Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster

Andreas Massouras1*, Sebastian M. Waszak1*, Monica Albarca-Aguilera1, Korneel Hens1, Wiebke Holcombe1, Julien F. Ayroles2, Emmanouil T. Dermitzakis3, Eric A. Stone4, Jeffrey D. Jensen5, Trudy F. C. Mackay4, Bart Deplancke1*

1 Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 2 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, 3 Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, 4 Department of Genetics, North Carolina State University, Raleigh, North Carolina, United States of America, 5 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Abstract

Understanding the relationship between genetic and phenotypic variation is one of the great outstanding challenges in biology. To meet this challenge, comprehensive genomic variation maps of human as well as of model organism populations are required. Here, we present a nucleotide resolution catalog of single-nucleotide, multi-nucleotide, and structural variants in 39 Drosophila melanogaster Genetic Reference Panel inbred lines. Using an integrative, local assembly-based approach for variant discovery, we identify more than 3.6 million distinct variants, among which were more than 800,000 unique insertions, deletions (indels), and complex variants (1 to 6,000 bp). While the SNP density is higher near other variants, we find that variants themselves are not mutagenic, nor are regions with high variant density particularly mutation-prone. Rather, our data suggest that the elevated SNP density around variants is mainly due to population-level processes. We also provide insights into the regulatory architecture of gene expression variation in adult flies by mapping cis-expression quantitative trait loci (cis-eQTLs) for more than 2,000 genes. Indels comprise around 10% of all cis-eQTLs and show larger effects than SNP cis-eQTLs. In addition, we identified two-fold more gene associations in males as compared to females and found that most cis-eQTLs are sex-specific, revealing a partial decoupling of the genomic architecture between the sexes as well as the importance of genetic factors in mediating sex-biased gene expression. Finally, we performed RNA-seq-based allelic expression imbalance analyses in the offspring of crosses between sequenced lines, which revealed that the majority of strong cis-eQTLs can be validated in heterozygous individuals.

Introduction

An important challenge in biology is to elucidate the relationship between genetic and phenotypic variation [1]. The increasing availability of comprehensive genome sequences of both human [2,3] and model organism populations [4,5] constitutes an important step towards meeting this challenge. The Drosophila Genetic Reference Panel (DGRP) is an example of such a recently emerging population resource, consisting of 192 sequenced wild-derived inbred Drosophila melanogaster lines [6,7]. Drosophila is a premier model organism to understand genome function given the availability of powerful and cost-effective genetic tools and resources [8–10]. In addition, its genome is small, but highly polymorphic [7,11–15], which has already proven helpful in studying the molecular basis of morphological evolution [16,17]. Moreover, linkage disequilibrium (LD) decays quickly across the genome [7,18], which is favorable to elucidating the relationship between genotypic and phenotypic variation at high resolution.

This requires that we genotype all classes of segregating variants (i.e., insertions, deletions, and structural variants) and not only the commonly studied single nucleotide polymorphisms (SNPs) to investigate their effect on phenotypes. Here, we use an integrative approach to derive and characterize a genome-wide, nucleotide-resolution catalog of variants including SNPs, insertions, deletions, complex substitutions, and structural variants in 39 DGRP lines. We then investigate the impact of naturally occurring genetic variation on adult gene expression, revealing novel insights into the regulatory architecture of gene expression variation in Drosophila.

Results

Variant discovery

We generated a catalog of sequence variants using whole-genome Illumina next-generation sequencing data from 39 inbred lines.
**Author Summary**

One of the principal challenges in current biology is to understand the relationship between genetic and phenotypic variation. The increasing availability of genomic variation maps of human as well as of model organism populations (mouse and Arabidopsis) constitutes an important step towards meeting this challenge. However, despite its excellent track record as a premier model to understand genome function, no genome-wide variation data beyond single-nucleotide variants and microsatellites are currently available for D. melanogaster. Here, we present a comprehensive, nucleotide-resolution catalogue of variants of various types (single-nucleotide, multi-nucleotide, and structural variants) for 39 wild-derived inbred D. melanogaster lines based on high-throughput sequencing. This catalogue confirms that non-SNP variants account for more than half of genomic variation, allowing us to provide new insights into the non-random distribution of variants in the Drosophila genome. We further present genome-wide cis-associations with gene expression based on whole adult fly microarray data, revealing significant associations for about 2,000 genes. Most associations are sex-specific, providing evidence for a decoupling of the genomic, regulatory architecture between males and females.

lines from the Drosophila Genetic Reference Panel (DGRP) for which gene expression data are available for young adult flies [6]. First, we used PrInSeS-G [19], which uses de novo local assembly to generate a preliminary list of variant calls for each DGRP line; this tool works by first detecting regions with a fluctuation in sequencing coverage, then using a short fragment from the reference sequence as a seed to build a contig by extending it using the reads; the contig extending stops when another short fragment from the reference sequence is encountered. We then re-aligned each resulting genome against the reference genome to present variants in a coherent way among lines and to reduce variant fragmentation. Finally, we developed and used a genotyping algorithm, which uses the combined variant call set for all DGRP lines to improve variant discovery. This is particularly useful for genomes with low read coverage (Figure S1). This strategy allowed us to identify more than 3.6 million unique sequence variants across all 39 DGRP lines including 2.8 million SNPs, 0.6 million indels, and 0.2 million complex sequence variants (Figure 1A, Table 1). To validate our variant calls, we used five distinct approaches. First, we compared our variant catalogue to indel data from the FLYSNPdb database [20]. We validated 713 deletions and 923 insertions on the breakpoint and complete sequence level, thus covering in total 45% of the FLYSNPdb content. Second, we compared our SNP calls with those published by Mackay et al. [7], and found that 94% of calls made in that study match (Table S1). Third, we used whole genome Roche-454 reads available for the same genomes to validate our SNP calls, resulting in confirmation of 98.7% of SNPs compared (Table S2). Fourth, we sequenced mRNA from three DGRP lines (see also below) and differentially aligned the reads to their respective transcriptomes and the reference transcriptome in order to validate variants in coding sequence. Thus, we examined, on average, 147,453 SNPs and 12,084 non-SNPs per line that were covered by mRNA sequencing reads, confirming on average 98.9% of all variant types within coding sequence for these lines (Table S3). The false discovery rate (FDR) is 0.5% for SNPs and 1.5% for indels and complex variants. Moreover, the FDR remains stable for indels up to 30 bp in size but increases to 8.9%

