J Mol Evol (2007) 64:272–283
DOI: 10.1007/s00239-006-0169-9

JOURNAL OF MOLECULAR EVOLUTION

A New Retroposed Gene in *Drosophila* Heterochromatin Detected by Microarray-Based Comparative Genomic Hybridization

Chuanzhu Fan, Manyuan Long

Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

Received: 20 April 2006 / Accepted: 17 August 2006 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. A genomic pattern of new gene origination is often dependent on a genomic method that can efficiently identify a statistically adequate number of recently originated genes. The heterochromatic regions have often been viewed as genomic deserts with low coding potential and thus a low flux of new genes. However, increasing reports revealed unexpected roles of heterochromatic regions in the evolution of genes and genomes. We identified recently retroposed genes that originated in heterochromatic regions in *Drosophila*, by developing microarray-based comparative genomic hybridization (CGH) with multiple species. This new gene family, named *Ifc-2h*, originated in the common ancestor of the clade of *D. simulans*, *D. mauritiana*, and *D. sechellia*. The sequence features and phylogenetic distribution indicated that *Ifc-2h* resulted from the retroposition from its parental gene, *Infertile crescent* (*Ifc*), and integrated into heterochromatic region of common ancester of the three sibling species 2 million years ago. Expression analysis revealed that *Ifc-2h* had developed a new expression pattern by recruiting a putative regulatory element from its target sequence. The distribution of indel variation in *Ifc-2h* of *D. simulans* and *D. mauritiana* revealed a significant sequence constraint, suggesting that the *Ifc-2h* gene may be functional. These analyses cast fresh insight into the evolution of heterochromatin and the origin of its coding regions.

Correspondence to: Manyuan Long; *email:* mlong@uchicago.edu

Key words: *Drosophila* — Heterochromatin — *Ifc-2h* — Microarray-based comparative genomic hybridization — Retroposition — Selection sweep

Introduction

It has been widely held that gene duplication is a major source for new gene functions (Ohno 1970). The high frequency of gene duplication in eukaryotes has been revealed in many organisms (Long et al. 2003; Eichler and Sankoff 2003; Inoue et al. 2001; Rubin et al. 2000). However, in order to understand the molecular mechanism underlying the evolution of novel genes and their functions that shaped the present patterns of gene duplications and functional components of genomes, we have to depend on direct observation of dynamics of recent duplicate copies. The well-studied cases of young gene evolution from *Drosophila* add significant insight into the early processes of gene duplication. The relatively small genome size of *D. melanogaster* and its well-defined phylogeny with known divergence times provide a convenient context to find young genes in *Drosophila*. Previous studies have shown that duplicates detected in *Drosophila* are more likely to be functional than psuedogenes, in contrast to the situation in vertebrates (Petrov 2001; Petrov et al. 1996).

Since the first young gene was identified in *Drosophila* (Long and Langley 1993), a number of *Drosophila* new genes have been reported. These new genes showed elevated rates of sequence substitutions driven by adaptive evolution (Jones and Begun 2005;

Jones et al. 2005; Long and Langley 1993), new gene structures generated by various molecular mechanisms (Betran and Long 2003; Nurminsky et al. 1998; Wang et al. 2002, 2004b; Yi and Charlseworth 2000), and unexpected cellular and biochemical functions (Loppin et al. 2005; Zhang et al. 2004). The new genes described so far are all located in euchromatic regions of chromosomes. However, little is known about the origin and evolution of genes in heterochromatic regions.

Heterochromatic regions comprise a large portion of many genomes (approximately 30% of the *D. melanogaster* genome) (Hoskins et al. 2002), and heterochromatin is usually considered to be a genomic desert: gene-poor, mostly consisting of repetitive DNA and transposable elements and their relics (Makalowski 2003). However, these regions serve many biological functions, from gene silencing to accurate chromosome segregation (Bernard et al. 2001; Choo 2001; Dillon 2004). Recent studies, moreover, show that heterochromatin contains genes important for fertility and viability (Hoskins et al. 2002; Wakimoto and Hearn 1990). For example, genetic screens in *D. melanogaster* identified 14 and 12 vital loci in the heterochromatin of chromosomes 2 and 3, respectively (Marchant and Holm 1988). The *D. melanogaster* whole-genome shotgun assembly for heterochromatic sequences has defined 30 known protein coding genes with intron-exon structure and 267 predicted protein-coding genes (Hoskins et al. 2002). Therefore, whether and how these genes originated in heterochromatin are an interesting problem (Schulze et al. 2006).

Despite recent advances in molecular technology and the rapidly expanding databases from different organisms, finding new genes is still a challenging and time-consuming endeavor, especially in these organisms whose genomes have not been sequenced. Fluorescence in situ hybridization (FISH) has been applied in an effort to search for new genes in various organisms (Betran et al. 2002; Wang et al. 2004b), but it can only detect new genes in a part of genome; computational search of genomic data depends on the availability of genomic data that are currently limited to only a few model organisms.

Microarray-based comparative genomic hybridization (CGH) has been developed to survey DNA copy-number variation across a whole genome in cancer and other genomic disorder in humans and other model organisms (Barrett et al. 2004; Greshock et al. 2004; Pinkel et al. 1998). A standard protocol is to use the DNA from different cell populations to hybridize to the microarray designed from the genome of the same species (Pinkel and Albertson 2005). However, adopting this technology to detect the variation of gene copy number by gene gain or gene loss in different species is not straightforward, because one has to solve a problem in which the sequences of microarray often exhibit mismatches with the DNA of the tested species. A more diverged species would have more mismatches. The subgroup *Drosophila melanogaster*, containing nine closely related species, is well defined and carries divergence times ranging from 0.2 million to 15 million years ago (mya) (Lachaise et al. 1988, 2000). This phylogeny, thus, provides an ideal system to adopt the microarray-based CGH method to detect new gene candidate. In the present study, we report young gene candidates (derived less than 2 mya) that were initially identified from an Affymatrix oligonucleotide array comparative hybridization experiment. Unexpectedly, we found new retrogenes present in the heterochromatic region. Analysis of these new genes provides a first insight into the origin and evolution of coding regions within this important portion of a genome.

Materials and Methods

Stocks and DNA Extraction

Isofemale stocks of *D. melanogaster* (Oregon-R), *D. mauritiana*, *D. sechellia*, *D. yakuba*, *D. teissieri*, *D. santomea*, and *D. erecta* were raised in our laboratory for over 50 generations. In addition, 38 *D. simulans* strains were kindly provided by Chung-I Wu, Jerry Coyne, Peter Andolfatto, and Eviatar Nevo. These *D. simulans* lines are samples from 13 different localities worldwide: 19 are from France, 9 are from Africa, 4 from North America, 2 from Israel, 2 from South America, and 1 from Cook Island in the Pacific.

Genomic DNA of *D. melanogaster*, *D. simulans*, *D. mauritiana*, *D. sechellia*, *D. yakuba*, *D. tessieri*, *D. santomea*, and *D. erecta* was extracted using the Puregene DNA isolation kits (Gentra Systems) from 25–30 flies (for microarray hybridization, Southern blotting, and genomic DNA PCR amplification) or a single male fly (for the *D. simulans* population survey).

