Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines

Peter D. Keightley,[1] Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L. Blaxter

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

We inferred the rate and properties of new spontaneous mutations in *Drosophila melanogaster* by carrying out whole-genome shotgun sequencing-by-synthesis of three mutation accumulation (MA) lines that had been maintained by close inbreeding for an average of 262 generations. We tested for the presence of new mutations by generating alignments of each MA line to the *D. melanogaster* reference genome sequence and then compared these alignments base by base. We determined empirically that at least five reads at a site within each line are required for accurate single nucleotide mutation calling. We mapped a total of 174 single-nucleotide mutations, giving a single nucleotide mutation rate of 3.5×10^{-9} per site per generation. There were no false positives in a random sample of 40 of these mutations checked by Sanger sequencing. Variation in the numbers of mutations among the MA lines was small and nonsignificant. Numbers of transition and transversion mutations were 86 and 88, respectively, implying that transition mutation rate is close to 2× the transversion rate. We observed 1.5× as many G or C → A or T as A or T → G or C mutations, implying that the G or C → A or T mutation rate is close to 2× the A or T → G or C mutation rate. The base composition of the genome is therefore not at an equilibrium determined solely by mutation. The predicted G + C content at mutational equilibrium (33%) is similar to that observed in transposable element remnants. Nearest-neighbor mutational context dependencies are nonsignificant, suggesting that this is a weak phenomenon in *Drosophila*. We also saw nonsignificant differences in the mutation rate between transcribed and untranscribed regions, implying that any transcription-coupled repair process is weak. Of seven short indel mutations confirmed, six were deletions, consistent with the deletion bias that is thought to exist in *Drosophila*.

[Supplemental material is available online at www.genome.org.]

The rates and properties of spontaneous mutations are important for many questions in evolutionary biology and molecular evolution. For example, under neutrality, the rate of molecular evolution is expected to be equal to the mutation rate, so between-species molecular divergence can be used to date divergence times of species by assuming clock-like molecular evolution. Conversely, the rate of molecular divergence at silent sites between species can be used to estimate the mutation rate. However, this requires the assumption of neutrality, and values for the generation time and divergence dates of the species are also needed.

Mutation accumulation (MA) experiments are an alternative way to directly study new mutational variation. MA lines are started by subdividing a homozygous progenitor strain, then allowing spontaneous mutations to accumulate, often for many tens of generations. The lines are maintained by a form of close inbreeding (typically full-sib mating or selfing) that reduces the effectiveness of natural selection, so the rate of fixation of mutations is expected to be close to the mutation rate. The classic method to analyze MA experiments uses information from the phenotypic values of MA lines. For example, the mutation rate per genome can be estimated based on the changes of the mean and between-MA line variance for fitness over t generations of mutation accumulation (Bateman 1959; Mukai 1964). However, it is known that this method tends to underestimate the genomic mutation rate (Lynch and Walsh 1998). To study new mutations more directly, it is now feasible to search for mutations that arise in the genomes of MA lines using mutation detection technology or sequencing. MA-based molecular estimation of the mutation rate per site was pioneered by Mukai and Cockerham (1977), who searched for new allozyme variants in *Drosophila melanogaster* MA lines. However, only three band-morph variants were found, so the estimate of the mutation rate was imprecise. More recently, technology to scan parts of MA lines genomes for new mutations has been applied to mitochondrial and nuclear genomes of *Caenorhabditis elegans* (Denver et al. 2000, 2004) and *D. melanogaster* (Haag-Liautard et al. 2007, 2008). However, only very small proportions of each nuclear genome were scanned, and relatively few mutations were detected. The emergence of new high-throughput sequencing technologies makes it feasible to obtain nearly complete genome sequences for many organisms, including complex eukaryotes. New mutations can then be detected by among-MA line genome sequence comparison. For example, shotgun pyrosequencing has recently been used to obtain the genome sequences of MA lines of yeast (Lynch et al. 2008), and this has enabled estimation of the per-nucleotide mutation rate and the mutation spectrum.