for larger indels. Note that some variants could not be reliably assigned to the true or false positive categories due to low coverage. Finally, we validated 92 large indels (200 bp to 5.6 kb) in five DGRP lines using PCR (true positives 77.9%, false positives 7.2%, variant size different from predicted 14.9%, Table S4). Moreover, Sanger sequencing of 10 randomly selected large indels validated by PCR revealed a high breakpoint accuracy, whereby breakpoints of 9 out of 10 indels were perfectly reconstructed consistent with results based on query of smaller indels in the FLYSNPdb database. Together, these results indicate that our variant catalogue is of similar quality as variant data produced for other organisms using high-throughput sequencing since an overall false discovery rate of less than 10% and an overall accuracy of ~95% or higher for all variant types (i.e., SNPs, indels and complex variants) is in line with numbers reported by these other studies [4,21–23].

**Indels and complex variants**

Among all lines, we identified a SNP every 43 bp and an indel or complex variant every 144 bp, together contributing a genetic marker every 33 bp on average. These findings illustrate a remarkably high density of molecular polymorphisms in Drosophila consistently greater than in humans [3,24] and mice [4]. Deletions and complex variants affect in total 4.2% (3.0 Mb) of the reference euchromatic genome while insertions add 2.1 Mb (of which 0.5 Mb are in insertions ≥100 bp not found with similarity of 90% or above elsewhere in the reference genome). Non-SNP variants are thus a substantial source of genomic variation in Drosophila. Indel size ranges from 1 to 6,082 bp (Table 1, Figure 1B). Single base pair indels are abundant (32%) and small indels (2–10 bp) represent 50% of the indel repertoire. 6,419 structural variants (≥100 bp) represent 1% of all indels with a median size of 208 bp and encompass ∼2.5 Mb of sequence in total. 850 structural variants are homologous (≥90% sequence similarity) to another part of the assembled reference genome, representing either “young” variants or polymorphic forms of segmental duplications (i.e., copy number variants) [25]. Most (70%) of those structural variants correspond to a duplcon on the same chromosome, indicating that intra-chromosomal duplication events occur more frequently (Figure 1C). These observations support a recently proposed theoretical model governing the formation of segmental duplications in D. melanogaster, whereby, after a double strand break, a search for an ectopic homologous region which is preferentially located within the same chromosomal region triggers the repair mechanism [26].

The majority (73%) of complex variants are balanced multi-nucleotide substitutions and most (61%) constitute di-nucleotide substitutions. Balanced multi-nucleotide substitutions may arise as single nucleotide substitutions that happen to occur in adjacent positions or as multi-nucleotide mutational events [27]. Using simulation within all 1,000 bp genomic windows in all 39 lines, we find that there are on average 3.3 times more di- and 16.3 times more tri-nucleotide substitutions than the number of single nucleotide substitutions expected to be adjacent by chance, indicating that complex mutational events contribute substantially to balanced multi-nucleotide substitutions.

The allele frequency spectrum of indels and complex variants decays steeply, with 28.5% of all indels and complex variants present in only one line (Figure 1D). Comparison to neutral spectra (Figure S2) revealed an excess of low frequency alleles for deletions, consistent with previous results suggesting that more of these variants are under purifying selection [13]. We also observed a depletion of high frequency variants among indels and complex variants in protein-coding regions, splice junctions, and UTR
sequences (Figure S3), again suggesting the action of pervasive purifying selection. In contrast, we found 452 deletions, 873 insertions, and 824 complex variants present among all 39 lines (Table S5). To evaluate whether these variants represent rare alleles in the reference genome, population-specific variants, or artifacts in our variant calling workflow, we focused on all 30 such insertions and deletions that were present in protein-coding regions. We found that 77% of these variants exactly recapitulate an ancestral allele (D. simulans, D. yakuba, or D. ananassae) and 90% when allowing at most one mismatch between the indel and the ancestral allele (Table S6). Therefore, these variants predominantly constitute rare alleles in the reference genome. The photoreceptor gene Rh6 is an interesting example: evolutionary conservation analysis around the two observed 17 bp and 2 bp insertions in its coding sequence revealed that both insertions perfectly match the ancestral allele of seven out of 11 Drosophila species (Figure S4). Moreover, the resulting gene model supports an Rh6 cDNA clone of the OregonR/white strain, thus revealing that the reference genome harbors a rare 19 bp deletion, which truncates the gene.

