A portrait of copy-number polymorphism in *Drosophila melanogaster*

Erik B. Dopman* and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138

Contributed by Daniel L. Hartl, October 18, 2007 (sent for review September 18, 2007)

Thomas Hunt Morgan and colleagues identified variation in gene copy number in *Drosophila* in the 1920s and 1930s and linked such variation to phenotypic differences [Bridges CB (1936) *Science* 83:210]. Yet the extent of variation in the number of chromosomes, chromosomal regions, or gene copies, and the importance of this variation within species, remain poorly understood. Here, we focus on copy-number variation in *Drosophila melanogaster*. We characterize copy-number polymorphism (CNP) across genomic regions, and we contrast patterns to infer the evolutionary processes acting on this variation. Copy-number variation in *D. melanogaster* is nonrandomly distributed, presumably because of a mutational bias produced by tandem repeats or other mechanisms. Comparisons of coding and noncoding CNPs, however, reveal a strong effect of purifying selection in the removal of structural variation from functionally constrained regions. Most patterns of CNP in *D. melanogaster* suggest that negative selection and mutational biases are the primary agents responsible for shaping structural variation.

centrality | copy-number variation | deletion | duplication | gene expression

Copy-number polymorphism (CNP) has a dramatic impact on phenotypic variation within species. In humans, copy-variable regions account for >15% of the total detected genetic variation in gene expression (1), and some genes contributing to disease are contained within known duplication and deletion polymorphisms (2). In addition to its role in generating trait variation within species, CNP represents the raw material for gene family expansion and gene duplication between species. This raw material has apparently had a major role in evolution because 30–65% of genes in sequenced eukaryotes have been duplicated (3). On a larger scale, differences in the number, orientation, and distribution of chromosome segments are the most distinguishing features characterizing divergence in genome architecture between species. As in the case of gene duplication, the population genetic processes regulating CNP (and other variation) within species drive these exceptional differences in genome architecture (4).

Although there is ample incentive to uncover the properties and dynamics of CNP, other than in humans little is known about copy-number variation in natural populations. Open questions remain about how much CNP exists in species' genomes. The observation that two unrelated healthy individuals can differ from one another in copy number across their genome raises uncertainty about the existence of an archetypal number of copies for any particular gene. Related to issues of the extent of CNP are differences in the type of CNP that can be found. Namely, the frequency, degree of dominance, and size of CNPs are largely unknown, as are differences between duplication and deletion polymorphisms. Equally important are the locations, chromosomal properties, and DNA sequence composition of CNPs. Finally, of all of the major issues surrounding CNPs, our knowledge of the evolutionary implications and functional consequences is the most limited.

Here, we address these issues by characterizing how the structure of the sequenced *Drosophila melanogaster* genome varies among representative populations from across the species distribution. We focus on differences in copy number between the sequenced *Drosophila* reference strain and five wild-type isofemale fly strains from the United States (New York), West Africa (Cameroon), East Africa (Kenya), French Polynesia, and Europe (The Netherlands). To characterize CNP in *D. melanogaster*, we used microarray comparative genome hybridization (aCGH), a technique that has demonstrated utility for detecting differences in copy number across diverse species and platforms (5, 6). We define a CNP as a genomic segment that, as assayed by aCGH, differs in copy number between a wild-type strain and the sequenced *Drosophila* reference strain.

Results and Discussion

Our microarrays are spotted arrays with 21,413 PCRs from genomic material based primarily on the Heidelberg Assembly (Eurogentec). Most PCRs amplify open reading frames, but annotation with *Drosophila* genomic sequence v5.1 shows that coding, noncoding (intron or UTR), and intergenic regions are each represented. The median interval between the PCR probes is ≈4.1 kb and the mean probe length is 400 bp, which is closer to the optimal length for aCGH (≈140 bp) than other array platforms (e.g., BACs) (7). In total, 11,934 genes are represented on the array and on average there are 1.2 probes per gene.

The performance of aCGH was validated by self–self and male–female hybridizations. Probes were interpreted as revealing a copy-number difference if the standard error of the log-intensity ratio was beyond an intensity-ratio threshold. This threshold ratio was established by constraining the number of false positives to <1% in three replicate self–self hybridizations. Only 14 of 17,728 high-quality probes were beyond a critical threshold of ±0.3 unit of the log-intensity ratio, giving an estimated false-positive rate of 0.08% (Fig. 1). The adequacy of this threshold for detecting copy-number differences was confirmed in three replicate male (XY) versus female (XX) hybridizations by comparing the number of X-linked probes that were beyond the threshold (Fig. 1). Of 2,970 high-quality X-linked probes, 2,620 were greater than the threshold, yielding an estimate of 88% and 12% for the rates of true positives and false negatives, respectively. There is a slight difference in GC content between true positives and false negatives based on the X chromosome, but the magnitude of the difference is small and the effects on the proportion of false negatives is negligible [see supporting information (SI) *Methods*]. The error rate did not differ between probes in coding and noncoding regions, suggesting that any bias due to GC content (or another source) is

Author contributions: E.B.D. designed and performed research and analyzed data; and E.B.D. and D.L.H. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE9639).