Here, we report the shotgun sequencing of three *D. melanogaster* MA lines that had undergone an average of 262 generations of spontaneous mutation accumulation (Fernandez and López-Fanjul 1996). The MA lines are a subset of the lines that we previously studied using mutation detection by denaturing high-performance liquid chromatography (DHPLC) coupled with direct sequencing (Haag-Liautard et al. 2007, 2008). In our previous study, our estimate of the mutation rate per base-pair was substantially higher than an estimate based on between-species silent

[1]**Corresponding author.**
E-mail Keightley.genomeres2009@gmail.com; fax 44-(0)-131-650-6564.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.091231.109.

site divergence, and our estimate of nuclear genome-wide deleterious mutation rate (*U*) exceeded one event per generation. Here, we use Illumina (formerly Solexa) sequencing-by-synthesis (Bentley et al. 2008) to an average depth of coverage of 11 reads per site per line. We align sequencing reads from each MA line to the *D. melanogaster* reference genome using two aligners—MAQ and Novoalign—that use somewhat different algorithms. MAQ performs ungapped alignment, allowing two or three mismatches in the first 28 bases of each read (Li et al. 2008). Without paired-end reads, as is the case here, it is feasible only to map single-nucleotide mutations with MAQ. Novoalign aligns sequencing reads using the more computationally intensive Needleman-Wunsch algorithm with affine gap penalties, so there is the potential to map insertion–deletion (indel) mutations. From the genome alignments, we call mutations by comparing the genome alignments base by base. We verify a subset of the single-nucleotide mutations and the indel mutations by Sanger sequencing of PCR products. We also verify most of the single-nucleotide mutations by checking low-quality reads not used in the initial mutation calling. We scan a far higher proportion of the genome than our previous study, and our data set is sufficiently large as to allow tests of among-line variation in the mutation rate, accurate estimation of the transition:transversion ratio, and tests for chromosomal and context-specific variation in the mutation rate. We also find clear evidence for nonrandom errors among Illumina reads.

Results

Alignment of sequencing-by-synthesis reads to the reference *D. melanogaster* genome using MAQ

Sequencing reads of 36 or 50 bases in length were aligned to the reference *D. melanogaster* genome version 5.9 using the MAQ aligner (Li et al. 2008), allowing up to three mismatches per read to generate MAQ-3 alignments. The mean read depth for base reads that MAQ was able to align at any quality is 11.1, and this is fairly consistent across the three lines sequenced (Table 1). About 10% of base reads were unmappable at any quality threshold. The distributions of numbers of reads per base in MAQ-3 alignments are shown in Supplemental Figure 1. Excluding locations with read depth 0, the variance of read depth is an average of 2.2× higher than the mean depth. There is therefore substantially more variation in depth than expected under a Poisson distribution. MAQ reports the Illumina-derived base quality and a mapping quality for each base read. Base qualities in Illumina reads are analogous to *phred* scores (Ewing and Green 1998) and indicate the probability of the base call being an error. Mapping qualities are derived from the number of matches of identical quality found in the whole-genome reference and indicate the probability that a read truly maps to the position indicated. Low mapping quality scores indicate that there were >1 sites in the genome where the read could be placed. In all cases analyzed, we used high mapping quality (i.e., ≥40) and generated high base quality (*b20*; base quality ≥ 20) and low base quality (*b5*; base quality ≥ 5) MAQ-3 alignments. Mean read depths for these are 8.1 and 9.9, respectively (Table 1). The consensus of the three sequences differed from the *D. melanogaster* reference sequence at 0.27% of sites.

To obtain an empirical estimate of the sequencing error rate after filtering by MAQ, we computed the fraction of reads at a site in each line that disagree with the most frequent nucleotide at the site. We limited this analysis to sites at which there are ≥5 reads, since this is the threshold above which we empirically determined that mutations can be called with high confidence (see below). The estimated error rates per base read for the MAQ-3 *b20* and *b5* alignments are 1.2×10^{-3} and 1.0×10^{-2}, respectively. Many errors are therefore filtered from the data by MAQ, particularly at the higher base quality, because the sequencing error rate from the instrument is higher than these values. Within-line heterozygosity makes only a trivial contribution to these error rate estimates, because of the rarity of spontaneous mutations (Haag-Liautard et al. 2007) and the low effective population size within each MA line (i.e., two in this case), which causes mutations to have very short persistence times.