Indels and SNPs are not uniformly distributed across the genome, with autosomal centromeric regions as well as the X chromosome containing fewer variants compared to other genomic regions (Figure 1C). Several models have been proposed to explain this pattern, ranging from purifying selection to recurrent hitchhiking to demography [7,28,29], although we cannot rule out that low coverage or read mapping quality issues in those regions has affected variant discovery. Moreover, we found that genome-wide SNP, indel, and complex variant densities are strongly correlated on a 1 Mb scale (Spearman’s $r_{SNP}$, $r_{del} = 0.927$, $r_{SNP-indel} = 0.90$, $r_{SNP-complex} = 0.946$; Figure S5), suggesting that similar higher-order constraints such as local selection intensities, recombination, or mutation rates may act on all types of variants. To test the impact of recombination, we tested SNP, indel, and complex variant densities against recombination rates (1 cM/Mb), and found that recombination is strongly positively correlated with variant densities (Spearman’s $r_{SNPs} = 0.54$, $r_{insertions} = 0.55$, $r_{deletions} = 0.62$, $r_{complex} = 0.46$), which is line with long-held observations regarding the actions of both pervasive positive selection and background selection in Drosophila [30,31]. We further examined the SNP distribution around detected indels in more detail, as it has been proposed that indels may act as “mutators” of surrounding sequences [32,33]. Consistent with this hypothesis, we observed that the SNP density is elevated in close proximity to indels independent of indel type, dropping quickly to background levels (Figure 2A, 2B, and 2C). On average 26% (40%) of all SNPs per line are within 40 bp (100 bp) of a non-SNP variant. An increase in SNP density around indels has previously
<table>
<thead>
<tr>
<th>Categories</th>
<th>Conditions</th>
<th>Total number (unique in 39 lines)</th>
<th>Average number per line</th>
<th>Mean total nd per line</th>
<th>Mean total ni per line</th>
<th>Mean nd per variant</th>
<th>Mean ni per variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP - Single Nucleotide Polymorphism</td>
<td>nd = ni = 1</td>
<td>2,829,654</td>
<td>615,258</td>
<td>615,258</td>
<td>615,258</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Indel - Insertion/Deletion</td>
<td>nd = 0 or ni = 0</td>
<td>3,667,449</td>
<td>761,417</td>
<td>1,438,915</td>
<td>1,134,852</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>Insertion</td>
<td>nd = 0</td>
<td>617,856</td>
<td>104,681</td>
<td>697,031</td>
<td>393,905</td>
<td>6.7</td>
<td>3.8</td>
</tr>
<tr>
<td>Homopolymer</td>
<td>repeat size = 1, resulting homopolymer size $\geq 3$</td>
<td>92,967</td>
<td>16,436</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Microsatellite</td>
<td>$2 \leq$ repeat size $\leq 4$, resulting number of repeats $\geq 4$</td>
<td>15,603</td>
<td>2,346</td>
<td>-</td>
<td>9,336</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tandem Repeat</td>
<td>repeat size $&gt; 4$, resulting number of repeats $\geq 2$</td>
<td>11,234</td>
<td>2,126</td>
<td>-</td>
<td>15,978</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Segmental Duplication / Copy Number Variant</td>
<td>ni $\geq 100$ blat 90% match, 98% identity, bitscore $\geq 300$</td>
<td>315</td>
<td>30</td>
<td>-</td>
<td>12,429</td>
<td>-</td>
<td>414</td>
</tr>
<tr>
<td>Deletion</td>
<td>ni = 0</td>
<td>354,440</td>
<td>52,454</td>
<td>697,031</td>
<td>-</td>
<td>13.3</td>
<td>-</td>
</tr>
<tr>
<td>Homopolymer</td>
<td>repeat size = 1, reference homopolymer size $\geq 3$</td>
<td>105,675</td>
<td>17,803</td>
<td>25,805</td>
<td>-</td>
<td>1.4</td>
<td>-</td>
</tr>
<tr>
<td>Microsatellite</td>
<td>$2 \leq$ repeat size $\leq 4$, reference number of repeats $\geq 4$</td>
<td>17,661</td>
<td>2,256</td>
<td>9,105</td>
<td>-</td>
<td>4.0</td>
<td>-</td>
</tr>
<tr>
<td>Tandem Repeat</td>
<td>repeat size $&gt; 4$, reference number of repeats $\geq 2$</td>
<td>14,309</td>
<td>1,688</td>
<td>12,867</td>
<td>-</td>
<td>7.6</td>
<td>-</td>
</tr>
<tr>
<td>Segmental Duplication / Copy Number Variant</td>
<td>nd $\geq 100$ blat 90% match, 98% identity, bitscore $\geq 300$</td>
<td>11,234</td>
<td>2,126</td>
<td>-</td>
<td>15,978</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Complex</td>
<td>ni &gt; 0 and nd &gt; 0</td>
<td>219,939</td>
<td>41,478</td>
<td>126,626</td>
<td>125,690</td>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td>MNS - Multiple Nucleotide Substitution</td>
<td>nd = ni &gt; 1</td>
<td>161,432</td>
<td>30,973</td>
<td>69,799</td>
<td>69,799</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Inversion</td>
<td>nd = ni $\geq 4$, sequence replaced by its reverse complement</td>
<td>511</td>
<td>98</td>
<td>447</td>
<td>447</td>
<td>4.6</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table 1. Type classification of variants.
been observed between species [32,33]; here, we present genome-wide evidence of this effect within a single species. Thus, the accuracy of a traditional SNP-calling method based on simple read alignment may suffer materially since a considerable portion of the reads that would contribute to SNP calling may not be mapped correctly as they also contain indels; this problem is circumvented by the integrative variant calling approach employed here.

Interestingly, we observed a similar SNP density increase around other SNPs (Figure 2A), suggesting that either SNPs are also mutagenic or, more likely, that alternative explanations need to be considered for the increased SNP density around variants. We plotted the allele frequencies of SNPs around variants and found that high allele frequency SNPs cluster around other high allele frequency variants (including other SNPs) and, correspondingly, low frequency variants cluster together (Figure 2B and 2C). These findings provide compelling evidence that variants in general are likely not mutagenic, since otherwise we would expect to observe a greater number of rare, thus mostly recent variants, around high frequency alleles. Alternatively, this phenomenon could occur if some genomic regions are more susceptible to mutation than others [32,34]. We therefore examined the SNP density in the same DNA regions, but in those of the 39 lines which do not contain the focal variants. We found no material enrichment in SNP density in these regions (Figure 2A, 2B, and 2C), providing little evidence of locally increased mutation rate.

**Functional impact of indels and complex variants**

Non-coding regions that are conserved between species contain significantly fewer variants than other non-coding regions (Figure 2D, Mann-Whitney U test \( P = 2.2\times10^{-16} \)), confirming that these regions experience intra-species purifying selection as well. We obtained similar results when integrating histone modification data from the modENCODE Project [9]. We observe fewer indels affecting regulatory genomic regions - especially promoters (marked by tri-methylation of histone H3 at lysine 4, H3K4me3) - than unmarked non-coding regions, (Mann-Whitney U test \( P = 3.1\times10^{-14} \)) (Figure S6).