*To whom correspondence may be addressed. E-mail: edopman@oeb.harvard.edu or dhartl@oeb.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709888104/DC1.

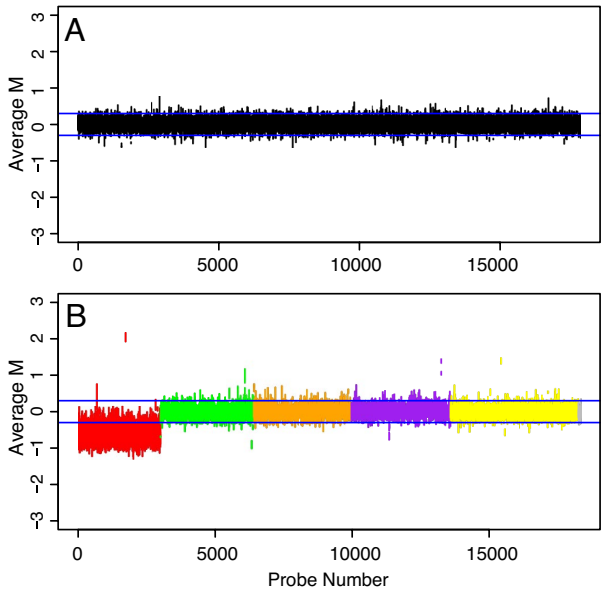

Fig. 1. Average and standard error of log-intensity ratios in self–self hybridization (*A*) and male–female hybridization (*B*) (±0.3 threshold is in blue). Red, X chromosome; green, chromosome 2L; orange, chromosome 2R; purple, chromosome 3L; yellow, chromosome 3R; gray, chromosome 4.

not systematic in its effect on coding and noncoding regions. Of 15,346 high-quality autosomal probes, 36 were beyond the ±0.3 ratio, providing a second estimate of 0.2% for the false-positive rate.

Other than X-linked probes, we assume that probes beyond the critical threshold in our male–female validation arrays represent false positives. In several instances, however, apparent false positives are likely recording real copy-number differences. A contiguous set of seven probes on chromosome arm 3L representing six genes (CG32022, CG6511, *Chorion Protein 18*, *Chorion Protein 15*, *Chorion Protein 16*, and *Paramyosin*) show beyond-threshold negative ratios. Chorion protein and adjacent genes (e.g., CG32022 and *Paramyosin*) are known to be amplified in the follicle cells of *D. melanogaster* females, where amplification of chorion genes is required for normal eggshell development and female fertility (8, 9). From these results we conclude that the false-positive rate predicted from male–female hybridizations is an upper limit. We can also conclude that copy-number differences can easily be detected from DNA extracted from heterogeneous cell populations, such as that from isofemale strains that are segregating for CNPs. Segregating variation is expected within isofemale strains because of heterozygosity contributed by the collected wild-type female and her multiple wild-type mating partners (10).

Unambiguous identification of duplications and deletions is challenging because copy-number changes are relative for aCGH data. By following the convention used for recent CNP assays of the human genome (2), we assign the less frequent or minor allele to the derived state. Minor alleles that are lower in copy number are interpreted as losses (deletions), whereas minor alleles that are higher in copy number are interpreted as gains (duplications). High-frequency CNPs will be misclassified by using this approach, but because >80% of CNPs in our sample are found in a single strain (singletons), most CNPs are likely to be properly classified based on frequency. For those probes in which a minor allele could not be determined (e.g., if the frequency was 0.5 after removal of low-quality probes), the probe was dropped from analyses where gain/loss determination was required.

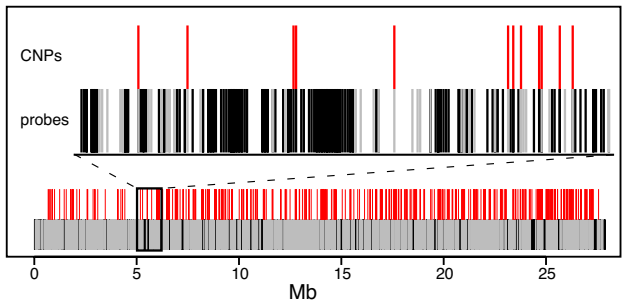

Fig. 2. Distribution of probes (black = coding, gray = noncoding + intergenic) and copy-number polymorphisms (red) on chromosome arm 3R. (*Inset*) Approximately 1 Mb of 3R is shown that illustrates clustering and noncoding bias.

