Meiotic Gene Conversion Tract Length Distribution Within the *rosy* Locus of *Drosophila melanogaster*

Arthur J. Hilliker,* George Harauz,* Andrew G. Reaume,† Mark Gray,‡ Stephen H. Clark† and Arthur Chovnick†

**Department of Molecular Biology and Genetics, University of Guelph, Guelph, Ontario, Canada N1G 2W1, †Department of Molecular and Cell Biology, The University of Connecticut, Storrs, Connecticut 06269-2131, and ‡Division of Reproductive Endocrinology, Department of Obstetrics and Gynecology, Tufts University School of Medicine, Boston, Massachusetts 02111*

Manuscript received August 10, 1993
Accepted for publication April 21, 1994

ABSTRACT

Employing extensive co-conversion data for selected and unselected sites of known molecular location in the *rosy* locus of *Drosophila melanogaster*, we determine the parameters of meiotic gene conversion tract length distribution. The tract length distribution for gene conversion events can be approximated by the equation $P(L \geq n) = \phi^n$ where P is the probability that tract length (L) is greater than or equal to a specified number of nucleotides (n). From the co-conversion data, a maximum likelihood estimate with standard error for ϕ is 0.99717 ± 0.00026, corresponding to a mean conversion tract length of 352 base pairs. (Thus, gene conversion tract lengths are sufficiently small to allow for extensive shuffling of DNA sequence polymorphisms within a gene.) For selected site conversions there is a bias towards recovery of longer tracts. The distribution of conversion tract lengths associated with selected sites can be approximated by the equation $P(L \geq n \mid \text{selected}) = \phi^n(1 - n + n/\phi)$, where P is now the probability that a *selected* site tract length (L) is greater than or equal to a specified number of nucleotides (n). For the optimal value of ϕ determined from the co-conversion analysis, the mean conversion tract length for selected sites is 706 base pairs. We discuss, in the light of this and other studies, the relationship between meiotic gene conversion and *P* element excision induced gap repair and determine that they are distinct processes defined by different parameters and, possibly, mechanisms.

MEIOTIC gene conversion is a non-reciprocal transfer of genetic information from one homologous non-sister chromatid to another resulting in non-Mendelian segregation ratios in individual meiotic tetrads. A fraction of gene conversions are also crossovers (*i.e.*, physical exchanges, recombinant for flanking markers in fine structure mapping experiments). In *Drosophila melanogaster* the *rosy* (*ry*) locus has proven instrumental in the study of intragenic recombination and gene conversion. There is a powerful selection system which allows large progeny samples to be screened for rare recombinant events including non-crossover associated conversions of *ry* locus selected marker sites (CHOVNICK *et al.* 1970).

In addition to the availability of a large number of mutant lesions within the *rosy* locus that are subject to selective recombination analysis, there are many non-selective polymorphic sites spread across the gene that serve as useful markers in the analysis of selected site recombinants [reviewed in HILLIKER and CHOVNICK (1981) and HILLIKER *et al.* (1988)]. Recent molecular analysis and sequencing of numerous *rosy* locus mutations and several wild-type alleles have identified the precise DNA sequence site changes that are the basis for both the selective and nonselective markers available for such studies (LEE *et al.* 1987; KEITH *et al.* 1987; GRAY *et al.* 1991; CURTIS and BENDER 1991; CURTIS *et al.* 1989).

Recently, DNA sequence studies of *rosy* locus intragenic recombinants (CURTIS *et al.* 1989; CURTIS and BENDER 1991) have contributed significantly to our appreciation of the nature and extent of these recombination events. For the present report, two points of interest are to be noted. (1) Of greatest significance is the demonstration that all of the conversions analyzed are consistent with the notion that gene conversions are continuous tracts of DNA, previously inferred only from genetic data (CHOVNICK *et al.* 1971; HILLIKER and CHOVNICK 1981). (2) Restricting attention to conversions not associated with crossovers, CURTIS *et al.* (1989) and CURTIS and BENDER (1991) determined lower and upper limits for the individual tract lengths of 27 conversions taken from two earlier recombination studies (CLARK *et al.* 1984; CARPENTER 1984) and excluding from consideration another 13 tract lengths obtained from meiotic mutant genotypes. Pooling these results of these studies, CURTIS and BENDER (1991) estimated an average conversion tract length of 1161 bp. Since only conversions that select for a functional *rosy* locus are recovered, these authors recognize possible sources of error in their estimate and offer a corrected estimate of 885 bp as the average conversion tract length.

In this report, we fit the conversion data collected over many years in the CHOVNICK laboratory to a model in which the conversion tract lengths follow a geometric