While the majority (>90%) of indels and complex variants fall into intronic or intergenic regions, 80% of all protein-coding genes are affected by indels or complex variants (Table 2). However, only 13% of non-SNP variants within coding regions result in exon disruptions or full gene deletions. More than 50% of indels leading to frame shifts are singletons versus 35% for non-frame shifting indels, indicating that purifying selection is acting against these
likely deleterious mutational events. This is illustrated by our
observation that the length of indels affecting coding regions is
highly biased for sizes that are multiples of three, a pattern that is
still visible even up to 51 bp (Figure 2E). Consistent with previous
inter-specific comparisons [35], we found that genes affected by
disruptive indels or complex variants, are significantly enriched for
the functional categories of sensory perception of taste and smell,
proteolysis, and innate immune response as well as pathways
related to food and drug metabolism (Tables S7 and S8).

Although the density of variants around genes is lower than the
overall background, we observed an increase in variant density
almost to background levels directly upstream of the transcription
start site (TSS), followed by a dramatic drop across the TSS
(Figure 3A). This is consistent with strong selective constraint at
the TSS but not the region immediately upstream, possibly
because this region has high AT content and it is typically depleted
of nucleosomes [36,37]. Indeed, we observed a strong correlation
between the local AT content and indel density around the TSS
(Figure S7A), strengthening earlier observations that sequence
context or chromatin structure may affect the indel rate [38,39].
We observed a similar, albeit less striking effect at the transcription
end site (TES) (Figure 3B and Figure S7B).

Identification and characterization of cis-eQTLs

We used published whole adult microarray gene expression data
[6] to perform association analysis between the expression and
variants within 10 kb of each gene. We initially identified 9,789
(9,434) genetically variable transcripts (FDR<0.001) in males
(females) after removing probes from the microarray analysis that
contain genomic variation. We then used the Kruskal-Wallis test
followed by permutation-based multiple test correction to find
significantly associated variants, termed cis-expression QTLs (cis-
eQTLs), which may point to underlying functional regulatory
elements. QTL studies have so far mostly associated bi-allelic
variants with gene expression levels [40,41]; given the availability
of a high-resolution, comprehensive catalog of sequence variants,
many of which (9.3%) are multi-allelic within the DGRP
population, we grouped expression levels according to the
corresponding allele as an input to each test. We conducted
~3.8 million tests for each sex, and restricted our association
analysis to variants present in at least three lines. We found 17,501
cis-eQTLs in 2,033 genes (26% of the 7,889 genes tested) at a false
discovery rate of ~10% (Figure 3C, 3D, and 3E; Tables S7 and
S10); and results for 1% and 20% FDR thresholds are listed in
Table S11), generating to our knowledge the first cis-eQTL map in
Drosophila.

Surprisingly, the majority of cis-eQTLs were found to be sex-
specific (58% specific to males and 17% to females; Figure 3C and
3D, and an example can be found in Figure 4A) with males having
more than two-fold more cis-eQTL-associated genes compared to
females. This result is not an artifact of the significance threshold,
as a plot of the underlying association P-values in both sexes
clearly reveals that the majority of cis-eQTLs are sex-specific
(Figure S8). Further, the variance in gene expression among
females is comparable to that among males as 56% of genetically
variable transcripts without any cis-eQTL-associations display
higher expression variance in females, making it unlikely that such
bias affects our findings. In addition, only 38% of transcripts
with male-specific cis-eQTLs exhibit greater gene expression variance
in females than in males, refuting the hypothesis that the more
than two-fold greater number of male- compared to female-
specific cis-eQTL-associated genes may be due to a greater
variability in gene expression in females compared to males.

Intriguingly, we found that male-specific cis-eQTL-associated
genes are significantly depleted from the X chromosome
($P=3.6e-11$, $\chi^2$ test), whereas female-specific cis-eQTL ($P=0.17$
and unbiased ($P=0.02$) genes show a more uniform chromosomal
distribution, perhaps consistent with the observed depletion of
male-biased genes on the X chromosome [42,43]. Sex-unbiased
cis-eQTL-associated genes were found to have more cis-eQTLs
than sex-biased ones (Figure S9) and the effect of cis-eQTLs found
in both sexes is larger (Mann-Whitney U test, $P=2.2e-16$; Figure
S10). Sex-unbiased cis-eQTL-associated genes have also a greater
tendency to be expressed only in somatic tissues (Figures S11 and
S12), possibly explaining why the underlying variants affect gene
expression in both sexes. Nevertheless, more than 20% of sex-
specific cis-eQTL-associated genes are also only expressed in
somatic tissues, indicating that sex-biased changes in gene
expression are pervasive [44]. Furthermore, we found that 20%
of male-specific cis-eQTL-associated genes are exclusively
expressed in the testes or the accessory gland in males, whereas
the expression of only few (5%) female-specific counterparts are
restricted to the ovaries or spermathecae in females. Approximately
60% of male-specific cis-eQTL-associated genes are expressed in female reproductive tissues (whereby 28% having
greater expression than in any somatic tissue) (Figure S13), yet
these genes do not yield any cis-eQTLs in females. In other words,
despite the fact that these genes are expressed both in males and
females, genetic mutations only produce cis-eQTLs in males,
indicating that these mutations do not significantly affect gene
expression in females. Together with the lower number of female-
specific cis-eQTL-associated genes in general, this suggests that the

---

### Table 2. Potential functional indels and complex variants in genes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Deletions</th>
<th>Insertions</th>
<th>Complex variants</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-frame</td>
<td>5,113 (2,582)</td>
<td>3,001 (1,568)</td>
<td>6,611 (3,614)</td>
<td>14,725 (5,499)</td>
</tr>
<tr>
<td>Frameshift</td>
<td>1,514 (1,011)</td>
<td>489 (364)</td>
<td>243 (188)</td>
<td>2,246 (1,313)</td>
</tr>
<tr>
<td>Gene disruption</td>
<td>85 (82)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>85 (82)</td>
</tr>
<tr>
<td>Stop gain</td>
<td>13 (13)</td>
<td>72 (56)</td>
<td>0 (0)</td>
<td>85 (69)</td>
</tr>
<tr>
<td>Stop loss</td>
<td>6 (6)</td>
<td>4 (4)</td>
<td>0 (0)</td>
<td>10 (9)</td>
</tr>
<tr>
<td>Splice-site disruption</td>
<td>80 (74)</td>
<td>79 (75)</td>
<td>24 (23)</td>
<td>183 (168)</td>
</tr>
<tr>
<td>UTR</td>
<td>20,357 (6,728)</td>
<td>15,020 (5,538)</td>
<td>8,026 (4,052)</td>
<td>43,403 (8,783)</td>
</tr>
<tr>
<td>Total</td>
<td>27,168 (8,456)</td>
<td>18,665 (6,655)</td>
<td>14,904 (6,533)</td>
<td>60,737 (11,186)</td>
</tr>
</tbody>
</table>

Number in brackets indicates the number of genes affected.
doi:10.1371/journal.pgen.1003055.t002
underlying regulatory architecture may be more constrained in females as compared to males.