Oligonucleotide Microarray Hybridization and Duplication Identification

Ten micrograms of genomic DNA was digested using DNase I. The digested DNA was then terminally labeled using Enzo BioArray terminal labeling kit with biotin-ddUTP to generate 3′ termini of the fragmentation products. The labeled DNA fragments (~100–150 bp) were hybridized onto The GeneChip *Drosophila melanogaster* Genome Array following the standard Affymetrix protocol (Affymetrix, Santa Clara, CA) of the Functional Genomics Facility, The University of Chicago. The hybridization intensity for each probe was calculated using the RAM algorithm (Wu et al. 2004) following array neutralization, and background correction in *R* using the Bioconductor Affy package (Gautier et al. 2004). The ratio of pairwise comparisons for each probe was calculated using intensity among eight species, and the median value of intensity fold in all probes for each feature was taken as threshold for gene duplication criterion.

Southern Hybridization

Two micrograms of genomic DNA was digested by *Hin*dIII and then transferred to a positively charged nylon membrane (Roche Molecular Biochemials) by Southern blotting. A digoxigenin-labeled

partial *Ifc-2h* sequence was hybridized to the membrane to confirm the copy numbers in different species.

Expression Analysis

Retrotranscription (RT)-PCR was used to analyze the expression profile in different developmental stages and tissues. Total RNA was extracted from *D. simulans* adult virgin females, males, 2-h-old eggs, second- and third-instar larvae, and pupae using a Qiagen total RNA extraction kit. For species of *D. mauritiana* and *D. sechellia*, total RNA was obtained from adult female and male flies. To amplify *Ifc-2h* cDNA in different species, species specific primers were designed based on genomic sequences of *D. simulans*, *D. mauritiana*, and *D. sechellia*.

Characterization of Duplicated Genes Using RACE

To characterize the gene structure, we applied rapid amplification of cDNA ends (RACE) assays and retrotranscription (RT)-PCR following the manufacturer's protocol. 5′ RACE was conducted using a kit and protocol from Ambion. As for the 3′ RACE, the adapter-linked oligo(dT) primer (Life Technologies) was used to synthesize the first strand of cDNA, and two forward specific primers were designed to amplify the 3′ end of the cDNA against the adapter primer (AUAP, Life Technology).

DNA Amplification and Sequencing

Both the parental and the young genes were amplified using gene-specific primers. The double-stranded PCR products were purified using the Qiagen PCR purification kit or Qiagen miniprep Gel purification system. Purified PCR products were directly sequenced using the Applied Biosystems 3730XL 96-capillary automated DNA sequencer. Flanking sequences were obtained using thermal asymmetric interlaced (TAIL) PCR (Liu and Whittier 1995).

DNA Sequence Analyses

Sequences were edited and assembled. Clustal X (Thompson et al. 1997) was used to align sequences for further analyses. Because the homologous sequences are highly similar and the multiple sequence alignment is robust, only limited manual adjustments were necessary.

Monte Carlo simulation was conducted to estimate the probability of obtaining the observed indel distributions in various gene regions of *Ifc-2h*. A random distribution of indels was hypothesized in which the numbers of indels were expected to be proportional to the lengths of the gene regions. An algorithm was written using C language in a UNIX environment. In the simulation, the observed number of indels was scaled to the lengths of the five gene segments (5′ flanking, 5′ untranslated region [UTR], coding, 3′ UTR, and 3′ flanking) and a random indel was generated using a pseudo-random number generator. A simulation was scored if any symmetrical patterns of the 5′ flanking region $>$ 5′ UTR $>$ coding region $<$ 3′ UTR $<$ 3′ flanking region appeared. The observed pattern is one of all these symmetrical patterns. The probability of the symmetrical pattern was estimated from 1 million simulations.

Basic population genetic analyses were implemented in DnaSP (Rozas et al. 2003). The sequence diversity was calculated using nucleotide diversity (π) (Nei 1987) and the population mutation parameter of Watterson's (1975) θ (denoted θ_w). Tests of deviation from neutrality were conducted using Tajima's (1989) and Fu and Li's (1993) *D*. Significance of Tajima's *D* and Fu and Li's *D* was assessed using coalescent simulation of 2000 replicates incorporated with the number of segregation sites. We calculated the above statistics for the coding sequences of *Ifc* and *Ifc-2h* genes and then generated statistical comparisons for values of sequence variations (π_S, sequence diversity for synonymous sites) between *Ifc* and *Ifc-2h* by computing the values of *Pn*(s), the probability that a sample of *n* alleles is associated with s polymorphic sites at a level of nucleotide variation expected under the hypothesis of neutrality (Hudson 1990; Wang et al. 2004a). The value of π used in calculations was 0.0292 (see Results and Discussion), which is close to the *D. simulans* genome average $\pi_S = 0.0306$ (Powell 1997). The mutation rates were inferred from *Ks* values of *Ifc* and *Ifc-2h*, which were calculated using sequence comparison between *D. simulans* and *D. mauritiana*. The minimum number of recombination events R_m (Hudson and Kaplan 1985) and the linkage equilibrium among all pairwise comparisons of polymorphic sites (Rozas et al. 2001) were estimated using DnaSP.

The phylogenetic analysis was performed using the neighbor-joining method with Jukes-Cantor distance implemented in PAUP*4.0b10 (Swofford 2002). Since the start and stop codons in *Ifc-2h* have disappeared from its parental gene *Ifc*, only the 369-bp *Ifc-2h* and *Ifc* fragment shared in all eight species was used in tests of molecular evolutionary hypotheses using the codon model (codeml) (Goldman and Yang 1994; Yang 1998) implemented in PAML (Yang 1997). For this analysis, a tree depicting *Ifc-2h* and *Ifc* phylogeny was used with the *D. melanogaster Ifc* gene sequence as outgroup (see Results). Since *Ifc-2h* appeared after the *D. simulans* and *D. melanogaster* lineages diverged (Fig. 5B), for the initial Codeml analyses, two models, (i) a single ratio (ω) for all branches and (ii) a free ratio (ω) for each branch, were employed first to determine whether the Ka/Ks ratios were indeed different among lineages. If so, subsequent tests with multiple-ratio models were conducted. Ka/Ks ratio differences among branches were evaluated by the maximum likelihood ratio test (LRT). Log likelihoods of the defined models were compared with a chi-square distribution with degrees of freedom equal to the difference in the number of variable parameters between the nested models. The numbers of synonymous and nonsynonymous substitution along each branch were calculated based on branch length (*t*) and Ka/Ks ratio (ω) together with the estimated transition/transversion ratio (κ) under the free-ratio model. Because the *D. sechellia Ifc-2h* gene shares a short coding region with the *D. simulans Ifc-2h* gene but a reasonably long region (297 bp) with the *D. mauritiana Ifc-2h* gene, we partitioned Ka/Ks analyses into two steps: (1) use the dataset for the 369-bp fragment shared by *ifc* and *ifc-2h* in all sequences except the *D. sechellia Ifc-2h* gene, and then (2) investigate constraint in *D. sechellia Ifc-2h*. We did a Ka/Ks analysis on the coding region of 297 bp fragment shared between *D. sechellia* and *D. mauritiana Ifc-2h* genes with one *Ifc* outgroup from *D. simulans*. We thus excluded the noncoding region following the premature stop codon in the *D. sechellia* genome from analysis.

