23:401–410.
18. Kohn MH, Fang S, Wu CI (2004) Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. *Mol Biol Evol* 21:374–383.
19. Akaike H (1974) New look at statistical-model identification. *IEEE Transact Auto Control* Ac19:716–723.
20. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nat Rev Genet* 9:102–114.
21. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res* 17:1919–1931.
22. Cheung VG, *et al.* (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437:1365–1369.
23. Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat Genet* 40:346–350.
24. Wayne ML, Pan YJ, Nuzhdin SV, McIntyre LM (2004) Additivity and *trans*-acting effects on gene expression in male *Drosophila simulans*. *Genetics* 168:1413–1420.
25. Ronald J, Brem RB, Whittle J, Kruglyak L (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet* 1:e25.
26. Wu CI, Palopoli MF (1994) Genetics of postmating reproductive isolation in animals. *Annu Rev Genet* 28:283–308.
27. Wu CI, Ting CT (2004) Genes and speciation. *Nat Rev Genet* 5:114–122.
28. Wu CI, Xu EY (2003) Sexual antagonism and X inactivation—the SAXI hypothesis. *Trends Genet* 19:243–247.
29. Parisi M, *et al.* (2003). Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* 299:697–700.
30. Gupta V, *et al.* (2006). Global analysis of X-chromosome dosage compensation. *J Biol* 5:3.
31. Montgomery E, Charlesworth B, Langley CH (1987) A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet Res* 49:31–41.
32. Wu CI, *et al.* (1995). Sexual isolation in *Drosophila melanogaster*: A possible case of incipient speciation. *Proc Natl Acad Sci USA* 92:2519–2523.
33. Cherbas L (2006) *Dendrimer Use and Hybridization Protocol* (The Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN), CGB Technical Report 2006-08.
34. Wu H, Churchill GA (2005) *R/MAANOVA: An Extensive R Environment for the Analysis of Microarray Experiments* (The Jackson Laboratory, Bar Harbor, ME).
35. Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 4:210.