We found that 5% of all cis-eQTL-associated genes do not exhibit detectable expression in any tissue of the corresponding sex (as represented in FlyAtlas [45]), indicating that the underlying genetic variation induces previously unreported gene expression in young adult flies. These cis-eQTLs show an enhanced effect compared to genes previously known to be expressed in adult fly tissues (Figure S14), thus strengthening the finding that these genes become strongly expressed given a specific genetic background. An example of this phenomenon involves Hsc70-2. This gene, ranking second among all female-specific associations, is known to be highly expressed in testes of adult males only [45,46]. However, we find that females from several DGRP lines also exhibit high Hsc70-2 expression (Figure 4A), an observation which was validated using RNA-seq-based allele imbalance analyses (see below).

Genomic distribution of cis-eQTLs

Although we considered only cis-associations, we identified 1,162 cis-eQTLs (6.6%) associated with two genes, and 58 with three genes. We found on average seven (median three) significant cis-associations per transcript (males and females combined) with the physical distance between two consecutive cis-eQTLs being 689 bp (median 121 bp), thus narrowing putatively causal variants down to a few hundred base pairs. However, cis-eQTLs associated with the expression of the same gene were found to be often in strong linkage disequilibrium (average $r^2 = 0.88$).

We examined the genomic location of significant associations relative to TSSs and TESs (Figure 3A and 3B). Both the TSS and TES show the most significant associations independent of the variant type. We observed a quasi-symmetric distribution around TSSs in marked contrast to TESs, where upstream associations are, on average, more significant than downstream associations. These findings are consistent with previous studies in other metazoans including humans [40,47,48]. The high density of genetic markers in this study affords however greater resolution.
Figure 4. Examples of cis-eQTLs and their associated genes. DGRP lines in (A) and (B) are grouped by their allele. Male and female expression levels are depicted in blue and dark pink, respectively. (A) Sex-biased cis-eQTL. A SNP (3R:8,875,391) is associated with higher gene expression levels in females only. (B) Indel-based cis-eQTLs associated with gene expression. Two insertions (7 bp, 3L:332,512; 1 bp, 3L:332,594, $r^2 = 0.20$) are associated with markedly different expression levels in males and females. (C) cis-association overview. Plot illustrating the variant and association data for a single gene (mthl9) on a rolling window basis. The gene is shown on the top track, with UTRs in grey and coding regions in black. Significant cis-eQTLs are drawn below and color-coded by significance for each sex separately (red most significant). Linkage ($r^2 = 0.5$) is shown by arcs, color-coded according to $r^2$, with higher values in red. Rows represent all 39 DGRP lines and the left column shows gene expression levels for each line and sex separately (red indicates the highest expression level and green the lowest). The grid contains a representation of variants in rolling 50 bp windows (successive windows overlapping by 45 bp) with net insertions in red, net deletions in blue, and variants not affecting the sequence size (mostly SNPs) in black. The height of each variant indicates the net size of variants with the window, up to 20 bp. The two shaded vertical bars mark the cis-eQTLs shown in (B).

doi:10.1371/journal.pgen.1003055.g004
clearly positioning the most significant associations within a window of less than 500 bp around the TSS and 1 kb upstream of the TSS, consistent with findings in yeast [49]. The results are also consistent with enrichment for significant associations in H3K4me3 regions (2.6-fold for females and 2.6-fold for males) as compared to the rest of the genome (Table S12). Non-coding conserved regions are not enriched for cis-eQTLs, in agreement with findings in humans, where eQTLs are underrepresented in such regions [50]. One explanation could be that variants in conserved non-coding regions that dramatically affected gene expression were purged from the genome. Regions marked by trimethylated histone H3 at lysine 27 (H3K27me3) are underrepresented for cis-eQTLs, consistent with the modification being a heterochromatin mark [51]. However, while we observe an enrichment of cis-eQTLs in functionally annotated regions, the majority of cis-eQTLs are outside of previously annotated regulatory regions. High-resolution maps of these cis-eQTL-derived putative regulatory elements (Figure 4C) may in this regard constitute a powerful resource for dissecting the regulatory architecture of gene expression in further detail.

Inheritance of gene expression
To validate cis-eQTLs genome-wide, we performed mRNA sequencing (RNA-seq) of reciprocal F1 female hybrids from two crosses involving three DGRP lines. Specifically, we aimed to identify whether transcripts predicted to be regulated by cis-eQTLs exhibit a significant allele-specific bias in gene expression. Since both alleles act in the cross in the same trans environment, differential expression in the F1 is a direct measure of cis-regulatory activity [52,53]. As a quality control step, we first analyzed the correlation of the expression data between the reciprocal crosses, obtaining a Spearman correlation coefficient of >0.99 for both F1362–765-initial/F1362–765-reciprocal and F1517–765-initial/F1517–765-reciprocal, respectively. We then tested on average ~7,100 transcripts for allelic imbalance in the two crosses. Using a 10% FDR, we found that 6% (443) of the tested transcripts show significant allele-specific gene expression (Figure 5A) with a median allelic expression difference of 1.5-fold. Of the cis-eQTLs for which the parental lines have different alleles, and whose population mean effect on microarray expression was two-, three- and four-fold among all 39 lines, we found that on average 39%, 60% and 75% of the associated transcripts respectively exhibited a significant allelic imbalance in the crosses (Figure 5B). These results are in line with similar, albeit smaller-scale analyses in yeast [49] and mouse [48], and thus provide support for our cis-eQTL map, revealing that the majority of strong cis-eQTLs can be validated in heterozygous individuals.

