The Estimation of the Number and the Length Distribution of Gene Conversion Tracts From Population DNA Sequence Data

Esther Betrán,* Julio Rozas,† Arcadio Navarro* and Antonio Barbadilla*

*Departament de Genètica i de Microbiologia, Facultat de Ciències, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain and †Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain

Manuscript received January 30, 1996
Accepted for publication January 24, 1997

ABSTRACT

DNA sequence variation studies report the transfer of small segments of DNA among different sequences caused by gene conversion events. Here, we provide an algorithm to detect gene conversion tracts and a statistical model to estimate the number and the length distribution of conversion tracts for population DNA sequence data. Two length distributions are defined in the model: (1) that of the observed tract lengths and (2) that of the true tract lengths. If the latter follows a geometric distribution, the relationship between both distributions depends on two basic parameters: $\psi$, which measures the probability of detecting a converted site, and $\phi$, the parameter of the geometric distribution, from which the average true tract length, $1/(1-\phi)$, can be estimated. Expressions are provided for estimating $\phi$ by the method of the moments and that of the maximum likelihood. The robustness of the model is examined by computer simulation. The present methods have been applied to the published $rp+9$ sequences of Drosophila subobscura. Maximum likelihood estimate of $\phi$ for this data set is 0.9918, which represents an average conversion tract length of 122 bp. Only a small percentage of extant conversion events is detected.

Studies of intra- and interspecific DNA sequence variation are providing extremely useful information to infer the mechanisms generating genetic variation at the population level. Gene conversion, the non-reciprocal transfer of information between homologous nonsister chromatids in individual meiotic tetrads, has been proposed to account for the observation of "mosaic sequences," the shift of small continuous segments of DNA among different sequences or haplotypes (Riley et al. 1988; Hammer et al. 1991; Hughes et al. 1993; Rozas and Aguade 1994; Popadić and Anderson 1995). Although conversion events are often intuitively inferred from mosaic sequences, there are, however, rigorous statistical methods for their detection. Stephens' method (1985) is based on the clustering pattern of variable sites among transferring haplotypes. This method is especially appropriate for small sample sizes, on the order of two or three sequences. Sawyer (1989) introduced a powerful and more general statistical test for detecting conversion, based on the number of segregating sites distinguishing any two regions of DNA. Both the Stephens and Sawyer approaches simply infer that gene conversion has occurred among the sampled sequences. They ignore the pattern, the number, and the true lengths of gene conversion events. Here, we provide for a sample of homologous DNA sequences: (1) a new method for detecting gene conversion tracts and (2) a statistical model to estimate the true length distribution of gene conversion tracts and the true number (observable plus hidden) of conversion events from the tracts detected by our method.

In their study of Drosophila melanogaster rosy locus, Hillyer et al. (1994) recovered extensive co-conversion events for selected and unselected heterozygous sites of known molecular localization by crossing strains with very close markers. Gene conversion tract lengths inferred from these and previous co-conversion data fitted a geometric distribution well (Engels 1994), from which the maximum likelihood estimation for $\phi$, the parameter of the distribution, and the mean conversion tract length (352 bp) were estimated. DNA sequence data from population surveys can also be used to estimate the parameter $\phi$ of the conversion tract length distribution. However, statistical inference from DNA population data is more complicated than in the rosy locus, because sequence data from sampled populations lack a priori differentiated (heterozygous) positions, and a new parameter giving the probability that a converted site is detected must be introduced.