Calling of nuclear genome single-nucleotide mutations

To identify candidate mutations, we compared the three MA line MAQ-3 *b20* nuclear genome alignments site by site. A candidate mutation was called if there was a difference in the reads between the lines and all reads within each line were in complete agreement or near complete agreement (see Methods). We counted numbers of candidate mutations and numbers of sites where all three MA lines had a valid consensus nucleotide (as defined in Methods) and recorded the minimum read depth at the site among the three lines. An estimate of the mutation rate per site is

$$\mu = \frac{no.\ of\ mutations\ called}{nt \times no.\ of\ sites}, \quad (1)$$

where n is the number of MA lines (i.e., 3), and t is the mean number of generations of mutation accumulation (i.e., 262). Mutation rate estimates as a function of minimum read depth are shown in Table 2. Estimated mutation rate is negatively related to read depth, clearly suggesting that the number of differences between lines is inflated by sequencing or mapping errors. The mutation rate estimates stabilize, however, for a read depth of 5 and above: The mutation rates for read depths 4 and ≥5 are

Table 1. Total number of bases and mean depth of coverage for all Illumina reads and two MAQ-3 genome alignments that use different base quality thresholds

	MAQ-3 alignment					
	All reads		*b20*		*b5*	
MA line	Bases	Mean depth	Bases	Mean depth	Bases	Mean depth
M126	1,253,080,339	10.4	928,458,566	7.71	1,139,342,821	9.46
M138	1,466,086,352	12.2	1,034,688,889	8.60	1,301,730,659	10.81
M158	1,271,228,590	10.6	948,325,280	7.88	1,144,690,323	9.51

Mean depth excludes sites at zero coverage.

Table 2. Numbers of candidate single base mutations detected, numbers of sites, and estimated mutation rate for sites at which three lines have a valid nucleotide in the MAQ-3 *b20* genome alignments

Depth	No. of mutations called	No. of sites	Mutation rate × 10^{-9}
2	448	7,101,073	80.27
3	57	10,232,862	7.09
4	50	13,023,269	4.88
5	42	14,642,636	3.65
6	33	14,631,553	2.87
7	30	13,069,142	2.92
8	27	10,518,971	3.27
9	21	7,659,559	3.49
10	7	5,112,505	1.74
11	8	3,081,450	3.3
12	3	1,700,030	2.25
13	1	872,906	1.46
14	1	416,031	3.06
15	1	188,520	6.75
16	0	83,586	0
17	0	38,308	0
18	0	19,691	0
19	0	11,453	0
≥20	0	32,229	0
Total ≥ 5	174	72,078,570	

Depth is the minimum read depth at each site among the three MA lines.

significantly different (Fisher's exact test; $P = 0.005$), whereas the rates for depths 5 and ≥6 are nonsignificantly different ($P = 0.19$). Assuming that sequencing or mapping errors occur randomly at a rate of 1.2×10^{-3} per base per read (as estimated above for the MAQ-3 *b20* alignment), the expected number of errors for sites with a minimum read depth of 2 is 10.2 (i.e., 7.1×10^6 [the number of such sites, Table 2] $\times$ $[1.2 \times 10^{-3}]^2$), whereas the observed number is 448. Similarly, the expected number of errors for sites with a minimum read depth of 3 is 0.02, whereas the observed number is 57. These excesses of observed numbers over expectation indicate that there must be a contribution from nonrandom sequencing or mapping errors, i.e., some sites are substantially more likely than others to have nucleotide miscalls. Using sites called at a minimum read depth of ≥5, the total number of mutations detected is 174, and an estimate of the mean mutation rate per site is 3.07×10^{-9}.

The above mutation-calling algorithm is likely to generate false negatives and underestimate the mutation rate because mutations are almost always associated with mismatches to the reference sequence, so reads containing mutations are less likely to be aligned at high quality. This is evident in our data, because mean read depth for mutants is less than that for wild type at sites where a mutation is called at depth ≥5 (Table 3). Sites containing mutations will therefore have a higher probability of falling below the threshold of five reads that we set for accepting a mutation. We used the difference in read depth between mutants and wild types to correct for missing mutations by resampling the data (see Methods). This predicts that the mutation-calling algorithm missed about 13% of mutations from the MAQ-3 alignments. The corrected single base mutation rate is then 3.46×10^{-9} (approximate 95% confidence limits 2.96×10^{-9} and 4.01×10^{-9}). A somewhat higher fraction of mutations is predicted to be missed from MAQ alignments that allow up to two mismatches per read (MAQ-2; Table 3), but the corrected mutation rate estimate is very similar. Using an alternative aligner (Novoalign), more mutations are called, and fewer are predicted to be missed, and the corrected mutation rate estimate also agrees closely with those from MAQ (Table 3).