Discussion
We present a comprehensive, genome-wide list of variants of all major types (i.e., single-, multi-nucleotide-, and structural variation) for a Drosophila wild-derived population, significantly expanding the catalog of variants that has been compiled to date for this model organism (e.g., [7,13–15,20]). Our analyses reveal extensive non-SNP variation among lines, with a high indel frequency (roughly every 150 bp) and indels in 80% of the protein-coding genes. The majority of indels do not alter the reading frame, but nevertheless reveal surprising tolerance for indels in coding regions. Our data are relevant to the inference of underlying mechanisms of varying mutation rates across genomes [32]. We show elevated SNP density around other SNPs as well as indels; that neighboring variants have similar allele frequencies; and that regions with indels are not particularly mutation-prone. These data suggest that this phenomenon reflects the existence of small haploblocks. While these may arise because of segregating mutational hotspots, a more likely explanation invokes the demographic history of the population. For example, admixture may generate such a haplotype structure to an extent largely dictated by the local genomic recombination environment [54]—potentially consistent with our observed correlation between recombination rate and variant density. Thus, while our data suggest that these population level processes may lead to incorrect inference regarding the mutagenic nature of indels, this will be a topic of further investigation in future studies during which alternative explanations (e.g., duplication hotspots) will also be explored.
The presented variant collection constitutes a useful resource for several scientific disciplines, including phylogenetics and population genetics, which are beyond the scope of this study. Here, we applied this resource to study the impact of natural polymorphisms on gene expression, identifying cis-eQTLs for >2,000 genes (10% FDR). Although further work is necessary to identify the causal variants underlying cis-eQTLs, the combined results provide general insights into the complexity and evolution of the cis-regulatory architecture of gene expression in Drosophila. More specifically, the data suggest that gene expression variation is governed by multiple cis-eQTLs of different variant types, that the most significant cis-eQTLs are tightly clustered around the TSS and immediately upstream of the TES, but that many cis-eQTLs are also located in regions currently devoid of regulatory annotations. These regions likely play an important role in gene regulation as their alteration has an expression impact that can be observed in whole adult fly expression profiles. In addition, the data suggest that the regulatory architecture is partly decoupled between males and females, at least within the tested settings, as underlying changes predominantly affect gene expression in males. However, cis-eQTLs common to both sexes typically have larger effects. This supports the hypothesis that intersexual ontogenetic conflict in D. melanogaster may stem from many sexually antagonistic alleles with low effect sizes [55]. Finally, the detection of many sex-specific cis-eQTLs indicates that genetic factors play an important role in sex-biased gene expression as changes in cis-regulatory sequences appear to significantly contribute to this phenomenon. While the relative significance of these changes and cis-regulatory variation in general in governing morphological evolution remains an important topic of debate [56,57], we believe that the presented catalog of cis-regulatory changes constitutes a valuable resource to further illuminate this discussion.

Materials and Methods

Variant prediction

We used Release 5 of the Drosophila melanogaster reference genome for all analyses.

Stage 1: Variant discovery. We used PrInSeS-G [19], in conjunction with BWA [58] to perform the initial variant discovery using whole-genome Illumina sequencing reads.

Stage 2: Variant refinement. We computed whole-genome alignments between the chromosomes of each line and the reference genome in order to reduce the amount of variant fragmentation and to optimize the representation of variants among lines (see Text S1).

Stage 3: Variant genotyping. The purpose of genotyping is to improve variant discovery for DGRP lines or genomic regions whose sequencing coverage was originally too low for effective de novo assembly by PrInSeS-G. The algorithm is described in detail in Text S1, and Figure S16.

Gene expression analysis

Whole-adult gene expression microarray raw data files (i.e., .CEL) for 39 inbred lines were downloaded from the EBI ArrayExpress Archive (accession number E-MEXP-1594). Briefly, Ayroles et al. [6] derived inbred lines from the Raleigh (USA) population by 20 generation full-sib mating and hybridized RNA from two independent pools (25 flies/line/sex) on Affymetrix Drosophila 2.0 microarrays. The raw data set consists of gene expression measurements for males and female in two replicates. We used a four-step pipeline to analyse the microarray data: Step 1. We obtained a custom probe set definition file (http://brainarray.mhnj.med.umich.edu/Brainarray/Database/CustomCDF/13.0.0/refseq.asp) for the Affymetrix Drosophila 2.0 platform, which is based on an updated set of 20,666 transcripts (UCSC RefSeq; June 22, 2010). We excluded probes that overlap with any type of sequence variants and probe sets that were covered by less than three probes after probe removal. In addition, we verified whether polymorphic duplicate regions may have contributed to gene expression variation. We aligned all insertions (>100 bp) to the annotated D. melanogaster transcriptome with the alignment program blat, retaining only high sequence similarity hits (>90%), since those would be the most likely candidates to cause problems with cross-hybridization on microarrays. We found only two insertions that could be mapped against annotated transcripts with both insertions targeting the same exon, demonstrating that duplication does not constitute an important confounding factor.

Stage 2. Raw microarray data were normalized across all sexes and lines with the Robust Multichip Average (RMA) algorithm as implemented in the R affy package (default settings).

Stage 3. We removed transcripts that were not or lowly expressed among lines using the Wilcoxon signed rank-based gene expression presence/absence detection algorithm (affy package, R). A transcript was classified as expressed in a single line if it was detected in either one or both sexes, respectively, and the requirement to be detected in both replicates. Finally, we only kept 16,985 probe sets (or 10,347 genes) that were classified as expressed in at least 4 out of 39 lines.

Stage 4. We used ANOVA on each transcript and sex separately to test whether it is genetically variable, i.e., has a significant line term, under a conservative FDR of 0.001 [6]. In total, we found 9,789 (6,239) and 9,434 (5,797) genetically variable transcripts (genes) in males and females, respectively, of which 6,745 (4,147) were variable in both sexes.

Tissue-specific gene expression

Tissue-specific gene expression data for young adult flies was obtained from FlyBase (ftp://flybase.org/flybase/associated_files/Gelbart.2010.10.13.tar.gz). A gene was considered to be expressed in a tissue if the expression level was above 50.