We can clearly detect mosaic sequences, and therefore infer conversion events, when gene conversion occurs between differentiated or divergent haplotypes. We will refer to these subsets of differentiated sequences as subpopulations, which obviously do not imply geographically separated populations. Such divergent subpopulations can be generated when crossing over is partial or completely suppressed between them, as in inversion systems of Drosophila (Rozas and Aguade...
Figure 1.—Nucleotide polymorphism in 34 analyzed sequences at the \( rph49 \) locus within two polymorphic arrangements in the \( O \) chromosome (\( O_1 \) and \( O_{2+} \)) in \( D. subobscura \). Functional domains of the region are indicated at the top. The 1528 nucleotides studied are distributed as follows: 841 in the 5' flanking region, 93 in exon 1, 62 in the intron, 309 in exon 2, and 213 in the 3' flanking region. The first and the last columns indicate the strain and the arrangement of the \( O \) chromosome of the strain, respectively. Nucleotides identical to the first sequence are indicated by a dot. Blocks of nucleotides converted from one gene arrangement to the other are depicted as a shaded box. (From Rozas and Aguade 1994.)

Consider, for example, the sequence data of Rozas and Aguade (1994). Thirty-four sequences for the \( rph49 \) locus were analyzed for two polymorphic chromosomal gene arrangements (\( O_2 \) and \( O_{2+} \)) in \( D. subobscura \) (Figure 1). Because (1) the \( rph49 \) gene is very close to a chromosomal inversion breakpoint where crossing over is mostly suppressed, and (2) the \( rph49 \) sequences from the two chromosome arrangements are partially differentiated, the authors were able to identify five gene conversion tracts from the observation of the mosaic sequences. This data set will be used to illustrate the statistical model described here.

The value of the model and estimators provided is twofold. Current estimates of conversion tract lengths are limited to a few experimental systems. In \( D. subobscura \), for example, estimates are almost exclusively restricted to the \( rosy \) locus. New sequence data are providing information of many regions of the genome, and a more complete picture of the tract length distribution along the genome can be obtained. From a population genetics point of view, knowledge of the number and
length of conversion tracts is crucial to determine the importance of gene conversion in generating haplo-
typic diversity (Ishi and Charlesworth 1977, Leslie and Watt 1986; Berry and Barbadilla 1997).

**THE MODEL**

The reasoning underlying the model is as follows: consider two sets of differentiated (i.e., exhibiting linkage disequilibrium) sequences (e.g., sequences derived from two different polymorphic inversions). A fraction of the tracts exchanged between each set, or subpopula-
tion, will produce detectable mosaic haplotypes. We will introduce first a method for the detection of these tracts that considers that two or more subpopulation-specific variants (diagnostic or informative nucleotides) must be transferred to distinguish a conversion event from muta-
tion. These detected or observed tracts will constitute the empirical basis of the whole model. We assume that each observed tract is produced by a single conversion event, i.e., each tract has not subsequently been broken by additional conversion or crossing-over events. This assumption holds for DNA regions with a moderate or low density of conversion or crossing-over events. Because informative nucleotides are a subset of all nucleotides, the observed tract length (measured as the distance between the outermost informative nucleo-
tides) will always be equal to, or less than, the true tract length. The relationship between observed and true tract length depends both on the frequency of informative nucleotides and the distance between the outermost informative nucleotides. The probability of detecting a conversion event can be determined if it captures at least two diagnostic nucleo-
tides (Stephens 1985). Informative (or diagnostic) nucleotides are those revealing the mosaic origin of a tract. Let \( \psi \) (psi) be a parameter measuring the probability of detecting a conversion event between any two sub-
populations at a given site (i.e., the probability of a site being informative). For example, a \( \psi \) value of 0.01 means that a given conversion tract has, on average, 1% of informative nucleotides showing the mosaic origin of the tract. Assuming an uniform distribution of \( \psi \) along the sequence, for a true tract of length \( n \), the probability of observing \( s \) informative sites within the tract is binom-
al as follows:

\[
P(S = s | N = n) = \binom{n}{s} \psi^s (1 - \psi)^{n-s}.
\]

(1)

Accordingly, the probability of detecting a conversion tract is

\[
P(S \geq 2 | N = n) = 1 - P(S < 2) = 1 - (1 - \psi)^n - n\psi(1 - \psi)^{n-1}.
\]