Promoter Prediction

We manually searched for putative promoter motifs in the 5′ flanking sequences of *Ifc-2h* (e.g., TATA-boxes) and further applied two programs (NNPP and McPromoter) to predict regulatory sequences. NNPP (Reese 2001) can predict the regulatory region by application of a time-delay neural network with a threshold of 0.8 (which is predicted to give a false-positive rate of 0.4% for NNPP). McPromoter (Ohler et al. 2002) is a statistical tool aiming at identifying the *exact* localization of eukaryotic RNA polymerase II transcription start sites based on specific models designed in *D. melanogaster* DNA (sequence sensitivity set for highest 65% and threshold of 0.8).

Results

Duplication Identification Using Microarray-Based Comparative Genomic Hybridization

Our CGH pilot study using a *D. melanogaster* line with known segmental duplication (z[1]w[118]; Dp(1;2)w[+]70h; Bloomington Drosophila Stock Center) suggested that a ratio >1.5 is likely indicative of a duplication in the *Drosophila* genome (Noe, Emerson, and Long, unpublished data). By examining the microarray intensity ratios in pairwise comparisons, we identified duplication candidates with ratios ≥ 1.5 in these species. We further applied genomic Southern hybridization and/or blast search to genomic sequence data of *D. simulans* and *D. yakuba* to confirm the duplicate copies and survey their phylogenetic distribution. By these means, we found a number of new duplicates in species of *D. melanogaster*, *D. yakuba*, and a sibling species clade consisting of *D. simulans*, *D. mauritiana*, and *D. sechellia* (Fan, Emerson, and Long, unpublished). For example, we identified 12 new candidate duplicates in the branch leading to *D. melanogaster* since its divergence from the *D. simulans* clades and 5 putative candidate new genes in the *D. simulans* branch since the common ancestor of that clade. We will focus on one peculiar gene in this paper, to further understand its origination process in a genomic location not previously expected and interpret the value of the genomic technology we developed.

Two Homologous Copies of Ifc *Gene Sequences Exist in* D. simulans, D. mauritiana, *and* D. sechellia

Our CGH comparisons found candidates yielding elevated fluorescence intensity ratios (threshold >1.5) in the clade containing the three sibling species *D. simulans*, *D. mauritiana*, and *D. sechellia*. Expression and evolutionary analyses of these candidates led to the discovery of a new gene copy of *infertile crescent* (defined as *Ifc-2h*) originated in the heterochromatic region of the second chromosome and its parental gene *infertile crescent* (*Ifc*). *Ifc*, containing a single intron, located in the large arm of the second chromosome (2L, 26B2).

The fluorescence intensity ratios of *D. simulans*, *D. mauritiana*, and *D. sechellia* to *D. melanogaster* are higher than those of the other four species (sim/mel = 1.74, mau/mel = 1.73, sec/mel = 1.56, yak/mel = 1.15, tei/mel = 1.42, san/mel = 1.49, ere/mel = 1.23), suggesting that multiple copies of *Ifc* gene sequence are likely to be present in *D. simulans*, *D. mauritiana*, and *D. sechellia*, and a single copy in the rest of species in which the ratio is slightly higher than unity. The control experiment that tested the *D. melanogaster* line with the known segmental duplicate genomic region ((z[1]w[118];Dp(1;2)w[+]70h) revealed that the intensity for the single copy fluctuated from 0.8 to 1.5 (Noe, Emerson, and Long, unpublished results). By blasting the *Ifc* sequence of *D. melanogaster* to genomic data of *D. simulans*, we found two homologous sequences to *Ifc*. One, with a single intron, apparently is the parental copy; the other, without an intron, is likely to have been derived via retrotransposition. A blast search against the *D. simulans* (strain white 501) sequences positioned the duplicate on the second chromosome within heterochromatin (chr2h_random_005 genomic scaffold). A Southern hybridization using *Hin*dIII digestion further confirmed the blast result. Only a single band was found in *D. melanogaster*, *D. teissieri*, *D. santomea*, and *D. erecta*. Two bands were found in *D. simulans*, *D. mauritiana*, and *D. sechellia*. The *D. yakuba Ifc* has a *Hin*dIII site in its intron, as shown by our PCR sequencing (unpublished data not shown), so *D. yakuba* actually has one copy though two bands (Fig. 1).

Both *Ifc* and *Ifc-2h* genomic sequences were amplified using locus-specific primers. Coding sequences of *Ifc* in the three species are conserved except for a few point mutations. In contrast, the sequences of *Ifc-2h* have a different structural pattern in the three species. In *D. simulans*, there are three deletions in *Ifc-2h* (Fig. 2 and Supplemental Data). The first deletion of 13 base pairs (bp), located 150 bp downstream of the start codon of *Ifc*, results in a premature stop codon by a change in reading frame. A short peptide sequence of only 56 amino acid residues could be redefined by this new stop codon. The second is the loss of the only intron of the parental gene, suggesting retroposition association. The third is a deletion of 375 bp located between the coding region and the 3′ UTR of *Ifc*, which removed the original stop codon. Therefore, a new start codon (92 nucleotides downstream of the deletion) and stop codon have to be defined if the new gene is functionally transcribed and translated (see Supplemental Data). This defines a coding region of 462 nucleotides encoding a peptide sequence of 154 amino acid residues, a more reasonable size of protein sequence than the first short peptide sequence. Most of this sequence is similar to *Ifc* in the same reading frame. In *D. mauritiana*, in addition to intron loss, we found three deletions (24, 30, and 42 bp, respectively) in the coding region, which do not disrupt the original reading frame, and one deletion in the 3′ UTR. In *D. sechellia*, *Ifc-2h* has a 122-bp deletion of a sequence between the two parental exons, so *Ifc-2h* in *D. sechellia* has the original (*Ifc*) start codon but a new stop codon (Fig. 2 and Supplemental Data). Given the premature stop codon that shortened the coding region to 297 bp in the *D. sechellia Ifc-2h*, there is a possibility that this *D. sechellia* gene may be degen-

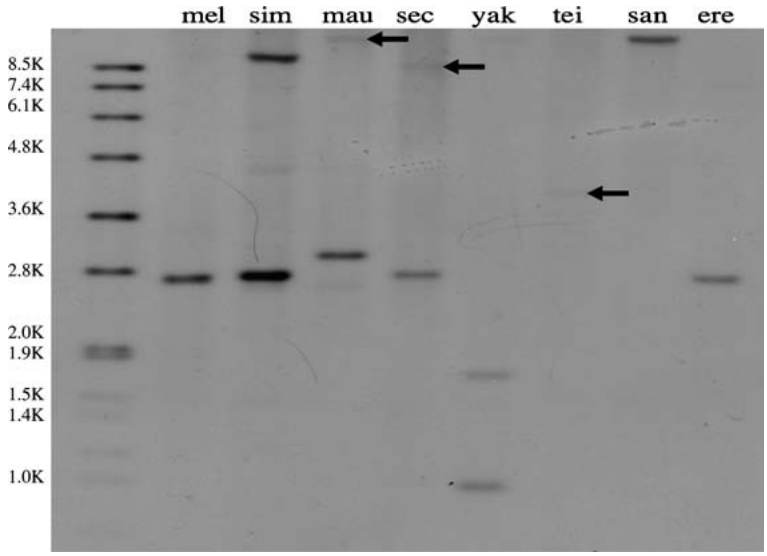

Fig. 1. Genomic Southern blot using a probe of *Ifc-2h* with *Hin*dIII digestion. Species names are shown above each lane. mel, *D. melanogaster*; sim, *D. simulans*; mau, *D. mauritiana*; sech, *D. sechellia*; yak, *D. yakuba*; tei, *D. teissieri*; san, *D. santomea*; ere, *D. erecta*.