CNP Frequency, Size, and Prevalence in *Drosophila*. In hybridizations using pooled DNA from ≈60 males from each of five wild-type strains and ≈60 males from the sequenced reference strain (four slides per strain), 8.6% of 18,384 high-quality probes were variable in at least one strain, and 99% showed only gain or only loss. CNP in *Drosophila* is apparently quite common with, on average, 436 CNPs per strain. Duplication and deletion CNPs are not equally abundant. When CNPs are polarized into major and minor alleles (1,465 probes), deletions outnumber duplications by ≈2:1 (987:478). Although a polymorphic deletion bias can be found in *Drosophila* (e.g., 11), Redon *et al.* (2) noted that the power to detect duplications could be lower as a result of a smaller ratio of relative change compared with deletions (3:2 versus 1:2). This may in part explain the excess of deletions detected by their platform and by ours. Singleton alleles account for 81% of all variable probes. Although some of these are false positives, this result suggests an appreciable level of between-strain or between-population differentiation for CNP variants.

Regions that are known to vary in copy number between flies in nature are detected as copy-number variable by aCGH. Different *Drosophila* lines possess different transposable element numbers and genomic distributions (12). Of 98 transposable elements represented on the microarray, 58 are variable between the sequenced strain and at least one of the five wild-type strains (SI Table 3).

Although our aCGH scan indicates that many chromosome segments are affected by CNP across the *Drosophila* genome, regions equal to or smaller than single genes are most susceptible to copy-number change. The median length between probes in coding regions is only 4.7 kb, but single-probe copy-number change accounts for 91% (1,440) of the total variation. Of those probes showing evidence for multiprobe change, the median length is ≈3 kb. The largest region showing copy gain is 12 kb on chromosome 2L. It includes two probes: one falls within the coding region of the gene *salm* and the other is located in the 3′ intergenic region. The largest region showing copy loss is ≈33 kb on chromosome arm 2R. The region includes two adjacent probes, both of which fall within an intronic region in the current annotation of the gene *luna*.

Genomewide Consequences of Mutation and Natural Selection for CNPs. Although large genomic regions are not commonly involved in copy-number variation in *D. melanogaster*, clusters of CNPs are found across the *Drosophila* genome ($P = 0.018$) (Fig. 2). This nonrandom distribution suggests the existence of chromosomal segments of structural instability. Structural variation "hot spots" have been found for human and chimpanzee (*Pan troglodytes*), in which ancient segmental duplications (regions of >1 kb with >90% sequence similarity), and included repeat regions, have been implicated in the formation of CNPs (6). Repetitive regions are believed to facilitate structural genomic

EVOLUTION

Table 1. Evolutionary rates for genes with and without copy-number variation

	d_N	Wilcoxon rank-sum test, *P* value	d_S	Wilcoxon rank-sum test, *P* value	d_N/d_S	Wilcoxon rank-sum test, *P* value
Duplication CNP	0.0094	0.71	0.1294	<0.01	0.0737	0.59
Monomorphic	0.0097	—	0.1211	—	0.0819	—
Deletion CNP	0.0139	<0.00000001	0.1281	<0.001	0.111	<0.000001

Nonsynonymous (d_N) and synonymous (d_S) rates. *P* values compare CNP genes with monomorphic genes.

variation, including segmental duplications, through nonallelic homologous recombination (13). In *D. melanogaster*, the tandem-repeats finder algorithm (TRF) identified that tandem repeats are significantly enriched in regions surrounding CNPs (one-sided Wilcoxon rank-sum test, $P < 1e\text{-}04$). This result suggests that repeated sequences may be responsible for generating clustered CNPs in *Drosophila* and that repetitive regions are important catalysts of structural variation among widely diverse species (13).

Recombination has demonstrated mutagenic effects in yeast, humans, and flies, especially in the presence of repeated regions (e.g., refs. 14–16). Recombination rate is significantly greater for genes whose coding region shows deletion polymorphism ($\bar{x} = 2.86$) compared with "monomorphic" genes (those lacking copy-number variation) ($\bar{x} = 2.43$) (Wilcoxon rank-sum test, $P < 1e\text{-}08$), suggesting that the process of homologous recombination is, in part, responsible for producing CNPs (17). Molecular analysis of small *Drosophila* deletions has indicated that approximately half are flanked by direct repeats of 2–7 bp in length (18), supporting a mechanism of slip-strand mispairing during DNA replication. If such events can also accompany repair synthesis during recombination, this could account for the association between recombination and deletion CNPs. However, the rate of recombination does not differ between genes with duplication CNPs ($\bar{x} = 2.52$) and those without copy-number polymorphism ($P = 0.3$).