We used Sanger sequencing of amplified PCR products to verify a random sample of the single-nucleotide mutations called by MAQ using amplicons containing the candidate mutations in the affected MA lines. Of 40 mutations tested, all were confirmed (Supplemental Table 1).

Checking nuclear genome mutations using the low quality genome alignment

We also generated nuclear genome alignments using a lower base quality cutoff (*b5*; base quality ≥5 in MAQ) and used these to check for possible miscalling of mutations in the high-quality *b20* data set. This *b5* data set has an average of 1.8 more reads per site than the high-quality alignment (Table 1). Although the *b5* data set has a higher overall number of errors, the frequency of errors per site is expected to be quite small. We classified a mutation as "negated" if >1 of the extra reads in the *b5* data agreed with the reference base at that site for the mutant line (rather than the mutant base). If a mutation was not negated, it was "corroborated" if >1 extra reads in the *b5* data agreed with the mutant base. By this criterion, there were no mutations negated, 132 corroborated, and 42 mutations for which the *b5* alignment was uninformative.

Properties of nuclear genome single-nucleotide mutations

For the purposes of this section, we assume that the 174 mutations called by the algorithm described above are genuine. Among the *D. melanogaster* chromosome arms, the mutation rate does not vary significantly (Table 4; χ^2 5 degrees of freedom [df] = 3.1; $P = 0.68$). In particular, the mutation rate does not differ significantly between the autosomes and the X chromosome (Table 4; Fisher's exact test, $P = 0.13$). The numbers of mutations called in the three MA lines were 56, 58, and 60 for M126, M138, and M158,

Table 3. Read depths for mutants and wild types at sites where a mutation was called at read depth ≥5, estimated fractions of mutations missed, and uncorrected and corrected mutation rates obtained from analysis of three genome alignments

		Mean read depth			Mutation rate × 10^{-9}	
Alignment	No. of mutations called	Mutants	Wild type	Fraction of mutations missed	Uncorrected	Corrected
MAQ-3	174	9.12	10.19	0.128	3.07	3.46
MAQ-2	157	8.69	10.14	0.182	2.90	3.55
Novoalign	204	9.92	10.28	0.038	3.44	3.57

MAQ-2 and MAQ-3 alignments allow up to two and three mismatches per read, respectively. They are both at base quality 20 and mapping quality 40.

Table 4. Numbers of mutations called, numbers of sites, and estimates of the mutation rate for the chromosome arms and the genome of *D. melanogaster* for MAQ-3 alignments

Chromosome arm	No. of mutations	No. of sites	Mutation rate × 10^{-9} (uncorrected)
X	18	10,307,812	2.22
2L	36	14,219,624	3.22
2R	34	13,143,584	3.29
3L	42	15,172,378	3.52
3R	42	18,429,633	2.89
4	2	805,539	3.16
Autosomes	156	61,770,758	3.21
Genome	174	72,078,570	3.07

respectively. The variation among the lines in the numbers of mutations called is therefore nonsignificant (χ^2 2 df = 0.14; P = 0.93). The lines were initially of the same inbred genotype, so the lack of significant variation suggests that there was no detectable effect resulting from an accumulation of new mutations that modified the mutation rate. However, we have low power to detect mutation rate variation. For example, we could detect 25% and 50% increases in the mutation rate in one line with 40% and 95% probability, respectively, at the 5% significance level under the assumption of Poisson-distributed mutation numbers.

We polarized the mutations using the major nucleotide called in each of the three MA lines. The matrix of mutational changes is shown in Table 5. Transition mutations make up slightly less than one half of the mutations (i.e., 86 transitions vs. 88 transversions), and this is similar to the ratio of transition to transversion substitutions at synonymous sites between species of the *D. melanogaster* group (Moriyama and Powell 1996). Our data therefore confirm a close to 2:1 transition:transversion mutation bias in *D. melanogaster*. There is a significant excess of mutations that decrease GC content (i.e., 80 G or C → A or T vs. 53 A or T → G or C mutations; χ^2 1 df = 5.5; P = 0.02), which implies that GC content of the genome is not at an equilibrium determined solely by mutation. Whole-genome sequencing in yeast has yielded similar observations (Lynch et al. 2008). The *Drosophila* genome is 43% GC, so the mutation rate from G or C → A or T is 2.0× that from A or T → G or C, and the predicted GC content at mutational equilibrium is therefore 33%. This is similar to the predicted equilibrium content (35%) of putatively neutrally evolving "dead on arrival" transposable elements (Petrov and Hartl 1999; Singh et al. 2005).