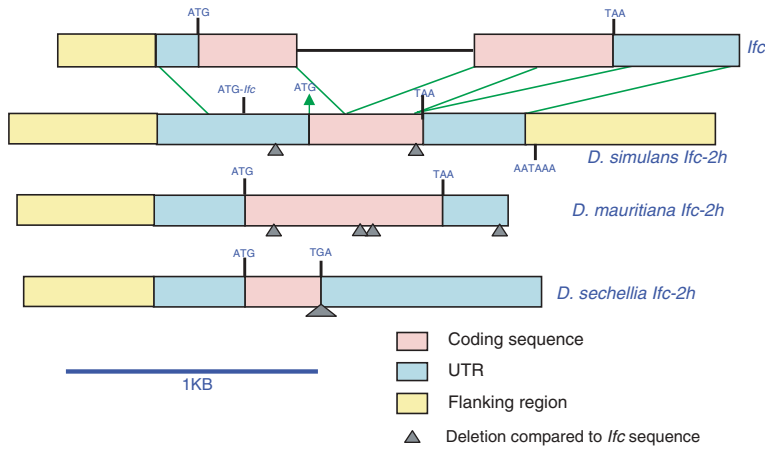

Fig. 2. Schematic sketch of gene structure of *Ifc* and *Ifc-2h*. Start codon, stop codons, and adenylation signal are shown.

erating while the other two new gene lineages are evolving new functions. We thus focused on the analyses of *Ifc-2h* in the other two species, *D. simulans* and *mauritiana*.

Ifc-2h *in* D. simulans, D. sechellia, *and* D. mauritiana *Is Transcribed*

Both *Ifc* and *Ifc-2h* are transcribed but show different expression patterns (Fig. 3). The parental gene, *Ifc*, is ubiquitously transcribed in all developmental stages of *D. simulans* (Fig. 3A). *Ifc-2h* is expressed in all three species (Figs. 3B, D, and E) and equally expressed in both head and body in *D. simulans* (Fig. 3C). The expression profile for developmental stages in *D. simulans* displays high transcription in egg, second-instar larva, and adult, but relatively low transcription in third-instar larva and pupae (Fig. 3B). These expression patterns are consistent with a possible functional divergence between the new gene and its parents. However, its expression profile alone is not evidence that the new gene is functional since some pseudogenes are expressed. Further evidence of functionality can be sought by examining the evolutionary constraints on the gene sequences.

Sequence Evolutionary Constraints in the Coding Region of Ifc-2h

Significantly functional constraint of *Ifc-2h* was observed in patterns of indel from polymorphism and divergence in coding and noncoding regions.

First, we compared fixed indel changes in *Ifc-2h* between *D. simulans*, *D. mauritiana*, and *D. sechellia*. These indels were determined by aligning all three *Ifc-2h* sequences (see Fig. 2). We identified 12 fixed indels from coding and noncoding regions, including 4 in the coding region (3 in *D. mauritiana*) and 8 in noncoding regions (Table 1). There are two distinct patterns associated with the two regions: the lengths of all four indels in coding regions are multiples of three. This pattern in *D. mauritiana* is significantly

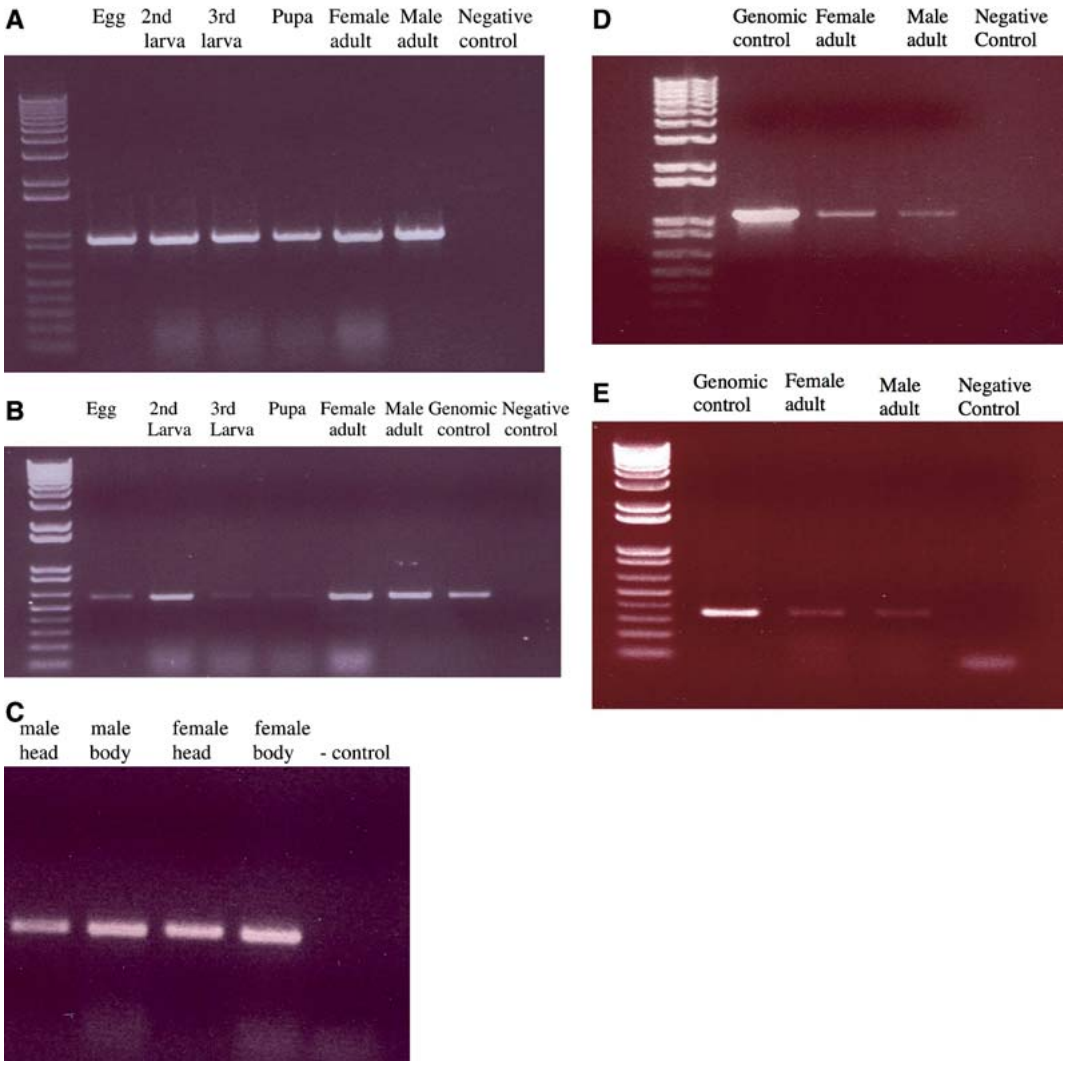

Fig. 3. RT-PCR for (**A**) *Ifc* in *D. simulans*, (**B**) *Ifc-2h* in *D. simulans*, (**C**) *Ifc-2h* in head and body for *D. simulans*, (**D**) *Ifc-2h* in *D. sechellia*, and (**E**) *Ifc-2h* in *D. mauritiana*.