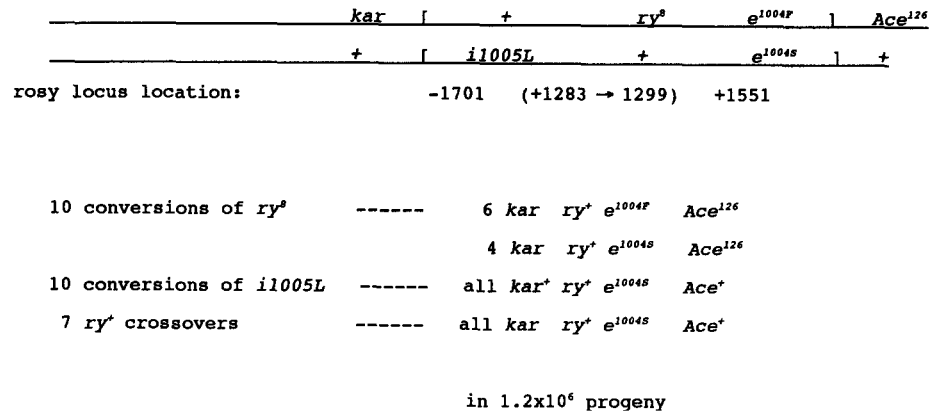

FIGURE 1.—An example of a *rosy* locus intragenic recombination experiment in which selected site conversions (ry^8) can be assayed for co-conversion of an unselected site (*e1004*). (See text for detailed discussion of this specific experiment.)

distribution. This model is fundamentally similar to that developed by GLOOR *et al.* (1991) in the analysis of gap repair path lengths following *P* element excision. The present report describes the model and employs this data base (involving 306 conversions recovered in experiments that sampled 44×10^6 progeny) to estimate the mean length of conversion tracts, and, indeed, to derive the actual frequency distribution of meiotic conversion tract lengths.

Recently, W. R. ENGELS and his colleagues have described a process of double-strand gap repair following *P* element transposase induced *P* element excision in Drosophila which bears a resemblance to meiotic gene conversion (ENGELS *et al.* 1990; GLOOR *et al.* 1991; NASSIF and ENGELS 1993; JOHNSON-SCHLITZ and ENGELS 1993). The relationship between this process and meiotic gene conversion is discussed below in the light of the results of the present study.

THE GENETIC SYSTEM

The *rosy* (*ry*) locus is located in the right arm of chromosome *3* of *D. melanogaster* at map position 52.0 approximately 5 cM from the centromere. Closely flanking markers are the *karmoisin* (*kar*) and *Acetylcholinestrase* (*Ace*) loci at map positions 51.7 and 52.2, respectively.

Intragenic recombination experiments involve large scale crosses of *rosy* heteroallelic females to tester males, and the progeny are reared on a selective medium containing purine (7*H*-imidazo[4,5-*d*]pyrimidine). The *rosy* locus encodes a peptide which, as a homodimer, functions as the enzyme xanthine dehydrogenase (XDH). The biochemical and regulatory features of this gene are reviewed elsewhere (DUTTON and CHOVNICK 1988). *rosy* mutant alleles are recessive, conditional lethals. Mutant individuals survive and reproduce vigorously on standard Drosophila media; however, they are unable to complete development on medium supplemented with an appropriate concentration of the selective agent, purine.

An example of a typical recombination experiment in which co-conversion of a selected and an unselected site can be monitored is presented in Figure 1. Females, heterozygous for the selective sites, ry^{i1005L} and ry^8, and the unselected electrophoretic mobility site alternatives, *e1004F* and *e1004S* are presented within brackets designating them as sites within the *rosy* locus. They are also heterozygous for the flanking markers, *kar* and Ace^{126}. The locations of the *rosy* locus sites on the molecular map are indicated. Thus, ry^{i1005L} is an extreme underproducer site variant, spontaneous in origin (MCCARRON *et al.* 1979), located at −1701 just 5′ of the transcription start site (CURTIS *et al.* 1989). The X-ray induced mutation, ry^8, is a 17-bp frameshift deletion (GRAY *et al.* 1991) located in exon 2. The location of the spontaneous electrophoretic polymorphism, ry^{e1004} (MCCARRON *et al.* 1979) was identified from a comparison of various wild-type and intragenic recombinant sequences (CURTIS *et al.* 1989).

Females were mated and progeny reared on selective medium permitting the survival of offspring receiving a ry^+ bearing recombinant chromosome. The tester males have third chromosomes with *rosy* mutant sites, identifiable flanking markers and multiple break rearrangements that serve to prevent subsequent recombination in the surviving progeny, each of whom receive one or another of the paternal chromosomes. In this experiment (Figure 1), 27 ry^+ recombinant survivors were recovered, scattered at random among the replicate crosses that sampled an estimated 1.2×10^6 total progeny. Total progeny sample was estimated from a count of total offspring in a portion of the replicate cultures reared in the absence of selective medium.

The surviving exceptional progeny were mated individually and subjected to an array of tests designed to characterize each of the maternally derived recombinant chromosomes (see HILLIKER and CHOVNICK 1981). Figure 1 presents the results of such tests by summarizing the genetic composition of the recombinant chromosomes. Seven of the ry^+ chromosomes were recombinant for the flanking markers, and clearly represent crossovers at sites between the selective markers ry^{i1005L} and ry^8. Another group of 10 carried all parental markers of the ry^{i1005L} bearing chromosome and are classified as conversions of $ry^{i1005L} \rightarrow \pm$. A third group of 10