To determine whether neighboring bases influence the spontaneous mutation rate, we counted the frequencies of bases preceding and following sites where a mutation had occurred. We treated both DNA strands as equivalent so that, for example, a wild-type A base that mutated to any other base preceded by G is treated as equivalent to a wild-type T that mutated to any base followed by C. The observed numbers of bases preceding or following specific wild-type bases do not differ significantly from expectation (Supplemental Table 2), suggesting that context effects, at least at the level of neighboring bases, are fairly weak. However, the numbers of observations are not large, so these tests lack power.

Single-nucleotide mutations classified by functional category

Using the *D. melanogaster* genome annotation, we classified the nuclear genome mutations according to whether they occurred in a constitutively or alternatively expressed exon, an intron, or intergenic DNA (Table 6). Exonic mutations were classified as synonymous or nonsynonymous. None of the nonsynonymous mutations generated a nonsense change. We computed expected numbers of mutations in the different categories by sampling mutational changes from the mutation matrix in proportion to the relative frequencies of the 12 possible mutational types (Table 5). We sampled a mutation by randomly sampling locations in the genome until the base at a location matched the wild-type base sampled from the mutation matrix. We generated 10,000 such randomly sampled mutations and then used these to calculate the relative frequencies of mutations in the different functional categories by interrogating the genome annotation. The numbers of nonsynonymous mutations and all other mutations (Table 6) are nonsignificantly different from their expectations (χ^2 1 df = 2.2; P = 0.13), but the ~25% deficit of nonsynonymous mutations is suggestive that selection may have prevented the fixation of strongly deleterious amino acid mutations in the MA lines. There is no evidence in our data of a transcription coupled repair process, since observed numbers of mutations in transcribed and intergenic DNA are very similar to expected (Table 6).

Mitochondrial genome mutations

Mean read depths for the mitochondrial genome were 816, 560, and 465 for M126, M138, and M158, respectively (cf. Table 1 for the nuclear genome). We previously employed DHPLC and direct sequencing to scan more than 50% of the mitochondrial genome for new mutations in a superset of the MA lines studied here (Haag-Liautard et al. 2008). Most of the mutations that we detected in our previous experiment were heteroplasmic within a line. In the three MA lines sequenced within the present experiment, our previous experiment detected a single G → A transition mutation at position 10,093 segregating in line M126 at an estimated frequency of 0.14. In this experiment, there are no fixed differences between the three MA lines, which agrees with our previous experiment. We set a lower limit of 200 reads at a site in a line and called a mutation if a minimum of 5% of reads differed from the consensus nucleotide at the site. This revealed two candidate mutations, including the G → A transition at position 10,093 mentioned above, which segregated in line M126 at a frequency of 0.12. This is very similar to our previous estimate. A second candidate mutation at position 18,984 (C → T) appeared to be segregating at frequencies of 60% and 62% in lines M126 and M138, respectively, and at a frequency of 50% in M158, although the number of reads in M158 was only 158. This may therefore represent an extremely mutable hotspot, or, more likely, a mapping artifact. We were unable to scan this site by DHPLC in our previous study (Haag-Liautard et al. 2008), because it is in the AT-rich region and is difficult to amplify.

Table 5. Matrix of numbers of single-nucleotide mutation types obtained by comparing MAQ-3 alignments

	To			
	A	**T**	**G**	**C**
From				
A		10	15	8
T	11		12	18
G	23	16		9
C	11	30	11	

Table 6. Mutations classified by functional category, along with expected numbers in MAQ-3 alignments

Category	No. of mutations	
	Observed	Expected
Nonsynonymous	18	24.0
Synonymous	8	8.67
Intronic	70	64.7
Genic	96	97.4
Intergenic	78	76.6