Table 1. Fixed indels identified in aligned *D. simulans*, *D. mauritiana*, and *D. sechellia Ifc-2h* sequences

Location	No. of indels found	Length of indels (bp)
Coding region	4	24, 30, 42, 375
Noncoding region (including flanking and UTR regions)	8	1 (2), 2, 6, 13, 33, 122, 182

different ($p = (1/3)^3 = 0.0369$) from random distribution found in the noncoding regions. This revealed significant evolutionary constraint to maintain a nondisrupted long reading frame in the coding region of *D. mauritiana* (Table 1).

Second, we examined polymorphic indels in the sequence data of 38 *D. simulans* alleles. We identified eight indel polymorphisms in the *Ifc-2h* gene region (Fig. 4). None of these polymorphisms are present in the coding region; all are located in the flanking region and UTR. Furthermore, a gradient appears in the distribution of indel numbers: flanking regions > UTR > coding region. This gradient is symmetrical on the 5′ and 3′ sides of the coding region, namely, the 5′ flanking region > 5′ UTR > coding region < 3′ UTR < 3′ flanking region. After scaling the distribution of these indels to the length of each segment, Monte Carlo simulation showed that the probability for random occurrence of all possible symmetrical gradient patterns including this observed one is significant ($p = 0.0263$). These analyses reveal that the *Ifc-2h* genes in *D. simulans* and *D. mauritiana* are evolutionarily constrained.

Single-Nucleotide Polymorphism and Divergence in Ifc-2h

We analyzed the polymorphism spectrum of the *D. simulans Ifc-2h* gene sequences, transcribed se-

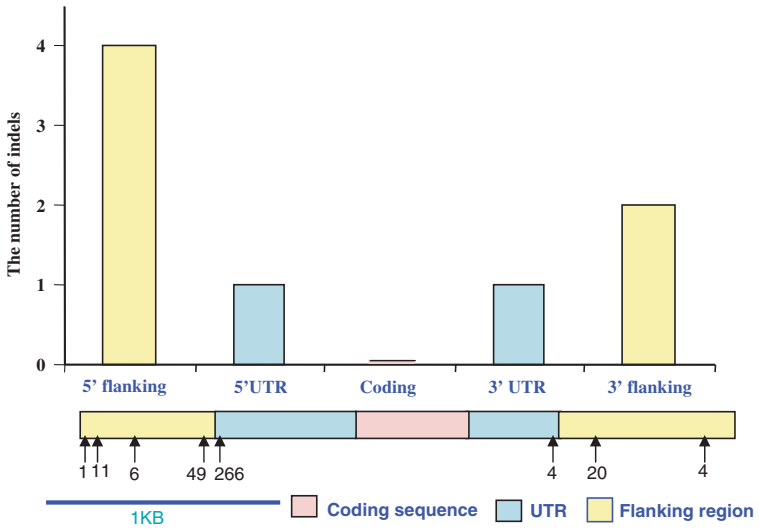

Fig. 4. *Ifc-2h* polymorphic insertions and deletions (indels) found in *D. simulans* population sequences. Arrows indicate the relative positions of indels. The number below the arrow indicates the indel size (bp).

Table 2. Population genetic analysis of *Ifc* and *Ifc-2h* in *D. simulans*

Region	L (bp)	*n*	N_{hap}	S	π	θ_w	Tajima's *D*	Fu & Li's *D*
Ifc-2h transcript sequence (coding region & UTRs)	1500	38	14	23	0.00249	0.00445	−1.48864* (***p* = 0.043**)	−2.8371* (***p* = 0.020**)
Ifc-2h coding sequence	375	38	8	9	0.00272	0.00571	−1.54503* (***p* = 0.043**)	−2.35108* (***p* = 0.013**)
Ifc-2h 5′ flanking	602	38	15	18	0.00869	0.00801	0.28262 ($p > 0.1$)	−0.95822 ($p > 0.1$)
Ifc-2h 3′ flanking	649	38	7	6	0.00180	0.00228	−0.57510 ($p > 0.1$)	0.38048 ($p > 0.1$)
Ifc	1571	20	13	62	0.00901	0.0112	−0.77255 ($p > 0.1$)	−1.38408 ($p > 0.1$)

Note. L, length of sequence; *n*, population size; N_{hap}, number of haplotypes; S, segregating sites. For three synonymous sites of *Ifc-2h*, Tajima's $D = 1.5660$ ($p > 0.10$); for six replacement sites of *Ifc-2h*, Tajima's $D = 1.1612$ ($p > 0.10$).

quences (including coding region and two UTRs), and flanking regions. We compared the pattern of molecular evolution of the transcript and flank regions.

Tajima's and Fu and Li's *D* for transcribed sequences were −1.49 and −2.83, respectively, which is deviated significantly from neutrality ($p = 0.04$ and 0.02) (Table 2). We note here that the partition of coding polymorphisms into replacement sites (six sites) and synonymous sites (three sites) led to nonsignificant Tajima's *D* values, which may be interpreted into small sample sizes. This may be interpreted as a consequence of positive Darwinian selection or association with a particular demographic history of the population (e.g., population expansion). However, the parental copy, *Ifc*, shows no usual sign of demographic history since its polymorphism is less biased (Table 2; Tajima's $D = -0.7726$, $p > 0.1$; Fu-Li's $D = -1.3841$, $p > 0.1$).

After testing the parental copy that is located distal to the heterochromatic regions near the centromere, we tested two regions directly proximal to the coding region of *Ifc-2h*. Tajima's *D* and Fu and Li's *D* show no significant deviation from neutrality in both the 5′ and the 3′ flanking regions ($p = 0.7$ and 0.2 for Tajima's *D* and Fu and Li's *D*, respectively) (Table 2). Thus, these different genomic regions seemingly show different evolutionary histories.