TABLE 1

Location of selected and nonselected sites within the *rosy* locus employed in this analysis

rosy allele	Molecular description
5	19-bp deletion from +294 to +312; null allele
8	17-bp deletion from +1283 to +1299; null allele
26	GG → T at +2804–5; null allele
41	Deletion of gly codon at +3095 to +3097; null allele
e111	Electrophoretic variant at +3557
201	TT insert at +737; null allele
204	GCC → GC at +685; null allele
e217	Electrophoretic variant at +736
406	G → A at +451; complementing null allele
e408	Electrophoretic variant at +3557
502	3-bp deletion from +683 to +685; null allele
e507	Electrophoretic variant at +736
e508	Electrophoretic variant at +3557
606	G → A at −468; complementing null allele
e1004	Electrophoretic variant at +1551
i1005	T → C at −1701; hypomorphic allele

carried outside flanking markers of the ry^8 bearing parental chromosome, and hence their classification as conversions, $ry^8 \rightarrow ry^+$. However, six of these ry^8 conversions carried the *e1004F* marker like the ry^8 parental chromosome, while four of the ten ry^8 conversions were *e1004S* indicating that they are co-conversions for the electrophoretic site located 252 bp 3′ to the ry^8 site. We infer that 40% of the conversions of $ry^8 \rightarrow ry^+$ included a DNA segment that extended downstream to include the *e1004* site.

The DNA characterizations and locations of selective sites as well as unselected sites utilized in the present report are summarized in Table 1 (LINDSLEY and ZIMM 1992).

RESULTS

Models for conversion tract length distribution: We first derived a simple model for conversion tract length distributions using proportionality arguments. Say that as a conversion tract is initiated and as the tract elongates, for each new nucleotide there is a probability (ϕ) that the conversion tract will continue and a probability ($1 - \phi$) that it will terminate. Therefore, the overall probability, P, that conversion tract length, L, will be a specific number of nucleotides (n), is $P(L = n) = (1 - \phi)\phi^n$ when there is no selective bias for larger conversion tracts. This is simply a geometric distribution of conversion tract lengths. The most frequent single class of tract lengths would be $L = 1$ nucleotide, although for high values of ϕ it would represent only a minute fraction of all conversion tracts. The mean value of n with respect to this distribution is $\phi/(1 - \phi)$. The probability that conversion tract length is equal to or greater than a specified number of nucleotides can be shown to be $P(L \geq n) = \phi^n$.

For conversions of selected sites in fine structure mapping experiments we should see a bias for large conversions. Conversion tracts are randomly initiated and terminated in Drosophila (CLARK *et al.* 1988; CURTIS *et al.* 1989; CURTIS and BENDER 1991). Hence for specific mutant sites assayed for conversion for ry to ry^+ the probability that a given conversion tract will include the site is proportional to its length. Thus, the probability that a conversion tract will include a selected site and therefore be *observable* can be shown to be $P(L = n \mid \text{selected}) = n\phi^{(n-1)}(1 - \phi)^2$. The mean value of n with respect to this distribution is $(1 + \phi)/(1 - \phi)$ (see APPENDIX). For large ϕ, this mean is approximately double that of the distribution associated with unselected conversions. It can also be shown that $P(L \geq n \mid \text{selected}) = \phi^n(1 - n + n/\phi)$. (It should be noted that boundaries are not of practical significance to conversion tract length distribution. Although the Drosophila genome is not present as a single circular DNA molecule but as four pairs of chromosomes; it is unlikely that a conversion tract of the length distribution observed in Drosophila would encounter a boundary such as the end of a DNA molecule or, a possible boundary, the heterochromatic-euchromatic junction.)

It should be noted that selection is not simply for inclusion for a particular site, as we have assumed above, but is also against inclusion of another nearby site in an intragenic recombination experiment involving two heterozygous heteroalleles. The conditional probability of a tract that contains the positively selected site but misses the negatively selected one thus depends on the distance between the sites. The derivation of these conditional probabilities is shown in the APPENDIX contributed by W. R. ENGELS. As the distance between positively and negatively selected sites becomes large, these probabilities become the ones described above.

Co-conversion data analysis: Using the model in which the conversion tract lengths follow a geometric distribution, the only variable is ϕ. One can estimate ϕ and also test the utility of the model by employing certain co-conversion data obtained in fine structure mapping experiments of the sort illustrated in Figure 1. In these experiments ry^+ conversions of specific *rosy* mutant alleles of known molecular location were examined for co-conversion of nonselected electrophoretic mobility site polymorphisms of known molecular location in the *rosy* locus. (In these experiments all ry^+ conversions were examined for co-conversion of the electrophoretic site.) If we plot the frequency of co-conversion of selected and nonselected sites it should fit the distribution defined by the equation $P(L \geq n) = \phi^n$ if this simple model is correct.

Table 2 summarizes a series of intragenic mapping experiments in which we were able to determine co-conversion frequencies of specific ry alleles and nonselected electrophoretic sites, all of known molecular location (Table 1). The frequency of co-conversion and the physical distance between the co-converted sites was then used to estimate ϕ.

TABLE 2

Co-conversion frequencies of selected and nonselected sites of known molecular location in the *rosy* locus

No. of base pairs	*rosy* allele	Electrophoretic site(s)	Co-conversions/ total conversions	Frequency of co-conversion
1	*201*	e217	3/3	1.00
51	*204*	e217	7/8	0.88
51	*502*	e507	58/71	0.82
252	*8*	e1004	4/10	0.40
285	*406*	e217	2/3	0.67
424	*5*	e217/e507	7/18	0.39
460	*41*	e111, e408, e508	22/80	0.28
547	*8*	e217, e507	5/22	0.23
752	*26*	e111, e508	1/19	0.05
1204	*606*	e507	1/19	0.05
3106	*406*	e408	0/6	0.00
3252	*i1005L*	e1004	0/47	0.00

There were 306 conversions in 44.28×10^6 progeny (44 million).