Genomic intervals in *D. melanogaster* that contain tandem repeats may contribute to a structural dynamism that predisposes some regions to copy-number variation (18). However, in our data, tandem repeats are primarily elevated near CNPs located in noncoding and intergenic regions ($P < 0.001$), rather than in coding regions ($P = 0.47$). This finding suggests that forces beyond the mutational processes by which they originate shape the distribution of CNPs in the *Drosophila* genome. For example, strong purifying selection in protein-coding regions would be expected to erode or constrain any underlying mutational bias that promotes copy-number variation. Our data support this notion in that a 36% reduction in the proportion of deletion CNPs in coding sequence (0.047) is observed compared with those in noncoding sequence (0.073) ($G = 17.85$, $P < 1e\text{-}04$, df = 1) (Fig. 2). Similar results have been found for deletion polymorphisms in humans (e.g., see ref. 19). Duplication polymorphisms involving coding regions (0.024) are reduced by 14% (noncoding regions: 0.028), but the reduction is not significant ($G = 0.89$, $P = 0.34$, df = 1).

If many deletion CNPs are deleterious and recessive, fewer are expected on the X chromosome than on autosomes, because hemizygosity of males uncovers the effects of otherwise recessive mutations, making them susceptible to selection (20). Consistent with the deleterious recessivity of some CNPs, deletion polymorphisms involving coding sequence tend to be preferentially located on autosomes. In particular, a 28% reduction of deletion CNPs in coding regions is observed in the X chromosome (X, 0.036; autosome, 0.05; $G = 6.24$, $P = 0.012$, df = 1). In contrast to the pattern for deletion CNPs in coding regions, the proportion of polymorphic deletions in noncoding regions does not differ between chromosomes, likely because selection is weaker in these regions (X, 0.088; autosome, 0.071; $G = 0.77$, $P = 0.38$, df = 1).

The chromosomal distribution of polymorphic duplications presents a challenge because the genomic position is known only for the copy in the reference sequence. Assuming that most duplication CNPs within species are tandem, or are otherwise located in the same chromosome, we can discern whether polymorphic duplications show autosomal predominance. Unlike deletion CNPs, however, the proportion of duplication CNPs does not differ among chromosome arms in coding or noncoding regions (coding,: $G = 2.01$, $P = 0.16$, df = 1; noncoding, $G = 1.08$, $P = 0.3$, df = 1). Along with the similar proportion of duplication CNPs between coding and noncoding regions, these results suggest that, compared with deletion CNPs, a larger proportion of gains involving functional sequences are selectively more nearly neutral and some possibly beneficial.

Selective Constraint in Copy-Variable Genes. Although many deletions in protein-coding regions may not contribute to polymorphism because of purifying selection, differences in the evolutionary pattern for partially deleted and monomorphic genes should reflect differences in selective constraint. Specifically, genes affected by polymorphic deletions may be more robust to mutations of all types, including those that alter the protein-coding sequence. We tested this idea by comparing the d_N/d_S ratios for orthologous genes between *D. melanogaster* and *D. simulans*, where d_N is the number of amino acid replacement substitutions per nonsynonymous site and d_S is the number of synonymous substitutions per synonymous site. Assuming neutrality for synonymous substitutions, a $d_N/d_S < 1$ indicates that amino acid change is selectively constrained. A d_N/d_S that is elevated, but still less than one, is generally interpreted as a relaxation of selective pressure.

We found that d_N/d_S ratios between genes with polymorphic deletions are significantly shifted toward higher values than those for monomorphic genes (Table 1). Although d_S is also significantly elevated, the magnitude of increase for the median d_S value (1.06) is much smaller than the magnitude of increase for d_N (1.43). Therefore, we interpret a higher d_N/d_S for deletion CNPs as stemming largely from an increased rate of amino acid replacement (comparison with *D. yakuba* yielded similar results). Although this pattern of DNA sequence evolution could be attributed to positive Darwinian selection, it is difficult to imagine why deletions would occur in such genes. A more parsimonious explanation for both observations is reduced selective constraint.

In contrast to deletion CNPs, d_N/d_S ratios (and d_N) between genes with polymorphic duplications in *D. melanogaster* did not significantly differ from d_N/d_S (and d_N) for monomorphic genes (Table 1). As with deletion CNPs, there is evidence for an increase in the rate of synonymous substitution, but the magnitude of increase was also relatively small (1.07). We conclude that parental copies of duplication CNPs within species do not have an obvious tendency for accelerated sequence evolution or for reduced selection pressure.

Essentiality and Centrality. In several species it has been demonstrated that proteins with greater centrality in biological networks evolve slowly and tend to be essential (21). It follows that genes with deletion polymorphisms, which experience weak constraint and evolve rapidly, may have the opposite properties.

We analyzed CNPs with regard to data available for protein–protein interactions in *D. melanogaster*. The interaction data are somewhat noisy owing to methodological artifacts that can bias the assay for any individual interaction (22). Nevertheless, among >10,000 *Drosophila* open reading frames tested for a physical interaction (e.g., ref. 23), we found that the proportion of deletion CNPs with at least one interaction is significantly reduced (240 of 557) compared with genes that lack CNPs (5,787 of 10,727) ($G = 25.04$, $P < 1e\text{-}06$, df = 1).