Nuclear genome short indel mutations

To estimate the nuclear indel mutation rate and characterize the properties of indel mutations, we generated alignments to the reference genome sequence using the Novoalign aligner (Novocraft Technologies), which generates gapped alignments using the Needleman-Wunsch algorithm with affine gap penalties. After processing the output by MAQ, we obtained locations of potential short indels, called relative to the reference genome, for each MA line. We compared these candidate indels to find indels that were unique to one line and called in >90% of reads crossing a putative indel. The estimated mutation rate, obtained by dividing by the mean number of sites per line at a given read depth (Supplemental Table 3), increases steeply at low read depth. This is presumably due to a contribution from errors, following a similar pattern to potential single-nucleotide mutations (Table 2). The numbers of mutations called at high read depth are quite small, so determining an empirical threshold depth for accepting indel mutations is difficult. The rates of mutations for depths 7 and 8 are nearly significantly different (Fisher's exact test, $P=0.06$), whereas contrasts between mutation rates for consecutively higher read depths are nonsignificant, so we set the threshold for accepting indels at a read depth ≥8. However, using this criterion, Sanger sequencing across 35 candidate indel mutations called at a minimum read depth of eight confirmed only seven of these (Table 7; Supplemental Table 4). This suggests that there is a high degree of nonrandom error associated with indel assignment for our data. Among the confirmed indels, deletions outnumber insertions by six to one (Table 7).

Discussion

The single-nucleotide mutation rate estimate from our study is 3.5×10^{-9} per site per generation, based on 174 mutations mapped in 60% of the euchromatic genome. This is similar to our estimate of 2.7×10^{-9} (based on only eight mutations mapped in 0.13% of the genome) that we previously obtained using DHPLC on a superset of the MA lines (Haag-Liautard et al. 2007). The consistency between these estimates lends supports to the contention that there is substantial mutation rate variation among *Drosophila* MA lines of different genotypes, because an estimate of the mutation rate from the Florida-33 *D. melanogaster* MA lines (Houle and Nuzhdin 2004) is 11.7×10^{-9} (Haag-Liautard et al. 2007), and the lower confidence limit for this estimate (5.9×10^{-9}) does not overlap with the upper confidence limit of our present estimate (4.0×10^{-9}). It is not possible to compare indel rates from the two studies, since in the present case we are unable to ascertain the rate for false negatives and estimating the number of sites scanned for indels is problematic. However, these results show clear evidence of a deletion bias, which is consistent with our previous study (Haag-Liautard et al. 2007) and with the pattern of within-species indel polymorphism in *D. virilis* (Petrov et al. 1996). Better prediction of indel mutations using Illumina technology may be possible with paired-end reads and/or longer reads.

Numbers of single-nucleotide mutations detected in each MA line are remarkably similar, which implies that the mutation rate was constant among MA lines over the course of the experiment. There is therefore no detectable effect of new mutation rate modifier mutations. We also failed to detect significant mutation rate variation among the chromosome arms. Although the X chromosome has the lowest mutation rate among the chromosome arms, hinting at male-biased mutation, its rate is nonsignificantly different from the autosomes. Higher nucleotide divergences have been observed on the X chromosome than on autosomes in comparisons of *D. melanogaster* group species (Begun et al. 2007). The magnitudes of these differences are quite small, however, and statistically indistinguishable from our data.

Our data show no significant neighboring base contextual effects on the mutation rate. There seem to be no previous reports of this phenomenon in *Drosophila*, in contrast to mammals where context-dependent mutation, particularly associated with methylated CpG dinucleotides, is important (Hwang and Green 2004; Siepel and Haussler 2004). Concerning the existence of a transcription-coupled repair process, divergences of transposable elements in *Drosophila* (Wang et al. 2007) and murids (Gaffney and Keightley 2006) are about 5%–10% higher in integenic than intronic regions, suggesting that such a process might operate. The numbers of mutations we observed in genic and intergenic categories are consistent with an effect of this magnitude (Table 6), although numbers of mutations are nonsignificantly different from their expectations based on equal mutation rates. We see some evidence for a reduction in the number of amino acid mutations below expectation, implying that a subset of these mutations (about one quarter) are strongly selected against, although this is not formally significant. The selection coefficients against these mutations would need to be greater than the reciprocal of the effective population size in the MA lines (i.e., >1/2) for selection to have an appreciable effect on fixation probability.