First of all, a computer program was written to find by numerical iteration the value of ϕ that best fit the observed data to $P(L \geq n) = \phi^n$. This value was found to be $\phi = 0.99736$, which would give a mean conversion tract length of 378 bp. In the APPENDIX (contributed by W. R. ENGELS), a maximum likelihood estimation that took into account the numbers of co-conversions and simple conversions of the selected sites, and not simply their ratio, yielded $\phi = 0.99717$, with a standard error of 0.00026. Figure 2 illustrates the fit of the co-conversion data to ϕ^n for this second value of ϕ. For the optimal ϕ value obtained by maximum likelihood, the mean conversion tract length is 352 bp. These results support the contention that the meiotic gene conversion tract length distribution in Drosophila can be approximated by the function $P(L \geq n) = \phi^n$.

Selected site conversion tract length distribution: The underlying or unselected conversion tract length frequency distribution is quite different from the distribution associated with *selected* site conversions, *i.e.*, conversions of mutant sites in intragenic mapping experiments. For $\phi = 0.99717$ the mean selected conversion tract length is approximately doubled to 706 bp, when the distance between the positions of positively and negatively selected sites is maximal (see APPENDIX). This estimate is in reasonable agreement with that of CURTIS and BENDER (1991) for mean selected conversion tract lengths (885 bp), based on analysis of 27 *ry* locus (non-crossover) conversions from crosses with multiple heterozygosities for DNA sequence polymorphisms within the *ry* locus (see also CURTIS *et al.* 1989).

Crossover-associated conversions: The parallel between meiotic gene conversion and crossing over [reviewed in HILLIKER and CHOVNICK (1981)] led us to postulate that all meiotic recombination in Drosophila has its origin in gene conversion with a fraction of gene conversions being resolved as crossovers (*i.e.*, physical exchanges). Indeed, intragenic mapping studies involving the use of half-tetrads have revealed that crossovers are often associated with gene conversion events (SMITH *et al.* 1970; CLARK *et al.* 1984; CURTIS *et al.* 1989). Although, CURTIS *et al.* (1989) obtained evidence that crossover-associated conversions are on average smaller than those not associated with crossovers, this is not due to a true size difference. First, as discussed above and also recognized by CURTIS *et al.* (1989), there is selection for non-crossover conversions to be large since larger conversions are more likely to convert a marker site to ry^+ than are smaller conversions. (However, for conversions occurring *between* selected markers which also result in crossing over and in the production of a ry^+ chromatid, there is no bias for larger conversions.) Second, CURTIS *et al.* (1989) inferred that one-half of crossovers would *not* be associated with an observable gene conversion, even if all crossovers have their origin in a gene conversion event. In one general class of molecular models, gene conversion involves the formation of a heteroduplex and the production of a single DNA strand recipient of a nonreciprocal transfer of information. When the heteroduplex dissociates, the "converted" (recipient) DNA single strand then base pairs with its original complementary single strand. They reasoned that the probability that the resultant mismatches are corrected to the recipient DNA (and thus donor) strand form is 50% and, thus, that 50% of gene conversion events should result in no net conversion. However, such "null" convertants are recoverable as ry^+ crossovers if the original conversion event occurred between the *rosy* heteroalleles and produced a crossover.

From the simple model and analysis presented in this report we would expect crossover and non-crossover-associated conversions obtained as ry^+ exceptionals in fine structure analyses to be associated with different conversion tract length distributions. Crossover-associated conversions should follow approximately the distribution $P(L \geq n) = \phi^n$ as among such conversions there is no selective bias against small conversion tracts. Non-crossover conversions should follow approximately the distribution $P(L \geq n \mid \text{selected}) = \phi^n(1 - n + n/\phi)$

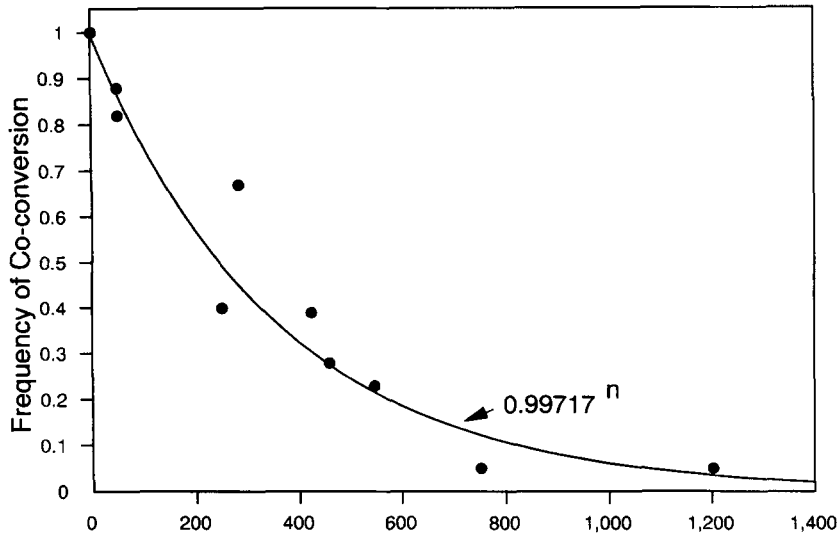

FIGURE 2.—Co-conversion frequencies–observed data (Table 2) and $P(L \geq n) = \phi^n$ using a maximum likelihood estimate with standard error for ϕ of 0.99717 ± 0.00026. The ordinate represents the proportion of co-conversions observed for sites a fixed number of nucleotides apart, plotted by black spheres, or, for the curvilinear lines, the probability $P(L \geq n)$ of co-conversion for two sites as a function of the number of nucleotides apart. The abscissa defines "n," the number of nucleotides by which two sites are separated.

and thus on average be larger than the crossover-associated conversions (see also APPENDIX).