(2)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi )</td>
<td>Probability that a converting tract elongates to an additional nucleotide</td>
</tr>
<tr>
<td>( \psi )</td>
<td>Probability of a site being informative of a conversion event</td>
</tr>
<tr>
<td>( N )</td>
<td>Random variable for true tract lengths</td>
</tr>
<tr>
<td>( n )</td>
<td>Length of a true conversion tract, ( P(N = n) = (1 - \phi) \phi^{n-1} )</td>
</tr>
<tr>
<td>( E(N) )</td>
<td>Expected true tract length, ( E(N) = 1/(1 - \phi) )</td>
</tr>
<tr>
<td>( S )</td>
<td>Random variable for observed informative sites in a tract</td>
</tr>
<tr>
<td>( s )</td>
<td>Number of informative sites in a tract</td>
</tr>
<tr>
<td>( S^* )</td>
<td>Random variable for observed informative sites higher than one in a tract</td>
</tr>
<tr>
<td>( L )</td>
<td>Random variable for observed tract lengths</td>
</tr>
<tr>
<td>( l )</td>
<td>Length of an observed tract determined by the outermost informative sites within a conversion tract</td>
</tr>
<tr>
<td>( L^* )</td>
<td>Random variable for observed tract lengths larger than one bp</td>
</tr>
<tr>
<td>( E(L^*) )</td>
<td>Expected length of observed tracts larger than one bp</td>
</tr>
<tr>
<td>( k )</td>
<td>Observed number of tracts</td>
</tr>
<tr>
<td>( E(k_s) )</td>
<td>Expected true number of tracts (observed plus hidden) exchanged between two subpopulations</td>
</tr>
<tr>
<td>( x_{ij} )</td>
<td>Relative frequency of the nucleotide variant ( i ) at site ( j ) in subpopulation ( 1 )</td>
</tr>
<tr>
<td>( p_{ij,a} )</td>
<td>Probability of observing the transferred variant ( i ) at site ( j ) from subpopulation ( 1 ) to ( 2 )</td>
</tr>
<tr>
<td>( m )</td>
<td>Length of the analyzed sequence</td>
</tr>
<tr>
<td>( \Phi )</td>
<td>Flux of conversion events per generation between two subpopulations</td>
</tr>
<tr>
<td>( E(T) )</td>
<td>Expected sum of the lengths of the branches of the sample genealogy (in generations)</td>
</tr>
<tr>
<td>( H )</td>
<td>Inter-subpopulation heterozygosity</td>
</tr>
<tr>
<td>( c )</td>
<td>Conversion rate per the analyzed region per generation</td>
</tr>
<tr>
<td>( \rho_c )</td>
<td>Conversion rate per base pair per generation</td>
</tr>
<tr>
<td>( S )</td>
<td>Number of segregating sites</td>
</tr>
<tr>
<td>( \mu )</td>
<td>Mutation rate per DNA sequence per generation</td>
</tr>
</tbody>
</table>

**TABLE 1**

<table>
<thead>
<tr>
<th>Definitions of symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
</tr>
<tr>
<td>( \phi )</td>
</tr>
<tr>
<td>( \psi )</td>
</tr>
<tr>
<td>( N )</td>
</tr>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>( E(N) )</td>
</tr>
<tr>
<td>( S )</td>
</tr>
<tr>
<td>( s )</td>
</tr>
<tr>
<td>( S^* )</td>
</tr>
<tr>
<td>( L )</td>
</tr>
<tr>
<td>( l )</td>
</tr>
<tr>
<td>( L^* )</td>
</tr>
<tr>
<td>( E(L^*) )</td>
</tr>
<tr>
<td>( k )</td>
</tr>
<tr>
<td>( E(k_s) )</td>
</tr>
<tr>
<td>( x_{ij} )</td>
</tr>
<tr>
<td>( p_{ij,a} )</td>
</tr>
<tr>
<td>( m )</td>
</tr>
<tr>
<td>( \Phi )</td>
</tr>
<tr>
<td>( E(T) )</td>
</tr>
<tr>
<td>( H )</td>
</tr>
<tr>
<td>( c )</td>
</tr>
<tr>
<td>( \rho_c )</td>
</tr>
<tr>
<td>( S )</td>
</tr>
<tr>
<td>( \mu )</td>
</tr>
</tbody>
</table>
Although $\psi$ is a well defined theoretical parameter, its estimation from DNA sequence data is a different issue, because the informative content of a site depends on the nature of sequence variation. Thus, when considering a site and any two subpopulations, there are two contrasting types of variation: (1) a monomorphic site, i.e., the same nucleotide variant in both subpopulations; and (2) a site with different fixed variants in each subpopulation. In the first case, sites are not informative, that is, the probability of detecting a converted site is equal to zero (sites 1 and 2 in Figure 2). In the second one, sites are completely informative, $P = 1$ (sites 3 and 4 in Figure 2). Variation at many sites is, however, intermediate among these two cases (sites 5–11 in Figure 2). In these instances, it is not evident how to infer whether parallel mutation or gene conversion is responsible for the polymorphisms.