We compared the polymorphism levels of *Ifc* and *Ifc-2h*, revealing that *Ifc-2h* has a much lower level of nucleotide variation than that of *Ifc*. The sequence diversity of *Ifc-2h* is significantly lower than that of *Ifc*: in the whole gene, 3.6 times lower ($\pi_T = 0.00249$ for *Ifc-2h* and $\pi_T = 0.00906$ for *Ifc*); at synonymous sites, an order of magnitude lower ($\pi_S = 0.00238$ for *Ifc-2h* and $\pi_S = 0.02819$ for *Ifc*, which is highly similar to the average value in the *D. simulans* genome; see Materials and Methods) (Table 3). And statistical comparison assuming neutrality (see Materials and Methods) indicated that the synonymous site diversity of the coding region *Ifc-2h* is significantly lower than the π_S of *Ifc* ($p = 0.000128$).

However, the nucleotide diversity in two flanking regions, especially the 3′ flanking region, is also lower than the nucleotide diversity of *Ifc*, suggesting that the gene region may not have a distinct evolutionary history that would have been defined by adaptive evolution. Furthermore, recombination analysis of *Ifc-2h* revealed low recombination in the entire genomic region; 5 was the minimum number of

Table 3. Polymorphism analysis of *Ifc* and *Ifc-2h* in *D. simulans*

Gene	L (bp)	*n*	S	π_T	θ_T	π_R	θ_R	π_S	θ_S
Ifc-2h	1500 (375 coding)	38	23	0.00249	0.00445	0.00283	0.00497	0.00238	0.00817
Ifc	1572 (837 coding)	20	62	0.00901	0.0112	0.00016	0.00044	0.02819	0.03757

Note. L—length of sequence; *n*—sample size; S—segregating sites; subscripts T, R, and S—total sites, replacement sites, and silent sites in the coding region.

Table 4. Likelihood values and parameters estimated under different maximum likelihood models

Branch (see Fig. 5A)	A: one ω	B: free ω	C: two ω	D: two ω	E: three ω	F: two ω
7..8	0.2215	1.8708	1.0166	1	1.8204	1
8..1	0.2215	1.0072	1.0166	1	0.8586	0.8948
8..2	0.2215	0.6978	1.0166	1	0.8586	0.8948
7..9	0.2215	0.9801	0.0526	0.0526	0.0524	0.0524
9..10	0.2215	0.0001	0.0526	0.0526	0.0524	0.0524
10..3	0.2215	0.0537	0.0526	0.0526	0.0524	0.0524
10..4	0.2215	1.0430	0.0526	0.0526	0.0524	0.0524
9..5	0.2215	0.0001	0.0526	0.0526	0.0524	0.0524
7..6	0.2215	0.0683	0.0526	0.0526	0.0524	0.0524
P	6	19	12	11	13	12
l	−712.73	−693.92	−694.54	−694.54	−694.45	−694.53
κ	2.37524	2.34638	2.38842	2.38480	2.34812	2.74958

Note. Model A: one ω for all branches. Model B: free ω. Model C: one ω for young duplicate, one ω for parental copy. Model D: young locus ω fixed at 1, one ω for parental gene. Model E: one ω for fast-evolving branch (7..8) of new gene, one ω for other branches of new gene, and one ratio for parental gene. Model F: as Model E, with fast-evolving branch of young locus ω fixed at 1. *P*, parameter; *l*, likelihood value; κ, transition/transversion ratio.

recombination events, R_m, in 2751 nucleotides in 38 alleles, which is much lower than the level of R_m in the *Adh* gene in a moderate-recombination region (Kreitman and Hudson 1991; Hudson and Kaplan 1985).

A linkage disequilibrium (LD) analysis in the gene region and flanking regions was conducted by chi-square test (Rozas et al. 2001, 2003) for 1081 pairwise comparisons in 47 polymorphic sites. One hundred fifty-six of 1081 pairwise comparisons showed a significant association; 85 of these comparisons remained significant after applying the Bonferroni procedure. Among the 85 significant pairwise comparisons, 28 are located between the 5′ flanking region and the transcription gene region, whereas 17 and 25 are found within the 5′ flanking region and the gene region, respectively. However, we found that the 3′ flanking region has a much lower LD: only 8 of 85 significant pairwise comparisons are between the 3′ flanking region and the gene region, and 2 reside within the 3′ flanking region. These data clearly indicate strong linkage disequilibrium between the gene region and the flanking regions, compared to the LD distribution within all three regions.

Finally, the McDonald-Kreitman test of evolution of *Ifc-2h* was not significant ($G = 0.024$, $p = 0.73476$) for the comparison between *D. simulans* and *D. mauritiana*, consistent with Ka/Ks ratio, did not reveal excess amino acid changes. Therefore, the protein sequence evolution may have been neutral.

Sequence Evolution of Ifc-2h

Divergence analyses were carried out using sequences from *Ifc* of *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* and *Ifc-2h* of *D. simulans*, *D. mauritiana*, and *D. sechellia*. The LRT test showed that the free-ratio model is significantly better than the one-ratio model (Table 4, Fig. 5), indicating the different ω values (Ka/Ks) along different branches.

Comparisons of Model A vs. C and Model A vs. E showed that *Ifc* evolved much slower than *Ifc-2h*. As shown in Table 4, the Ka/Ks ratio for *Ifc* is on average very low: 0.0526. Thus, *Ifc* evolved under strong purifying selection, e.g., high functional constraint. In contrast, *Ifc-2h* had accelerated evolution, resulting from either relaxed selection or Darwinian positive selection, which may be discriminated by the Ka/Ks ratio. Under neutrality, the nonsynonymous substitution rate should be equal to the synonymous substitution rate. If positive selection favors adaptive changes in a particular lineage, the Ka/Ks ratio may exceed 1. When we look at all branches of the *Ifc* and *Ifc-2h* gene tree, we find that the two segments (7..,8 from the origin of *Ifc-2h* till the separation of *Ifc-2h*;

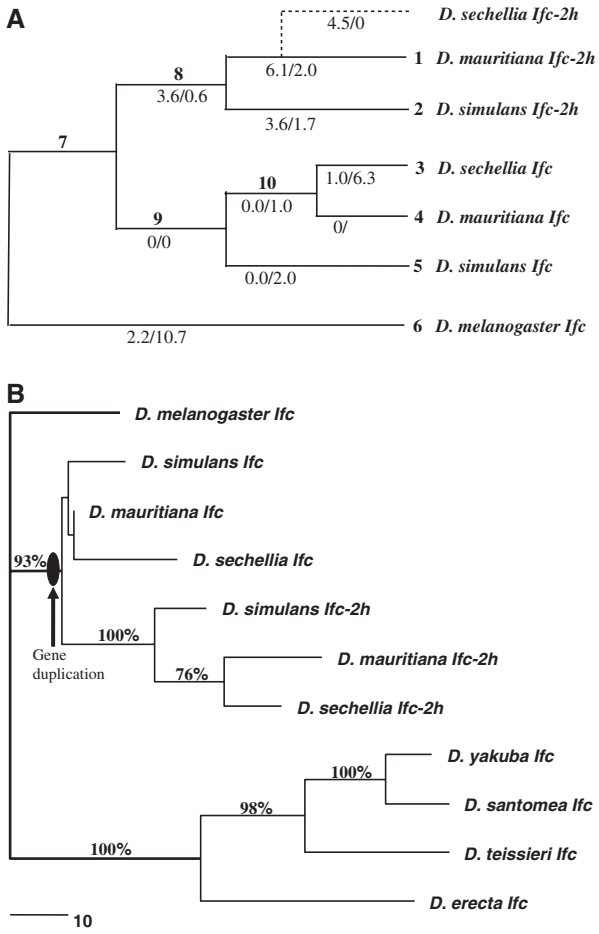

Fig. 5. **A** Gene and species genealogy used in the Codeml analysis. Estimated numbers of nonsynonymous (Ka)/synonymous substitutions (Ks) are shown under every branch. The values for the *D. sechellia Ifc-2h* gene (dashed line) were calculated using its 297-bp coding region aligned with the sequences of the *D. mauritiana Ifc-2h* gene and the *D. simulans Ifc* gene. **B** The neighbor-joining tree inferred from transcript sequences of *Ifc* and *Ifc-2h*. Bootstrap values are given above branches.