CURTIS *et al.* (1989) detected four *rosy* locus crossovers associated with gene conversions. Their estimate of the mean conversion tract length of these recombinants was 343 bp (with a minimum estimate of 156 bp). From our model, these conversions should be representative of the underlying (*i.e.*, unselected) conversion tract length distribution defined by the equation $P(L \geq n) = \phi^n$. The estimate of the mean of this distribution from the analysis of co-conversion data was 352 bp in good agreement with the CURTIS *et al.* (1989) estimate for crossover-associated conversions which was based on a relatively small sample.

DISCUSSION

We have described a geometric distribution of meiotic gene conversion tract lengths within the *rosy* locus of *D. melanogaster*. This distribution has an excellent fit with co-conversion data derived from fine structure experiments collectively yielding 306 conversions from over 44 million progeny. We demonstrate that the apparent difference in tract length between crossover and non-crossover-associated conversions (CURTIS *et al.* 1989; CURTIS and BENDER 1991) can be predicted by our simple model of conversion tract length distribution. We postulate that the conversion tract length distribution within the *rosy* locus is representative of the overall conversion tract length distribution throughout the euchromatic portion of the genome.

The underlying (*i.e.*, nonselective) conversion tract length distribution has a mean conversion tract length of 352 bp (for $\phi = 0.99717$). Nevertheless, much shorter tract lengths are common, *e.g.*, 13% of tract lengths would be less than 50 bp in length for the optimal ϕ. Thus, gene conversion can result in extensive shuffling and reshuffling of sequences within a gene over evolutionary time in populations in which there are numerous intragenic polymorphisms except for sites that are very close together. Selected site conversion tracts are biased toward larger lengths. We estimate that the mean conversion tract length within the *rosy* locus for selected conversions is 706 bp.

Conversion tract length is determined by ϕ, the probability of extension of a conversion tract on a nucleotide by nucleotide basis. Even with very high ϕ values (approaching $\phi = 0.999$) conversion tracts in excess of 3000 nucleotides are rare. Small reductions in ϕ have major effects on conversion tract length distribution. Thus, one could conceive of meiotic systems in which recombination as assayed by crossing over appears to occur in the absence of gene conversion. That is, the tract length would be so short that selected sites would be converted at a very low frequency.

Our analysis and those of CURTIS *et al.* (1989) and HILLIKER and CHOVNICK (1981) argue that gene conversion tracts are uninterrupted (continuous) in Drosophila. Nevertheless, one can conceive of situations in which conversion tracts may *appear* to be discontinuous. In Drosophila, as well as several fungi, gene conversions not associated with crossovers do not exhibit chromosomal interference [see HILLIKER and CHOVNICK (1981) and references therein]. In organisms with very high rates of meiotic recombination relative to genome size, such as many fungi, adjacent gene conversion events may result in apparent "patchy" single gene conversions.

The present study of meiotic conversion tract length distribution for the *rosy* locus in Drosophila is quite similar in logic to the analysis of the distribution of gap repair path lengths following *P* element excision (ENGELS *et al.* 1990; GLOOR *et al.* 1991; JOHNSON-SCHLITZ and ENGELS 1993). However, there are clear differences between meiotic gene conversion and mitotic gap repair. (1) The mean tract length of meiotic conversions is markedly less than that of mitotic gap repair tracts (352 *vs.* 1379 bp). (2) The frequency of mitotic gap repair is

highly sensitive to reduction by single-base mismatching within the homologous region (NASSIF and ENGELS 1992), whereas single base mismatches have no effect on meiotic gene conversion and associated intragenic crossovers within the *rosy* locus (HILLIKER *et al.* 1991). (3) Finally, mutants of the *mei-9* locus clearly affect meiotic gene conversion but have no effect on mitotic gap repair (CARPENTER 1982; BANGA *et al.* 1991) and mutants of the *mus(3)302* locus do not affect meiotic recombination but seriously disrupt mitotic gap repair (BANGA *et al.* 1991).

The authors are pleased to acknowledge research support from an operating grant from the Natural Science and Engineering Research Council of Canada to A.J.H. and from a research grant GM-09886 from the U.S. Public Health Service to A.C. We are grateful to W. R. ENGELS for contributing significant refinements to the mathematical analysis which are contained in the APPENDIX.

LITERATURE CITED

BANGA, S. S., A. VELAZQUEZ and J. B. BOYD, 1991 P transposition in Drosophila provides a new tool for analyzing post-replication repair and double-strand break repair. Mutat. Res. **255:** 79–88.

CARPENTER, A. T. C., 1982 Mismatch repair, gene conversion, and crossing over in two recombination-defective mutants of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **79:** 5961–5965.

CARPENTER, A. T. C., 1984 Meiotic roles of crossing-over and of gene conversion. Cold Spring Harbor Symp. Quant. Biol. **49:** 23–29.