Estimates of single base mutation rates are very similar from MAQ and Novoalign aligners if a correction is made for a difference in read depth between mutants and wild type (Table 3). However, our study suggests caution in using Illumina sequencing-by-synthesis for detecting rare SNPs (such as new mutations), because we found very clear evidence of nonrandom error caused by some sites having a higher than average probability of sequencing or mapping error. For example, the fraction of differences at sites showing a between-line difference at a minimum read depth of two is ~20 times higher than sites with a minimum depth of five. The error rate at these low coverage sites is much higher than the

Table 7. Indel mutations called by Novoalign that were confirmed by Sanger sequencing

Depth	Chromosome	Location	MA line	Indel
8	X	11,415,304	M158	−T
9	3L	8,947,586	M126	−CAC
10	2L	14,631,954	M126	−ATCG
11	3R	7,483,420	M126	−TA
11	2R	13,898,306	M158	−G
12	2R	6,468,287	M126	−GT
15	3L	13,103,866	M158	+G

average error per site would indicate, under the assumption of independent errors. This argues against assuming that errors are independent or using sites with fewer than five reads. We sequenced our MA lines to about twice the depth of the yeast MA lines recently sequenced using 454 Life Sciences (Roche) FLX technology (Lynch et al. 2008), and we set a higher minimum depth threshold for including a site in the analysis (five vs. three). Our analysis required higher stringency because the *Drosophila* genome is ~10× larger than the yeast genome and Illumina reads are shorter than Roche FLX. The method does not allow mutations to be mapped in repetitive or low complexity regions, because these are assigned low mapping quality scores. If the mutation process is unusual in these regions, then our estimate of the mutation rate and distribution of types will be biased. An obvious class that is expected to be missed is microsatellite mutations. Mutations in recently duplicated regions (copy number variants) would appear to be heterozygous within lines, so would not be detected. Furthermore, we showed that single-nucleotide mutations are less likely to be mapped to the reference genome than wild-type reads, so reads containing indels would suffer from this problem to an even greater extent. Although imperfect, the extremely high throughput of this and other genome sequencing methods makes previous mutation detection methods redundant, at least for detecting single-nucleotide mutation rate in MA lines. It also brings closer new possibilities, such as the genome sequencing of parents and their offspring (Kondrashov 2008) to estimate the mutation rate in individuals sampled from natural populations and circumvent biases that may arise from mutation accumulation in inbred lines.

Methods

Mutation accumulation lines

A homozygous progenitor for the MA lines ("Madrid lines") was generated with the aid of balancer chromosomes (Caballero et al. 1991). This method should preclude the possibility of residual heterozygosity in the MA line progenitor. We found no evidence of this in our previous study that involved scanning the genome of a large superset of current Madrid MA lines at 277 genomic locations. The MA lines were then maintained by full-sib mating or double first cousin mating until generation 47 and by full-sib mating until generation 262 (Fernandez and López-Fanjul 1996). Genomic DNA samples from pools of 25 flies each from three MA lines (M126, M138, and M158), obtained by Maside et al. (2001), were analyzed in this study. In our previous study, in which we scanned a small proportion of the genome by DHPLC for new mutations (Haag-Liautard et al. 2007), we detected no nuclear genome mutations in the three MA lines chosen for this study.

Whole-genome shotgun sequencing

Genomic DNAs from the three MA lines were used as template to prepare libraries for Illumina sequencing (Bentley et al. 2008), following the manufacturers' protocols. Random reads of lengths 36 or 50 bases were generated on an Illumina GAI instrument. Each MA line was sequenced until it was estimated that ~10× coverage in high-quality reads had been achieved (~0.75 of a GAI flowcell run).

Alignment to the reference *D. melanogaster* genome

We used MAQ (version maq-0.6.8_x86_64-linux) to align shotgun reads from each MA line to 120,381,546 euchromatic bases in the *D. melanogaster* version 5.9 genome, while specifying a mapping quality of ≥40 and two different base qualities. MAQ alignments were performed using default parameters, except that we varied the number of mismatches allowed per read. Most of the analysis we report refers to MAQ alignments that allow up to three mismatches per read (denoted MAQ-3). We also examined alignments that allow up to two mismatches per read (denoted MAQ-2). The outputs were converted to "pileup" format for single-nucleotide mutation calling. In this format, each line corresponds to a base of the reference genome and gives the numbers of reads of each nucleotide that align with it, along with their base and mapping qualities.

We also used Novoalign (version 1.06) to align reads from each MA line to the reference genome. The reference sequence was indexed using "novoindex" with *k*-mer length = 14 and step size = 2, otherwise using default parameters. The output was then converted to MAQ's ".map" format, and this was then further converted to "pileup" format for single-nucleotide mutation calling. Short indels were predicted using the "indelpe" command of MAQ.