Cis-associations

We grouped overlapping variants for all 39 lines as alternative alleles. For each such group of variants, we calculated associations with any transcript whose either end is within 10 kb, and for which microarray expression data was available. We employed two statistical methods: the first one, Alignment Score Association, is linear regression between the rank of alignment scores of all alleles and the rank of expression; in this context we defined alignment score to be the maximum of the number of bases inserted and that removed by the variant. Note that, since we used a non-parametric model, there was little advantage of using a more complex alignment score. This method thus associates the size of the variants rather than the exact genotype, which is important for the cases where more than two alleles have variants of different lengths. For the second method, Allele Association, we grouped the rank of expression of each line by its allele at each variant locus; we then performed a Kruskal-Wallis test. For both methods we corrected for multiple tests by repeating each test for 10,000 random permutations of gene expression, in a similar fashion to [40]. We then took for each permutation the lowest P-value for all tests in the same transcript, thus obtaining a vector of 10,000 such P-values. This vector was then sorted, and we obtained the adjusted P-value threshold from the value indexed by our unadjusted threshold. For example, the adjusted (multiple-test corrected) P-value threshold for a nominal threshold of 0.05 is the...
variants, hence we only presented the results of the Allele
approximately the same significance measure for bi-allelic
methods was more than 95% for all metrics, as both give
time. We note that the overlap between the two association
of the above workflow will produce a slightly different list of
threshold is used for each transcript for the same nominal
actually passes. This, of course, implies that a different adjusted
imbalance in F1
Variant validation
We used five complementary methods to validate our variant
calls. First, we compared our variant catalogue against publicly
available indel data from the FLYSNPdb database, which features
indels (1 bp to 360 bp) from five previously established D.
melanogaster lines (but none of the DGRP lines) [20]. Second, we
compared our calls with those published by Mackay et al. [7]. We
used two methods (comparing all variants and comparing SNPs
not within 30 bps of non-SNPs, Table S1). In many cases, SNP
calls by the other study overlapped or were close to indel calls,
which indicates a possible false positive since methods based on
direct read alignment are confounded by indels. Third, we aligned
Rho-454 whole genome sequencing reads from Mackay et al.
available for all 39 lines [7] to validate our SNP calls (Massouras et
al., in preparation), including those that (a) were originally
supported by two thirds or more of the Illumina reads and (b)
were covered by at least two Rho-454 reads (Table S2). Fourth,
we sequenced mRNA from three DGRP lines at the young adult
stage (see below and main manuscript for results) (Table S3).
Finally, we examined 92 large indels (200 bp to 5.6 kb) in five
DGRP lines using PCR after which we randomly selected 10 for
subsequent Sanger sequencing to validate the predicted variant
breakpoints (Table S4).

Variant validation using RNA-seq and allelic expression
imbalance in F1
All fly stocks were grown at 25°C and a 12 h light-dark cycle on
corn-meal fly medium. We collected virgin flies from three of the
DGRP lines (lines 362, 517, and 765) during three days and we set
up the following crosses: R 362×517 (male) and reciprocal and Q 517× Q
765 (and reciprocal). We performed RNA-seq to assess transcript
expression profiles of 3–5 day old F1 females. 25 flies per
sample were frozen between 1–3 pm. Total RNA was extracted using
the combined Trizol/RNeasy (Qiagen, http://www.qiagen.
com/) protocol. RNA quality was measured using RNA Labchips
and Bioanalyzer from Agilent Technologies (http://www.chem.
agilent.com/). RNA-seq libraries were prepared using the RNA-
True seq kit (Illumina). Prepared libraries were sequenced with an
Illumina Genome Analyzer 2 DNA Sequencing Platform (GTG,
Lausanne). We used the UCSC transcript annotation to derive the
sequences of the annotated transcriptome from the reference
Genome. We mapped the reads to both the genomes and the
annotated transcriptomes; any reads aligning to more than one
genomic mapping coordinate were discarded. The tool we built
specifically for this purpose takes as input the sequencing reads,
two genome haplotypes, and the UCSC-annotated transcript
coordinates in order to derive two transcriptomes. For every read
in turn it attempts to align it to both transcriptome haplotypes.
Reads that align to the same position in a transcript of the two
haplotypes are then checked to see if the alignment overlaps with
any variants. If it does, and the variants are different between the
two haplotypes, the read is deemed to “support” the allele that
produces the lowest number of mismatches. If the read aligns to
only one transcriptome haplotype, it is also deemed to support the
Corresponding allele. In this way the tool measures allele-specific
differential gene expression in the same line.

For the parental lines, we used the aforementioned tool,
supplying it with the Stage 3 genome (see Variant prediction
paragraph for more details) and the reference genome. For the
crosses we gave it the two parental Stage 3 genomes. In each case,
for each position with a variant we counted the number of reads
best aligning to each allele. As a result, for the parental lines, this
method provides a measurement of the true positive and false
positive rate with regard to our variant discovery, since it measures
the differential alignment of reads to the reference sequence
and the target haplotype of the line. For the F1s, this method provides
a measurement for differential expression for each transcript.

For variant validation, we considered a variant true positive
(TP) when either all aligned reads support this variant, as opposed
to the reference or when the number of reads that support it
compared to that supporting the reference results in P<0.05 in a
two-tailed binomial test. We considered a variant a false positive
(FP) when either all aligned reads supported the reference at this
position, or the number supporting the reference was significant
using the same binomial test.

Allelic expression imbalance analysis
We tested for significant allelic expression imbalance between
alleles in F1 using binomial exact tests (two-sided). About half of all
tested transcripts were covered by at least 200 discriminative
reads, thus providing a reasonable power to detect even small
allelic imbalances (≥1.5-fold allelic expression differences). P-
values were corrected for multiple hypothesis testing with the
Benjamini & Hochberg procedure (implemented in the function
p.adjust, R). Only transcripts that were tested for cis-associations in
females and that were covered by at least 20 informative reads (i.e.,
allele1+allele2≥20) were tested for significant allelic imbalance.
Transcripts that passed with 10% FDR in both reciprocal crosses
were considered to have a significant allelic imbalance in gene
expression.