CHOVNICK, A., G. H. BALLANTYNE, D. L. BAILLIE and D. G. HOLM, 1970. Gene conversion in higher organisms: half-tetrad analysis of recombination within the rosy locus of *Drosophila melanogaster*. Genetics **66:** 315–329.

CHOVNICK, A., G. H. BALLANTYNE and D. G. HOLM, 1971 Studies on gene conversion and its relationship to linked exchange in *Drosophila melanogaster*. Genetics **69:** 179–209.

CLARK, S. H. C., S. DANIELS, C. A. RUSHLOW, A. J. HILLIKER and A. CHOVNICK, 1984 Tissue-specific and pretranslational character of variants of the *rosy* locus control element in *Drosophila melanogaster*. Genetics **108:** 953–968.

CLARK, S. H., A. J. HILLIKER and A. CHOVNICK, 1988 Recombination can initiate and terminate at a large number of sites within the *rosy* locus of *Drosophila melanogaster*. Genetics **118:** 261–266.

CURTIS, D. and W. BENDER, 1991 Gene conversion in Drosophila and the effects of the meiotic mutants *mei-9* and *mei-218*. Genetics **127:** 739–746.

CURTIS, D., S. H. CLARK, A. CHOVNICK and W. BENDER, 1989 Molecular analysis of recombination events in Drosophila. Genetics **122:** 653–661.

DUTTON, F. L., and A. CHOVNICK, 1988 Developmental regulation of the *rosy* locus in *Drosophila melanogaster*, pp. 267–316 in *Developmental Biology*, Vol. 5, edited by L. W. BROWDER. Plenum, New York.

ENGELS, W. R., D. M. JOHNSON-SCHLITZ, W. B. EGGLESTON and J. SVED, 1990 High frequency *P* element loss in *Drosophila* is homologue dependent. Cell **62:** 515–525.

GLOOR, G. B., N. A. NASSIF, D. M. JOHNSON-SCHLITZ, C. R. PRESTON and W. R. ENGELS, 1991 Targeted gene replacement in *Drosophila* via P element-induced gap repair. Science **253:** 1110–1117.

GRAY, M., A. CHARPENTIER, K. WALSH, P. WU and W. BENDER, 1991 Mapping point mutations in the *Drosophila rosy* locus using denaturing gradient gel blots. Genetics **127:** 139–149.

HILLIKER, A. J., and A. CHOVNICK, 1981 Further observations on intragenic recombination in *Drosophila melanogaster*. Genet. Res. **38:** 281–296.

HILLIKER, A. J., S. H. CLARK and A. CHOVNICK, 1988 Genetic analysis of intragenic recombination in *Drosophila*, pp. 73–90 in *The Recombination of Genetic Material*, edited by K. B. LOW. Academic Press, San Diego.

HILLIKER, A. J., S. H. CLARK and A. CHOVNICK, 1991 The effect of DNA sequence polymorphisms on intragenic recombination in the *rosy* locus of *Drosophila melanogaster*. Genetics **129:** 779–781.

JOHNSON-SCHLITZ, D. M., and W. R. ENGELS, 1993 P element-induced interallelic gene conversion of insertions and deletions in *Drosophila*. Mol. Cell. Biol. **13:** 7006–7018.

KEITH, T. P., M. A. RILEY, M. KREITMAN, R. C. LEWONTIN, D. CURTIS *et al.*, 1987 Sequence of the structural gene for xanthine dehydrogenase (*rosy* locus) in *Drosophila melanogaster*. Genetics **116:**67–73.

KENDALL, M. G., and A. STUART, 1973 *The Advanced Theory of Statistics*. Charles Griffin & Co., London.

LEE, C. S., D. CURTIS, M. MCCARRON, C. LOVE, M. GRAY *et al.*, 1987 Mutations affecting expression of the *rosy* locus in *Drosophila melanogaster*. Genetics **116:** 55–66.

LINDSLEY, D. L., and G. ZIMM, 1992 *The Genome of Drosophila melanogaster*. Academic Press, New York.

MCCARRON, M., J. O'DONNELL, A. CHOVNICK, B. S. BHULLAR, J. HEWITT *et al.* 1979 Organization of the *rosy* locus of *Drosophila melanogaster*: further evidence in support of a *cis*-acting control element adjacent to the xanthine dehydrogenase control element. Genetics **91:** 275–293.

NASSIF, N., and W. ENGELS, 1993 DNA homology requirements for mitotic gap repair in Drosophila. Proc. Natl. Acad. Sci. USA **90:** 1262–1266.

SMITH, P. D., V. G. FINNERTY and A. CHOVNICK, 1970 Gene conversion in *Drosophila*: non-reciprocal events at the maroon-like locus. Nature **228:** 441–444.

WEIR, B. S., 1990 *Genetic Data Analysis*. Sinauer Associates, Sunderland, Mass.

WOLFRAM RESEARCH, 1993 *Mathematica*. Wolfram Research, Inc., Champaign, Ill.

Communicating editor: C. C. LAURIE

APPENDIX

Analysis of Conversion Tract Data with a Geometric Tract Length Distribution

William R. Engels

Genetics Department, University of Wisconsin-Madison, Madison, Wisconsin 53706

Estimation of φ by maximum likelihood: I will assume that the numbers of co-conversions and conversions in Table 2 come from a binomial distribution whose parameter is φ^k, where k is the number of base pairs between the selected and nonselected sites. As will be demonstrated below, the geometric distribution of tract lengths leads to φ^k being the binomial parameter even when the conversion tracts are subject to selection.