We need, therefore, two algorithms: one to estimate $\psi$ and other to detect gene conversion tracts from different sequence data. They are provided in APPENDIX A. The algorithms, which should be viewed as operational, have been implemented in the version 2.0 of DnaSP program (ROZAS and ROZAS 1997). When the algorithms were applied to ROZAS and AGUDE’s data (1994), five different conversion tracts were detected and their observed length, measured as the distance between the outermost informative site within a tract, were as follows: 12, 17, 39, 65 and 122 bp (APPENDIX A). The average observed length of the five tracts was 51 bp. The value of $\psi$ for the whole region (1518 bp), computed from Equation A4 in the APPENDIX, was 0.00847.

### Relationship between observed and true tract length distributions: Given a true conversion tract of length $n$, the probability of observing any tract of length $L = l$ will depend on $\psi$ as shown in Figure 3. The expression for this conditional probability, when $L > 1$, is

$$P(L = l | N = n) = \frac{(n - l + 1)\psi^2(1 - \psi)^{n-l}}{1 - (1 - \psi)^n - n(1 - \psi)^{n-1}\psi}.$$  

Note that the probability of observing a tract of length $l$ is the probability of $n - l$ noninformative sites flanking out the two informative sites. The probability values for $L = 0$ (or $S = 0$), and $L = 1$ (or $S = 1$), are $(1 - \psi)^n$ and $n(1 - \psi)^{n-1}\psi$, respectively. As “tracts” of length 0 are not observable, and those of length 1 cannot be distinguished from parallel mutation events, we will assume that only tracts larger than 1 bp (i.e., with two or more informative sites) can be observed. The range of observed values of $L$ must therefore be redefined for values $\geq 1$, obtaining the probability function of the new variable $L^*$, which has been normalized by the factor $1/P(L \geq 2)$.

$$P(L^* = l | N = n) = \frac{(n - l + 1)\psi^2(1 - \psi)^{n-l}}{1 - (1 - \psi)^n - n(1 - \psi)^{n-1}\psi}.$$  

The expected value of the observed tract length, $E(L^*)$, for a given true tract length $n$ is

$$E(L^* | N = n) = \sum_{l=2}^n lP(L^* = l | N = n)$$

$$= \frac{[\psi(2\psi - 1) - 2(\psi^2 - 2\psi + 1)](1 - \psi)^n}{\psi[(n\psi - \psi + 1)(1 - \psi)^n + \psi - 1]}.$$  

Figure 4a shows the relationship between expected length of the observed tracts (ordinate), and the true tract lengths (abscissa) for different values of $\psi$. The average observed tract is an increasing monotone function of the true tract length. In addition, however, the expected length of the observed tracts is also positively correlated with the $\psi$ value. If $\psi \approx 0.05$, the observed lengths are much lower than the true ones. In this case, the slope relating both variables is low, meaning that slight variation in the observed values can result in large differences in the inferred true tract lengths.