8...1 in the subsequent lineage) have Ka/Ks ratios >1 (Table 6), indicating potential positive selection acting *Ifc-2h*. However, the model comparison tests (Model C vs. D and Model E vs. F) did not suggest a significant difference in Ka/Ks for these two branches (Table 5).

5′ Regulatory Region of Ifc-2h

Searching the flanking sequences of the *Ifc-2h* gene, we first found a putatively conserved TATA-box motif (TATAGAAGAAAA) in *D. simulans*, *D. mauritiana*, and *D. sechellia* located 26 bp upstream of the transcription start sites (see Supplemental Data). It has 67% sequence identity with the TATA-box motif of *Ifc* (TATAGTTTTAAAA). A similar TATA motif has also been found in heterochromatic genes (Yasuhara et al. 2005). Furthermore, one putative promoter sequence region (ATTTCCCCCTTAAAACCCCC TCGCGTCTTCATACGGACCTAGCGCAGCGA) located +98 from the transcription start site was identified by both NNPP (with a high score of 0.97) and McPromoter prediction (see Supplemental Data). As discussed above, *Ifc-2h* is located in a gene-poor region containing primarily transposable element sequences. We searched several thousand base pairs of the 5′ flanking region up, and were unable to identify any genes. Therefore, *Ifc-2h* must have co-opted an intergenic sequence that is similar to a promoter sequence.

Table 5. Model comparison using likelihood ratio test (LRT)

Models compared	Degrees of freedom	LRT statistic	χ^2 value ($0.05 < p < 0.01$)
A vs. B	13	37.62**	22.36–34.53
A vs. C	6	36.38**	12.59–22.46
A vs. E	7	36.56**	14.07–24.32
C vs. D	1	0.000	3.84–10.83
E vs. F	1	0.16	3.84–10.83

Discussion

Based on the structure and sequences (Fig. 2) of *Ifc* and *Ifc-2h*, we can infer the series of events that led to the origination of the new gene *Ifc-2h*. First, in the ancestral species of *D. simulans, D. sechellia*, and *D. mauritiana*, an mRNA transcribed from *Ifc* was reverse-transcribed and inserted into a heterochromatic region. Second, the newly generated *Ifc-2h* recruited nearby regulatory regions and other regions to become an actively transcribed gene. Third, the structure and sequence of *Ifc-2h* further evolved by accumulating new mutations including major changes such as indels and coding substitutions. Fourth, the major evolutionary including background selection and genetic drift continue to act on mutations to characterize the gene structures in the three *Drosophila* species.

Our analyses of the between-species fixed indels and within-species indel polymorphisms in *D. simulans* reveal that the coding region sequence is under strong constraint to maintain reading frame. Furthermore, we have identified the new specific expression profiles that the *Ifc-2h* gene evolved. These data suggest that purifying selection is acting on the standing variation of *D. simulans*, suggesting that the *D. simulans Ifc-2h* is likely a functional gene. However, the *D. sechellia* copy is likely degenerating to a pseudogene, because a premature nonsense mutation in its coding region shortened the reading frame extensively. The intact long reading frame of *Ifc-2h* in *D. mauritiana* with all three in-frame deletions suggests that this copy is more likely a functional gene.

Table 6. Detailed output identifying parameters by free-ratio model in Codeml

Branch	t	S	*N*	Ka	Ks	Ka/Ks (ω)
7..8	0.034	91.5	277.5	0.0130	0.0069	1.8708
8..1	0.065	91.5	277.5	0.0218	0.0217	1.0072
8..2	0.043	91.5	277.5	0.0129	0.0185	0.6978
7..9	0.000	91.5	277.5	0.0000	0.0000	0.0000
9..10	0.008	91.5	277.5	0.0000	0.0111	0.0001
10..3	0.059	91.5	277.5	0.0037	0.0684	0.0537
10..4	0.000	91.5	277.5	0.0000	0.0000	0.0000
9..5	0.017	91.5	277.5	0.0000	0.0223	0.0001
7..6	0.105	91.5	277.5	0.0080	0.1168	0.0683

Note. κ (ts/tv) = 2.34638. t, branch length; S, total number of synonymous substitutions; *N*, total number of nonsynonymous substitutions; Ka, number of nonsynonymous substitutions per site; Ks, number of synonymous substitutions per site.

The analysis of variation, recombination, and linkage disequilibrium in the gene regions and flanking regions does not support the conclusion from the polymorphism spectrum distribution that the gene region and flanking regions are associated with different evolutionary histories. The negative Tajima's *D* may be a consequence of stochastic fluctuation of the spectrum of polymorphisms, rather than a signal from selective sweep (Braverman et al. 2005). The reduced level of variation detected in *Ifc-2h* and its flanking regions more likely results from background selection (Charlesworth et al. 1993; Nordborg et al. 1996). Furthermore, both McDonald-Kreitman test and likelihood analysis of the Ka/Ks ratio in evolution of *Ifc-2h* sequences reveal a neutral evolution, consistent with the conclusions of nucleotide polymorphism analyses that estimated a high proportion of nonsynonymous polymorphisms (Table 3).

These results are reminiscent of an important functional gene, fibronectin, which is involved in many important cellular processes but is associated with low sequence constraint, with a Ka/Ks ratio close to unity (Solidar et al. 2004). Thus, in general, a Ka/Ks ratio of unity may indicate neutral evolution but does not necessarily implicate pseudogene. Meanwhile, the nucleotide diversity of *Ifc-2h* at replacement sites was not lower than that at synonymous sites ($\pi_R = 0.00283$ vs. $\pi_S = 0.00238$; Table 3), suggesting that the replacement and synonymous mutations are similarly constrained. The peculiarity of *Ifc-2h* and fibronectin in the distribution of selective constraint is that the constraint is on the maintenance of the reading frame rather than the particular amino acid sequence.

Given the fact that heterochromatic regions in *Drosophila* generally have low rates of recombination, and commensurately low levels of polymorphism due presumably to Hill-Robertson effects (McVean and Charlesworth 2000), the efficacy of selection is also low. Therefore, after a new gene landed in the heterochromatic regions, the new mutations in the gene that might push the new gene toward a novel function have a higher selection threshold to overcome in order to be fixed by selection and not be eliminated by drift. This makes the origination of new gene functions in heterochromatin more difficult than for genes in euchromatin. Here we provide an example of a gene that may be overcoming this increased challenge and demonstrate that retroposition is a viable mechanism in heterochromatin.