The values of k for the 12 cases in Table 2 are: 1, 51, 53, 268, 285, 442, 462, 563, 753, 1204, 3106 and 3252. Let c_i be the number of co-conversions of the selected site and the electrophoretic site in the ith experiment, and d_i be the number of simple conversions of the selected

site only. The likelihood of the entire data set is given by the product of binomial probabilities:

$$L(\varphi) = \prod_{i=1}^{12} \binom{c_i + d_i}{c_i} \varphi^{k_i c_i}(1 - \varphi^{k_i})^{d_i}.$$

It is more convenient to work with the log of the likelihood,

$$\ln[L(\varphi)] = [\text{constant}] + \sum_{i=1}^{12} k_i c_i \ln(\varphi) + \sum_{i=1}^{12} d_i \ln(1 - \varphi^{k_i}). \quad \text{(A1)}$$

Setting the derivative of this expression equal to zero and solving numerically for φ yields the estimate ($\pm$SE):

$$\hat{\varphi} = 0.99717 \pm 0.00026 \quad \text{(A2)}$$

The standard error is the square root of the variance estimate, obtained from

$$V(\hat{\varphi}) \approx \frac{-1}{(d^2/d\varphi^2)\ln[L(\varphi)]|_{\varphi=\hat{\varphi}}},$$

where

$$\frac{d^2}{d\varphi^2}\ln[L(\varphi)] = \frac{\sum_{i=1}^{12} k_i c_i}{\varphi^2} + \sum_{i=1}^{12} \frac{d_i k_i \varphi^{k_i-2}(1 - k_i - \varphi^{k_i})}{(1 - \varphi^{k_i})^2}.$$

Reviews of the method of maximum likelihood are available (KENDALL and STUART 1973; WEIR 1990). Computations were performed with the computer program *Mathematica* (Wolfram Research, 1993).

Conditional distribution of conversion tracts under selection: The scheme in Figure 1 allows recovery of conversion tracts that restore one mutant site to the wild-type sequence, but do not extend far enough to convert the other site, which is wild type, to the mutant homolog. In the following analysis I will derive expressions for the probability distribution of tract length under this kind of selection, as well as its cumulative distribution and expectation.

Simple selection for inclusion of a given point can be obtained from these results as a special case by letting the distance between the positively and negatively selected sites approach infinity.

Let m be the distance in base pairs between the positively and negatively selected sites, and use the symbol to $a\bar{b}$ indicate the condition for restoration of ry^+ expression, *i.e.*, inclusion of one site but not the other. We wish to obtain the conditional probability of a conversion tract of length n given $a\bar{b}$. That is:

$$P(n \mid a\bar{b}) = \frac{P(n \cap a\bar{b})}{P(a\bar{b})}. \quad \text{(A3)}$$

Without loss of generality, I shall assume that the positively selected site is located to the right of the negatively selected one, and that each of G nucleotide sites in the genome is equally likely to serve as the left end of a conversion tract. Any conversion tract that is recovered in the screen must have its left end between the two selectable sites, and it must extend through the right site. For a position j base pairs to the left of the positively selected site ($j \leq m$), the probability is $1/G$ that a given conversion tract will lie with left end there, and the probability that the tract will extend at least to the positively selected site is φ^j. Since the placement and the length of the tract are assumed to be determined inde-