Unlike deletion CNPs, the proportion of duplication CNPs involved in at least one interaction is not reduced ($P = 0.22$). However, of those genes with ≥1 interaction, polymorphic gains are less likely to be central (in the sense of graphical connectivity and betweenness) in the protein interaction network (one-sided Wilcoxon rank-sum test, $P < 0.04$, SI Table 4). Centrality for polymorphic losses and monomorphic genes does not differ ($P > 0.69$). Reduced network centrality for duplication CNPs in *Drosophila* is consistent with the result in yeast in which gene duplications are negatively correlated with gene-product connectivity (24). Indeed, genes with close paralogs are also less central in the *D. melanogaster* interactome (closeness, $P < 0.01$). From these results it appears that genes with weak network centrality may be more likely to spawn duplicates that are retained across both short and long time scales.

In addition to the degree of network interaction, a gene's propensity to exhibit copy-number variation may be informative about its essentiality. In *Drosophila*, genes with polymorphic deletions are less likely to be lethal when genetically perturbed (48 of 557 vs. 1,412 of 10,727) ($G = 10.77$, $P < 0.01$, df = 1), suggesting their dispensability. The proportion of genes with duplication CNPs that have lethal alleles (32 of 294) is also reduced, but the reduction is not significant ($G = 1.37$, $P = 0.24$, df = 1).

It is perhaps to be expected that a gene's propensity to segregate for deletions and its dispensability are both negatively correlated with the level of network centrality as well as the strength of selective constraint. Essential genes in fly, yeast, and worm tend to be centrally located in interaction networks, where evolutionary constraint is higher (21). Genes with high centrality may be more constrained during evolution because protein-coding changes, including deletions, might impair the ability of a protein to form dependable network interactions (25). Areas of low or no connectivity in protein interaction networks, populated in *Drosophila* by genes with a greater likelihood to exhibit deletion and duplication polymorphisms, may experience reduced pleiotropy (26), and consequently may be more robust to nonsynonymous and structural mutation as well as less constrained during evolution.

Expression Polymorphism and Tissue Specificity. Compared with monomorphic genes, a greater proportion of genes with CNPs are duplicated in *D. melanogaster* (14% gain CNP, 13% loss CNP, 9% monomorphic, $G > 6.88$, $P < 0.01$, df = 1). Copy-variable genes may have a greater propensity to be duplicated because the evolutionary rate of gene duplication is higher or because natural selection acts on CNPs to favor the retention of paralogs, potentially because of positive Darwinian selection. Perhaps the easiest way to envision selection shaping copy-number polymorphism is if differences in gene dosage translate into differences in transcription, and ultimately, into protein concentration. For example, in humans, positive selection appears to have favored an increase in protein level and copy number of a salivary amylase gene in populations with a history of a high-starch diet (27). In *D. melanogaster*, genes with CNPs contribute disproportionately to gene-expression polymorphism (Wilcoxon rank-sum test, $P < 0.01$) (28, 29), suggesting that the phenotypic raw material for selection to act on may exist for some copy-polymorphic genes because of dosage effects on transcription.

Although genes with copy-number variation are more variably expressed among strains, they have a narrower breadth of gene expression among tissues. Genes with CNPs have appreciable transcript levels in fewer tissues (median = 6) than monomorphic genes (median = 9) (Wilcoxon rank-sum test, $P < 1e\text{-}07$). This translates into a significant reduction in the likelihood that genes with CNPs are expressed in more than one tissue (median Simpson's Diversity Index $D_{CNP} = 0.67$, $D_{mono.} = 0.76$, Wilcoxon rank-sum test, $P < 1e\text{-}07$). Both results suggest that CNP occurs in genes that are more tissue-specific in their expression patterns rather than in widely expressed genes that might have housekeeping functions. Of those genes representing the top 25% of the most specific genes (≥79% expressed in a single tissue) the proportion of copy-variable genes significantly differs among *D. melanogaster* tissues ($G = 18.51$, $P = 0.03$, df = 9). Copy-variable genes are most abundant in the midgut (12%) and in male accessory glands (15%) (SI Table 5).

The midgut is the principal site for secretion of digestive enzymes, digestion, and absorption in insects, but it is also the central entry point for viruses, hormones, bacteria, and toxins (30). Indeed, all three protein families largely responsible for detoxification of insecticides (31) are represented by copy-variable genes that have midgut expression and functions that are associated with insecticide metabolism or toxin response [*Cyp6g1* (32), *para* (33), and *GstD2*, *GstD3* (34)]. Of the 28 CNP genes with midgut specificity, defense response and transport are both heavily represented processes. Male accessory glands contain proteins that are transmitted to females during reproduction, the genes of which have been shown to be under intense antagonistic coevolution between males and females (35, 36). Among the 24 CNP genes showing specificity to the accessory glands are genes whose products are involved in sperm competition, female postmating behavior, and defense response.