Our array-based CGH analysis provided an efficient experimental genomic method to detect young genes in closely related species of model organisms, though we have to consider two challenges for adopting this method to detect new genes in different species. First, the GeneChip oligo array was designed from genomic sequences of *D. melanogaster*, so the divergence between *D. melanogaster* and other *Drosophila* species may compromise hybridization signals. We found that 1% sequence divergence can lead to a 5%–8% decline in signal in genomic hybridization. Hybridization performs better with species evolutionarily closer to the source species from which the microarray oligonucleotide chip was developed. Second, we also noted here that the signals from duplicate copies vary over a considerable range. Paralogous duplicates in the related species anticipate the divergence from the sequence of probes. The degree of divergence will be correlated with the time of the duplication event and the degree of functional constraint on the genes. We have observed that new genes tend to evolve at an accelerated rate early in their evolution, so the signal intensity may drop considerably, hampering their detection.

Conclusions

Our results show that *Ifc-2h* originated recently through the retroposition and might be a functional protein coding gene. Subsequently, the new gene evolved new expression patterns by recruiting new regulatory elements from the target sequences, revealing the rise of new functions in the sibling

species of *Drosophila* within 1 million–2 million years. Further, our study demonstrates that in *Drosophila* not only does the "genome desert" (heterochromatin) have fertile "gene oases," but also new genes may originate and add to existing gene populations in the heterochromatic portion of the genome. Although it is unknown whether or not the *Ifc-2h*/*Ifc* system represents a general scenario for the origination processes of many other heterochromatic genes, its origination and evolution provide new insight into this interesting but less known portion of the genome. Finally, our array-based CGH analysis provided an efficient experimental genomic method for detecting young genes in closely related species of model organisms.

Acknowledgments. The authors thank a number of people (Chung-I Wu, Jerry Coyne, Peter Andolfatto, and Eviatar Nevo) for provision of fly strains; Xinmin Li for performance of the microarray hybridization; members of the Long laboratory for valuable discussions and inputs, particularly, J. J. Emerson and Ying Chen for data analyses and Janice Spofford and Roman Arguello for critical reading of the manuscript; Gerald Wyckoff for discussion about the evolution of the fibronectin gene; and two anonymous reviewers for their suggestions, especially for the interpretation of the detected biased spectrum of polymorphisms. This work was supported by NIH and NSF grants to M.L.

References

Barrett M, Scheffer A, Ben-Dor A, et al. (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci USA 101:17765–17770

Begun D (1997) Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*. Genetics 145:375–382

Bernard P, Maure JF, Partridge JF, Genier S, Javerzat JP, Allshire RC (2001) Requirement of heterochromatin for cohesion at centromeres. Science 294:2539–2542

Betran E, Long M (2003) *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics 164:977–988

Betran E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. Genome Res 12:1854–1859

Braverman J, Lazzaro BP, Aguade M, Langley CH (2005) DNA sequence polymorphism and divergence at the erect wing and suppressor of sable loci of *Drosophila melanogaster* and *D. simulans*. Genetics 170:1153–1165

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303

Choo K (2001) Domain organization at the centromere and neocentromere. Dev Cell 1:165–177

Dillon N (2004) Heterochromatin structure and function. Biol Cell 96:631–637

Fu Y, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M, Weber BL (2004) 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. Genome Res 14:179–187

Hoskins R, Smith CD, Carlson JW, et al. (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biol 3:RESEARCH0085

Hudson R (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–42

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

Inoue K, Dewar K, Katsanis N, Reiter LT, Lander ES, Devon KL, Wyman DW, Lupski JR, Birren B (2001) The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. Genome Res 11:1018–1033

Jones C, Begun DJ (2005) Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci USA 102:11373–11378

Jones C, Custer AW, Begun DJ (2005) Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics 170:207–219

Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the *Adh* and *Adh*-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565–582

Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol 22:159–225

Lachaise D, Harry M, Solignac M, Lemeunier F, Benassi V, Cariou ML (2000) Evolutionary novelties in island: *Drosophila santomea*, a new *melanogaster* sister species from Sao Tome. Proc R Soc Lond Ser B 267:1487–1495

Liu Y, Whittier RF (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674–681

Long M, Langley CH (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260:91–95

Loppin B, Lepetit D, Dorus S, Couble P, Karr TL (2005) Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr Biol 15:87–93

Makalowski W (2003) Genomics. Not junk after all. Science 300:1246–1247

Marchant G, Holm DG (1988) Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. II. Vital loci identified through EMS mutagenesis. Genetics 120:519–532

McVean GAT, Charlesworth B (2000) The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155:929–944

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nordborg M, Charlesworth B, Charlesworth D (1996) The effect of recombination on background selection. Genet Res 67:159–174

Nurminsky D, Nurminskaya MV, De Aguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396:572–575

Ohler U, Liao G, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the *Drosophila* genome. Genome Biol 3:research0087.1-0087.12

Ohno S (1970) Evolution by gene duplication. Springer, Berlin

Petrov D (2001) Evolution of genome size: new approaches to an old problem. Trends Genet 17:23–28

Petrov D, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. Nature 384:346–349

Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. Nat Genet 37 (Suppl):S11–S17

Pinkel D, Segraves R, Sudar D, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211

Powell J (1997) Progress and prospects in evolutionary biology: the Drosophila model. Oxford University Press, New York

Reese M (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput Chem 26:51–56

Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158:1147–1155

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

Rubin G, Yandell MD, Wortman JR, et al. (2000) Comparative genomics of the eukaryotes. Science 287:2204–2215

Schulze S, McAllister B, Sinclair D, Fitzpatrick K, Marchetti M, Pimpinelli S, Honda B (2006) Heterochromatic genes in *Drosophila*: A Comparative analysis of two genes. Genetics 173:1433–1445

Solidar A, Paschall JE, Malcom CM, Wyckoff GJ (2004) The SPEED toolkit: a resource for evolutionary analysis in human genetic studies. Am Soc Hum Gen Annu Meet, Toronto, Ontario, Canada, October

Swofford D (2002) PAUP: phylogenetic analysis using parsimony, version 4.0b10. Sinauer Associates, Sunderland, MA

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

Thompson J, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Wakimoto B, Hearn MG (1990) The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*. Genetics 125:141–154

Wang W, Brunet FG, Nero E, Long M (2002) Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA 99:4448–4453

Wang W, Thornton K, Emerson JJ, Long M (2004a) Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. Genetics 166:1783–1794

Wang W, Yu H, Long M (2004b) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet 36:523–527

Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. Johns Hopkins Univ Dept Biostat Working Papers 1:1–26

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yasuhara J, DeCrease CH, Wakimoto BT (2005) Evolution of heterochromatic genes of *Drosophila*. Proc Natl Acad Sci USA 102:10958–10963

Yi S, Charlesworth B (2000) A selective sweep associated with a recent gene transposition in *Drosophila miranda*. Genetics 156:1753–1763

Zhang J, Dean AM, Brunet F, Long M (2004) Evolving protein functional diversity in new genes of Drosophila. Proc Natl Acad Sci USA 101:16246–16250