The increase in percent-validated (Figure 5B) was tested by
permutation analysis. For each “cis-eQTL strength cutoff” we
obtained a random set of transcripts from all tested cis-eQTL
transcripts) and calculated how many of those where validated in
F1. We repeated this sampling procedure 1,000 times and
obtained an empirical P-value for each “cis-eQTL strength cutoff”
by counting how many random transcript sets scored higher than
the real set of transcripts and divided by the number of
permutations (i.e., 1,000).

Recombination estimates
The Fiston-Lavier [59] recombination rate calculator was used to
estimate the recombination rate (cM/Mb) in 1 Mb windows
along each chromosome (only chromosome 4 excluded). The rate
at the center of each interval was used.

Genomic distribution of cis-eQTL-associated genes
Genes were grouped into four different categories (i.e., genes
with no-, male-specific-, female-specific-, and sex-unbiased
cis-eQTLs) and within each category the percentage of genes located on
the X chromosome was calculated. For cis-eQTL associated genes,
we calculated whether the fraction of X-linked genes deviates
significantly from non-cis-eQTL associated genes using the
Chi-square test with Yates’ correction.
Data availability
The full catalog of variants can be obtained from dgrp.epfl.ch/downloads. The RNA-seq data are available in the ArrayExpress Database using accession number E-MTAB-1266.

Supporting Information
Figure S1 Results of variant genotyping. Size (reference bases in thousands, left y-axis) of variants after Stage 2 (re-alignment, yellow) and Stage 3 (genotyping, blue), ordered by sequencing coverage (red line, right y-axis). This figure illustrates that the genotyping stage improves the consistency of variant calling materially, particularly for low-coverage genomes.

Figure S2 Number of variants by minor allele count. The number of insertions, deletions, complex variants and segmental duplications/copy number variants discovered is plotted next to the number expected under the neutral hypothesis. Deletions and especially segmental duplications are more enriched for low allele counts, which suggests that they are more under negative selection than the other variant types.

Figure S3 Allele-frequency spectrum of non-SNP variants by genomic location.

Figure S4 Evolutionary conservation analysis of insertions in the Rh6 coding sequence. Both insertions were predicted in all 39 lines and are also present in seven out of eleven Drosophila species other than D. melanogaster. Moreover, the resulting gene model supports an Rh6 cDNA clone of the OregonR/white strain. Thus, the reference genome has a likely rare allele, which disrupts a splice site and introduces a premature stop codon resulting in a truncated protein.

Figure S5 Genome-wide correlation between SNP and non-SNP densities. The concentration (variants per kb) of SNPs is correlated with deletions (blue), insertions (red), and complex variants (orange). Densities were calculated in non-overlapping genomic bins of 50 kb across all autosomes and the X chromosome.

Figure S6 Variant concentration in histone modified regions. The histone marks for adult flies were obtained from modENCODE. SNPs are in black/grey, indels and complex variants in purple.

Figure S7 Variant density and AT content (A) near TSS, (B) near TES. The inlet shows the same plot between 10 kb up- and downstream of respectively the TSS and TES (spline-smoothed). Blue dots depict variant density and red dots AT content (%).

Figure S8 Comparison of \( P \)-values for sex-specific associations. (a) Female-specific and (b) male-specific association \( P \)-values have mostly no significant counterpart for the other sex, ruling out the possibility that they labeled sex-specific only for marginally failing to meet the significance threshold.

Figure S9 Correlation between sex-bias and number of \( \text{cis} \)-associations. Transcripts were grouped by the number of \( \text{cis} \)-eQTLs and the y-axis indicates the percentage of transcripts for which \( \text{cis} \)-associations were detected in one sex only. Linear regression fits are plotted for both sexes separately.

Table S1 Comparison of SNP calls to those by Mackay et al., Nature, 2012.

Table S2 Validation of SNPs using Roche-454 whole-genome sequencing data.

Table S3 Validation of variants in exons by RNA-Seq.
testing was performed with the Benjamini & Hochberg correction procedure and a FDR of 0.05. Genomic Variation and Impact on Gene Expression

Table S4  Validation of 92 variants >200 bp using PCR in five DGRP lines. (XLSX)
Table S5  Allele count frequency distribution by variant type. Average number of variants per line. (XLSX)
Table S6  Evolutionary conservation analysis of 30 indels. This table contains information about the ancestral allele state of all indels that were predicted in 39 lines and which intersected with protein coding sequences. Multiple sequence alignment around the indel position was obtained from the UCSC genome browser (database: dm3, table: multiz15way) and D. simulans, D. yakuba, and D. ananassae were used as an outgroup. (XLSX)
Table S7  Functional enrichment (GO) for genes affected by disruptive non-SNP variants. Gene ontology enrichment (biological processes) was tested with DAVID (http://david.abcc.ncifcrf.gov/). Multiple hypothesis testing was performed with the Benjamini & Hochberg correction procedure and a FDR of 0.05. (XLSX)
Table S8  KEGG pathways for genes affected by disruptive non-SNP variants. Enrichment in KEGG pathways was tested with DAVID (http://david.abcc.ncifcrf.gov/). Multiple hypothesis testing was performed with the Benjamini & Hochberg correction procedure and a FDR of 0.05. (XLSX)
Table S9  List of significant associations detected in females. (XLSX)
Table S10  List of significant associations detected in males. (XLSX)
Table S11  Comparison of cis-eQTLs by false discovery rate threshold. (XLSX)
Table S12  Concentration of variants and cis-eQTLs in selected regions. Histone marks for adult females and males were obtained from modENCODE. (XLSX)
Text S1  Supplementary Information. (DOC)

Acknowledgments

The authors wish to thank Maria Gutierrez-Arcelus, Zoltan Kutalik, Frederic Schuitz, and Daniel Wegmann for their assistance and very helpful discussions, and Vital-IT High Performance Computing for their computational resources.

Author Contributions

Conceived and designed the experiments: AM BD. Performed the experiments: MA-A KH WH. Analyzed the data: AM SMW. Wrote the paper: AM SMW BD. Experimental and technical support as well as discussion: JFA JDJ ETD EAS TFCM BD.

References