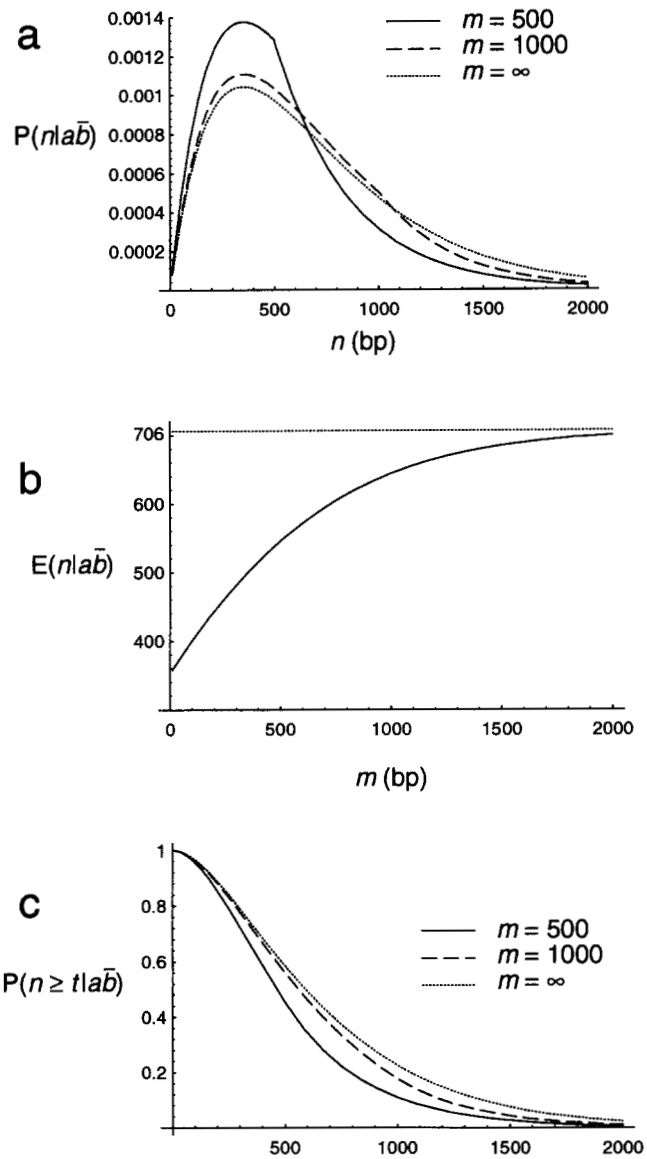

FIGURE 3.—(a) $P(n \mid a\bar{b})$, the probability of a conversion tract of length n given positive/negative selection, as given by Equation A5. The distance between the positively and negatively selected sites is m, and $\hat{\varphi} = 0.99717$. (b) The average conversion tract length as a function of m. The maximum likelihood estimate, $\hat{\varphi} = 0.99717$, from Equation A2 was substituted into Equations A6 and A7. (c) The probability of a conversion tract being as long or longer than a given size, t, from Equation A8 with $\hat{\varphi} = 0.99717$.

pendently, we can multiply these probabilities and compute $P(a\bar{b})$ by summing the product over all positions between the two selectable sites. Thus

$$P(a\bar{b}) = \frac{1}{G}\sum_{j=0}^{m} \varphi^j = \frac{\varphi(1-\varphi^m)}{G(1-\varphi)}. \quad \text{(A4)}$$

If $n < m$, the selection scheme requires that the left endpoint lie within n bp of the positively selected point, which has probability n/G. If $n \geq m$, the left endpoint can be anywhere between the two sites, which has probability m/G. We can again invoke independence between length and position to obtain the unconditional probability, $P(n \cap a\bar{b})$, that a given site has length n and satisfies the selection criteria. Thus,

$$P(n \cap a\bar{b}) = \begin{cases} \frac{n}{G}\varphi^n(1-\varphi) & \text{if } n < m \\ \frac{m}{G}\varphi^n(1-\varphi) & \text{if } n \geq m \end{cases}.$$

Combining this with Equations A3 and A4 and simplifying gives the desired conditional probability:

$$P(n \mid a\bar{b}) = \begin{cases} \frac{n\varphi^{n-1}(1-\varphi)^2}{1-\varphi^m} & n < m \\ \frac{m\varphi^{n-1}(1-\varphi)^2}{1-\varphi^m} & n \leq m \end{cases} \quad \text{(A5)}$$

Note that this probability is independent of the genomic constant, G. Figure 3a shows this probability plotted with the maximum likelihood estimate, $\hat{\varphi}$, from Equation A2. When m is large, it approaches $P(n \mid a) = n\varphi^{n-1}(1-\varphi)^2$, the probability conditioned on simple selection for one site.

The average tract length under positive/negative selection is obtained by summing $nP(n \mid a\bar{b})$ over all allowable values of n to yield:

$$E(n \mid a\bar{b}) = \frac{1+\varphi+\varphi^m(\varphi m - \varphi - m - 1)}{(1-\varphi)(1-\varphi^m)}. \quad \text{(A6)}$$

See Figure 3b for a plot of this expectation using the maximum likelihood value of $\hat{\varphi}$. For the case of simple selection for one site, we allow m to approach infinity, yielding:

$$E(n \mid a\bar{b}) = \frac{1+\varphi}{1-\varphi}. \quad \text{(A7)}$$

With the maximum likelihood estimate, $\hat{\varphi} = 0.99717$ obtained above, the average tract length from A7 is 705.7 bp.

To calculate the conditional probability of a tract being least as large as t base pairs, we sum the above expression from t to infinity. The result is:

$$P(n \geq t \mid a\bar{b}) \quad \text{(A8)}$$

$$= \begin{cases} \frac{\varphi^{m-1}(m-\varphi-2m\varphi+m\varphi^2)+\varphi^{t-1}(t+\varphi-t\varphi)}{1-\varphi^m} & t < m \\ \frac{\varphi^{t-1}(m-m\varphi)}{1-\varphi^m} & t \geq m \end{cases}$$

A plot of this probability is shown in Figure 3c for $\hat{\varphi}$, the maximum likelihood estimate. It shows that the presence of a negatively selected site has a substantial effect when m is small, but becomes negligible when m is sufficiently large. As m goes to infinity, *i.e.*, the case of simple selection for one site, the probability in Equation A8 approaches the limit $\varphi^{t-1}(t+\varphi-t\varphi)$.

Now consider the probability of inclusion of a non selected site such as the electrophoretic marker *e1004* in Figure 1, conditioned on the positive/negative selection at the other two selectable sites. Using g to symbolize the inclusion of the non selected site separated from the nearest selected site by k base pairs, we can use the same arguments as above to obtain:

$$P(g \mid a\bar{b}) = \frac{P(g \cap a\bar{b})}{P(a\bar{b})} = \frac{\frac{1}{G}\sum_{i=1}^{m}\varphi^{k+i}}{P(a\bar{b})}.$$

Using Equation A4 to substitute for $P(a\bar{b})$ and simplifying leads to

$$P(g \mid a\bar{b}) = \varphi^k,$$

thus justifying our use of φ^k as the binomial parameter in the maximum likelihood procedure described above.