The observation that copy-variable genes are overrepresented in the midgut and accessory glands may not be coincidental. Both tissues have a high level of environmental interaction that can dramatically impact fitness: the midgut meets challenges associated with ingestion of potentially harmful substances, whereas the accessory gland meets challenges associated with ensuring paternity and fecundity. Similarly, genes in yeast that localize to the cellular periphery, which are likely to have an environmental interaction in this single-celled organism, are more highly duplicated than genes with intracellular function (37). The relationship between the propensity for copy-number variation and environmental interaction has been argued to result from positive selection for a diversity of proteins with extracellular functions to meet the challenges encountered in a spatially and temporally changing environment [e.g., immunity genes, transporters, receptors, enzymes in secondary metabolism, stress response genes (4, 37)].

Among all copy-variable genes in *D. melanogaster*, genes whose products localize to the extracellular region and the plasma membrane are overrepresented; genes that are underrepresented have products whose functions localize intracellularly, to the cytosol and to the nucleus (Table 2). In regard to biological process, genes with CNP are enriched for functions including generation of precursor metabolites and energy, carbohydrate and lipid metabolism, transport, cell signaling, and response to biotic stimulus. Genes whose products are involved in cell proliferation, protein biosynthesis, nucleic acid metabolism, and transcription are underrepresented among those with CNPs. Many of the same functions are enriched or underrepresented in gene duplicates in *D. melanogaster* (Table 2 and SI Table 6). Similar functional patterns characterize copy-number polymorphism and interspecific gene duplicates in diverse species, including single-cell (yeast) and multicellular organisms (humans) (37–39).

A Look Ahead. Some of the patterns identified here for *D. melanogaster* and elsewhere for other species support a partial adaptive explanation for copy-number diversity. Indeed, under certain conditions, some genes have been proven to confer

Table 2. Statistically significant ($P < 0.05$) over- or underrepresentation of GO-Slim categories in *D. melanogaster* CNPs

GO ID	Representation	Description	Classification
GO:0006118	Over	Electron transport*	bp
GO:0006629	Over	Lipid metabolism*	bp
GO:0006091	Over	Generation of precursor metabolites and energy*	bp
GO:0009607	Over	Response to biotic stimulus*	bp
GO:0006811	Over	Ion transport*	bp
GO:0007267	Over	Cell–cell signaling*	bp
GO:0005975	Over	Carbohydrate metabolism*	bp
GO:0006810	Over	Transport*	bp
GO:0005576	Over	Extracellular region*	cc
GO:0005886	Over	Plasma membrane*	cc
GO:0008283	Under	Cell proliferation*	bp
GO:0006996	Under	Organelle organization and biogenesis*	bp
GO:0007275	Under	Development*	bp
GO:0009653	Under	Morphogenesis*	bp
GO:0007154	Under	Cell communication	bp
GO:0019538	Under	Protein metabolism	bp
GO:0008152	Under	Metabolism	bp
GO:0044238	Under	Primary metabolism	bp
GO:0006464	Under	Protein modification	bp
GO:0009790	Under	Embryonic development*	bp
GO:0007165	Under	Signal transduction	bp
GO:0006412	Under	Protein biosynthesis*	bp
GO:0006350	Under	Transcription*	bp
GO:0007049	Under	Cell cycle*	bp
GO:0016043	Under	Cell organization and biogenesis	bp
GO:0009058	Under	Biosynthesis	bp
GO:0015031	Under	Protein transport	bp
GO:0006139	Under	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism*	bp
GO:0050789	Under	Regulation of biological process*	bp
GO:0005829	Under	Cytosol	cc
GO:0043234	Under	Protein complex	cc
GO:0005856	Under	Cytoskeleton	cc
GO:0005634	Under	Nucleus	cc
GO:0005623	Under	Cell	cc
GO:0005737	Under	Cytoplasm	cc
GO:0043226	Under	Organelle*	cc
GO:0005622	Under	Intracellular*	cc

*GO term showing significant over- or underrepresentation for *D. melanogaster* gene duplicates. Biological process (bp) and cellular component (cc) controlled vocabularies.

greater fitness as the number of gene copies changes (4). However, the adaptive potential for CNP is moderated by evidence that reduced functional constraint and mutational bias are likely the dominant evolutionary forces shaping this variation. Confirming copy-number variation and identifying those genes that are targets of positive Darwinian selection represent a major goal for the population genetics of structural variation.

The genomics era has primarily concentrated on the single-nucleotide polymorphism (SNP) as the most biologically relevant feature of the genome. It is becoming increasingly clear, however, that a structurally dynamic genome is common across species and that this structural dynamism has functional and evolutionary consequences. What is still unclear is the extent to which chromosomal changes other than copy-number variation, and in particular inverted regions, contributes to structural variation. Because both copy-number polymorphisms and chromosomal inversions play important roles in heritable disease, adaptation, and speciation (4, 40–42), both should receive thorough attention in the effort to properly characterize genomes and genomic variation.