Calling of mutations

We compared each site in the three MA line genome alignments in turn. For each line, we assigned a valid consensus nucleotide if the same nucleotide was present at ≥90% of reads, otherwise the consensus nucleotide at that site was flagged as invalid. The results are hardly affected if 100% agreement is enforced (data not shown). We then compared sites for which there were valid consensus nucleotides in all three MA lines. A candidate mutation was called if the consensus nucleotide of one line disagreed with the consensus nucleotides of the other two lines, which themselves had to agree. The minimum number of reads among the lines with valid consensus nucleotides at the site was recorded.

Because mutations almost always mismatch with the reference sequence, reads containing mutations are less likely than wild types to be aligned at high quality. We confirmed this by calculating the mean read depth for sites containing mutants called at a depth of ≥5 and their corresponding wild types (Table 3). Let δ be the difference in mean read depth between wild types and mutants at these sites. To estimate the effect of an underrepresentation of mutant reads on the number of mutations called, we sampled 1,000,000 sites at random from each of the three MA line genome alignments. For sites where the number of reads $r_1 \geq 5$, we computed $r_2 = r_1 - x$, where x is a Poisson deviate with parameter δ. The fraction of such sites at which $r_2 < 5$ is our estimate of the fraction of mutations missed.

Checking of mutations by Sanger sequencing

We checked a random sample of 40 of the single-nucleotide mutations by sequencing PCR products that included the sites of the candidate mutations in each of the three MA lines. In the case of indels, we sequenced the line containing the candidate indel mutation and one of the other two MA lines.

Acknowedgments

We thank Carlos López-Fanjul and Aurora García-Dorado for generously providing samples of *D. melanogaster* MA lines; and Aurora García-Dorado, Cathy Haag-Liautard, and Donald Smith for helpful comments on the manuscript. Illumina sequencing was performed in the *Genepool*, which is funded by the Darwin Trust of Edinburgh, the UK Natural Environment Research Council, the Scottish Universities Life Science Alliance, and the School of Biological Sciences, University of Edinburgh.

References

Bateman AJ. 1959. The viability of near-normal irradiated chromosomes. *Int J Radiat Biol* **1:** 170–180.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Pohl Y-P, Hahn M, Nista PM, Jones CD, Kern AD, Dewey CD, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol* **5:** e310. doi: 10.1371/journal.pbio.0050310.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Caballero A, Toro MA, López-Fanjul C. 1991. The response to artificial selection from new mutations in *Drosophila melanogaster*. *Genetics* **127:** 89–102.

Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK. 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* **289:** 2342–2344.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430:** 679–682.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res* **8:** 186–194.

Fernandez J, López-Fanjul C. 1996. Spontaneous mutational variances and covariances for fitness-related traits in *Drosophila melanogaster*. *Genetics* **143:** 829–837.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genet* **2:** e204. doi: 10.1371/journal.pgen.0020204.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445:** 82–85.

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. 2008. Direct estimation of the mitochondrial DNA mutation rate in *D. melanogaster*. *PLoS Biol* **6:** e204. doi: 10.1371/journal.pbio.0060204.

Houle D, Nuzhdin SV. 2004. Mutation accumulation and the effect of *copia* insertions in *Drosophila melanogaster*. *Genet Res* **83:** 7–18.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci* **101:** 13994–14001.

Kondrashov AS. 2008. Another step toward quantifying spontaneous mutation. *Proc Natl Acad Sci* **105:** 9133–9134.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci* **105:** 9272–9277.

Maside X, Bartolome C, Assimacopoulos S, Charlesworth B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: In situ hybridization vs Southern blotting data. *Genet Res* **78:** 121–136.

Moriyama EN, Powell JR. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* **13:** 261–277.

Mukai T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics* **50:** 1–19.

Mukai T, Cockerham CC. 1977. Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster*. *Proc Natl Acad Sci* **74:** 2514–2517.

Petrov DA, Hartl DL. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci* **96:** 1475–1479.

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

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21:** 468–488.

Singh ND, Arndt PF, Petrov DA. 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169:** 709–722.

Wang J, Keightley PD, Halligan DL. 2007. Effect of divergence time and recombination rate on molecular evolution of *Drosophila* INE-1 transposable elements and other candidates for neutrally evolving sites. *J Mol Evol* **65:** 627–639.

Received January 15, 2009; accepted in revised form April 28, 2009.