Methods

Microarray Comparative Genomic Hybridization. To maximize the detection of germ-line copy variation, DNA was extracted from 30 males per line (QIAamp DNA Mini Kit; Qiagen). At least two extractions were combined for shearing to 0.2–1 kb by using a GeneMachines Hydroshear. DNA labeling was performed by using the Invitrogen BioPrime Plus Array CGH Labeling System (see *SI Methods*).

Hybridizations and washes were performed according to Pollack (43) for a minimum of four replicates (two dye-swaps) per isofemale line. The arrays were scanned on an GenePix 4000B Scanner (Axon Instruments) and the images were analyzed by using GenePix Pro 6. Quality-control criteria were applied both manually and by using the LIMMA library (v2.4.13) in R (v2.2.1) (44, 45). Features with at least two high-quality measurements (in $\geq$2 slides) were retained. Sequential spatial and intensity normalization of raw intensity data were performed, followed by estimation of mean ratio of intensity relative to the sequenced strain. The ratio of signal intensities for each strain relative to the sequenced strain were obtained by calculating the mean and standard error ratio of the feature across slides.

Statistical Analyses. Statistical analyses were conducted in R (45). Clustering of CNPs was tested by a 10-probe sliding window that moved in one-probe intervals within chromosome arms (with TE

probes removed). Clustering was defined as windows having two or more CNPs. The number of windows with ≥2 CNPs was tested by randomizing CNP order 1,000 times within chromosome arms. *P* values were the number of randomized sets that had more extreme values compared with the real data. Window size (decreasing to five-probe length) had little effect on results (*P* values were more extreme).

An increase in the number of repetitive regions (anchored by their middle position) in a 50-kb window surrounding CNP probes versus monomorphic probes was used to test for an effect of repeat region (in the sequenced strain) on CNP (TEs were excluded). The results from TandemRepeatFinder and RepeatRunner were used in these tests (46). RepeatRunner tracks were not elevated surrounding CNPs ($P > 0.14$).

Recombination rate was estimated for each gene by using the value *R* (47), and evolutionary rates were obtained from the pipeline result of Gnad and Parsch (48). In all comparisons between copy-variable and monomorphic genes, duplicate measurements were eliminated within groups. In comparisons of the proportion of CNP, we contrast probes in annotated genes because intergenic probes were thought to represent both unannotated genes and nonfunctional segments.

We tested for differences in physical network centrality by using the interaction dataset from BIOGRID (49). Our final list contained 21,665 interactions for 6,852 unique genes (see *SI Methods*). Centrality measures were calculated by using PAJEK (50). Systematic identification of gene essentiality is not available for *D. melanogaster*. As a proxy, we use the number of experimentally induced or naturally occurring lethal alleles annotated in Flybase (e.g., ref. 21). A total of 23,384 lethal alleles are annotated from 8,465 genes.

Gene ontology terms for probes revealing copy change were identified by using GOToolBox (51). Annotations were constructed on a generic slim hierarchy created on August 1, 2007 (ftp://ftp.geneontology.org/pub/go/GO_slims/). A hypergeometric test with Benjamini and Hochberg adjustment was used to assess significance for genes showing copy-number change by comparing the terms for genes represented on the microarray.

The method of Davis and Petrov (52) was used to identify close paralogs in *D. melanogaster*. Reciprocal BLASTp searches identified paralogs (*e*-value $< 10^{-9}$) and were alignable >60% of the protein length. We also required that all paralog pairs had ≥50% identity of the aligned region. This search resulted in 700 paralogous gene pairs.

Estimates of gene expression polymorphism used the variances of gene expression from Meiklejohn *et al.* (28, 29) (4,440 genes). Spatial patterns of gene expression were obtained from FlyAtlas (53). In total, 13,478 different FBgn are represented.

We thank Mohamed Noor and Christian Landry for providing critical reviews of this manuscript and Suzy Renn for allowing us to use some unpublished male-female hybridization data. E.B.D. thanks Pierre Fontanillas, Rob Kulathinal, and Suzy Renn for helpful discussions, encouragement, and assistance. Members of the D.L.H. laboratory were instrumental in the development of the microarray, and Scott Rifkin helped with annotations. The fly lines were provided by Charles Aquadro, Peter Andolfatto, and Frank Jiggins. The research was supported by National Institutes of Health Grant GM068465 (to D.L.H.). E.B.D. is supported by an National Institutes of Health Kirschstein-NRSA Postdoctoral Fellowship 1 F32 GM080090-01.

1. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird, C. P., de Grassi A, Lee C, *et al.* (2007) *Science* 315:848–853.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen WW, *et al.* (2006) *Nature* 444:444–454.
3. Zhang JZ (2003) *Trends Ecol Evol* 18:292–298.
4. Kondrashov FA, Kondrashov AS (2006) *J Theor Biol* 239:141–151.
5. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) *Nat Genet* 23:41–46.
6. Perry GH, Tchinda J, McGrath SD, Zhang JJ, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, *et al.* (2006) *Proc Natl Acad Sci USA* 103:8006–8011.
7. Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA (2006) *Nucleic Acids Res* 34:445–450.
8. Orr W, Komitopoulou K, Kafatos FC (1984) *Proc Natl Acad Sci USA* 81:3773–3777.
9. Claycomb JM, Benasutti M, Bosco G, Fenger DD, Orr-Weaver TL (2004) *Dev Cell* 6:145–165.
10. Imhof M, Harr B, Brem G, Schlotterer C (1998) *Mol Ecol* 7:915–917.
11. Ometto L, Stephan W, De Lorenzo D (2005) *Genetics* 169:1521–1527.
12. Vieira C, Lepetit D, Dumont S, Biemont C (1999) *Mol Biol Evol* 16:1251–1255.
13. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) *Trends Genet* 21:673–682.
14. Montgomery EA, Huang SM, Langley CH, Judd BH (1991) *Genetics* 129:1085–1098.
15. Strathern JN, Shafer BK, McGill CB (1995) *Genetics* 140:965–972.
16. Toffolatti L, Cardazzo B, Nobile C, Danieli GA, Gualandi F, Muntoni F, Abbs S, Zanetti P, Angelini C, Ferlini A, *et al.* (2002) *Genomics* 80:523–528.
17. Dvorak J, Yang ZL, You FM, Luo MC (2004) *Genetics* 168:1665–1676.
18. Petrov DA, Lozovskaya ER, Hartl DL (1996) *Nature* 384:346–349.
19. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) *Nat Genet* 38:75–81.
20. Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory* (Harper and Row, New York).
21. Hahn MW, Kern AD (2005) *Mol Biol Evol* 22:803–806.
22. Chiang T, Scholtens D, Sarkar D, Genetlman R, Huber W (2007) *Genome Biol* 8:1–13.
23. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al.* (2003) *Science* 302:1727–1736.
24. Li L, Huang YW, Xia XF, Sun ZR (2006) *Mol Biol Evol* 23:2467–2473.
25. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) *Science* 296:750–752.
26. Promislow DEL (2004) *Proc R Soc London Ser B* 271:1225–1234.
27. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, *et al.* (2007) *Nat Genet* 39:1256–1260.
28. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) *Mol Biol Evol* 22:1345–1354.
29. Meiklejohn CD, Parsch J, Ranz JM, Hartl DL (2003) *Proc Natl Acad Sci USA* 100:9894–9899.
30. Nation JL (2002) *Insect Physiology and Biochemistry* (CRC, Boca Raton, FL).
31. Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, Sharakhova MV, Unger MF, Collins FH, Feyereisen R (2002) *Science* 298:179–181.
32. Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, Tijet N, Perry T, Heckel D, Batterham P, *et al.* (2002) *Science* 297:2253–2256.
33. Pittendrigh B, Reenan R, ffrenchConstant RH, Ganetzky B (1997) *Mol Gen Genet* 256:602–610.
34. Enayati AA, Ranson H, Hemingway J (2005) *Insect Mol Biol* 14:3–8.
35. Mueller JL, Ram KR, McGraw LA, Qazi MCB, Siggia ED, Clark AG, Aquadro CF, Wolfner MF (2005) *Genetics* 171:131–143.
36. Ram KR, Wolfner MF (2007) *Integr Comp Biol* 47:427–445.
37. Prachumwat A, Li WH (2006) *Mol Biol Evol* 23:30–39.
38. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) *Genome Biol* 3:1–9.
39. Nguyen DQ, Webber C, Ponting CP (2006) *PLoS Genet* 2:198–207.
40. Noor MAF, Feder JL (2006) *Nat Rev Genet* 7:851–861.
41. Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) *Hum Mol Genet* 15:57–66.
42. Hoffmann AA, Sgró, CM, Weeks AR (2004) *Trends Ecol Evol* 19:482–488.
43. Pollack JR (2002) in *Microarrays: A Molecular Cloning Manual*, eds Bowtell D, Sambrook J (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), pp 363–369.
44. Smyth GK (2005) in *Computational Biology Solutions Using R and Bioconductor*, eds Genetlman R, Carey V, Dudoit S, Irizarry R, Huber W (Springer, New York), pp 397–420.
45. Team RDC (2007) R (R Foundation for Statistical Computing, Vienna, Austria).
46. Smith CD, Shu SQ, Mungall CJ, Karpen GH (2007) *Science* 316:1586–1591.
47. Hey J, Kliman RM (2002) *Genetics* 160:595–608.
48. Gnad F, Parsch J (2006) *Bioinformatics* 22:2577–2579.
49. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) *Nucleic Acids Res* 34:535–539.
50. Batagelj A, Mrvar A (1998) *Connections* 21:47–57.
51. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B (2004) *Genome Biol* 5:1–8.
52. Davis JC, Petrov DA (2004) *PLoS Biol* 2:318–326.
53. Chintapalli VR, Wang J, Dow JAT (2007) *Nat Genet* 39:715–720.