The unconditional probability function of observing a tract of length $l$ must introduce the distribution of the random variable of true tract lengths ($N$), a geometric distribution with parameter $\phi$:

$$P(L^* = l) = \sum_{n=1}^\infty P(L^* = l | N = n)P(N = n)$$

$$= \sum_{n=1}^\infty P(L^* = l | N = n)\phi^{n-2}(1 - \phi).$$  

Note that the sum starts at $n \geq 2$ because of the restriction that observed tracts are equal to or longer than...
2 bp, so \(P(N = n)\) has been normalized by the factor \(1/\phi\), so that the sum over its range equals unity.

The expected value of the observed tract lengths is as follows:

\[
E(L^*) = E[E(L^* \mid N = n)] = \sum_{n=2}^{\infty} E(L^* \mid N = n) \phi^{n-2}(1 - \phi). \tag{7}
\]

Figure 4b shows a graphical representation of expression (7) for different \(\psi\) and \(\phi\) values. For most of the range of both \(\psi\) and \(\phi\), the expected observed values are considerably lower than the expected true values.

**Estimation of \(\phi\):** The basic parameter describing the true distribution of tracts is \(\phi\). Expression (7) allows the estimation of \(\phi\) by the method of moments. If \(\psi\) and the average observed tract length \(E[L^*]\) are estimated from sequence data, then \(\phi\) can be estimated by solving Equation 7.

The maximum likelihood (ML) estimator has better statistical properties than those of the method of moments and is typically preferable (see Kendall and Stuart 1973; Weir 1990 for details). Given \(k\) observed conversion events of lengths \(l_1, l_2, \ldots, l_k\), the likelihood of an observed data set is given by

\[
L(\phi) = \prod_{i=1}^{k} P(L^* = l_i). \tag{8}
\]

The \(\phi\) value that maximizes \(L(\phi)\) can be found setting its derivative (or alternatively that of \(\ln [L(\phi)]\)) equal to zero and solving for \(\phi\). No explicit expression can be derived from (8), and \(\phi\) must be estimated by iterative or graphic methods. The asymptotic variance of the
estimate can be obtained from the second derivative of \( \ln [L(\phi)] \) (see \textit{Weir 1990}).

In \textit{Rozas and Aguade's} example, the average observed length of the five detected tracts, \( E(L^*) \), is 51 bp. For the estimation of \( \phi \) by the method of moments, \( E(L^*) \) and \( \psi \) must be substituted in expression (7) by their respective estimates, and \( \phi \) can be solved by successive numerical iterations. The estimated value was \( \hat{\phi} = 0.9909 \), which corresponds to an expected true length associated with these values the probabilities of unobserved events are derived from Equations 2 and 3.

According to expression (6), the probability of observing a length \( l \) is a function of both \( \psi \) and \( \phi \) parameters. The number of informative sites associated to an observed length has not been considered as an additional variable. In APPENDIX B is shown that this variable does not provide in fact new information for the estimation of parameter \( \phi \).

**Undetected gene conversion events:** Consider now the related yet independent problem of nondetecting a given conversion event in a sample of sequences. The probability of a conversion event being undetected and that of the expected true length associated with these unobserved events are derived from Equations 2 and the geometric distribution, as follows:

\[
P(S \text{ or } L = 0 \text{ or } 1) = \sum_{n=1}^{\infty} nP(L < 2)\phi^{n-1}(1 - \phi)
\]

\[
= \frac{(1 - \phi)(1 - \phi + 2\phi\psi - \phi\psi^2)}{(1 - \phi + \phi\psi)^2}
\]

(9)

\[
E(N \mid L = 0 \text{ or } 1) = \sum_{n=1}^{\infty} nP(L < 2)\phi^{n-1}(1 - \phi)
\]

\[
= \frac{(1 - \phi)(1 - \phi + 3\phi\psi - 2\phi\psi^2)}{(1 - \phi + \phi\psi)^3}
\]

(10)

Figure 5, a and b, plot expressions (9) and (10) for different values of \( \psi \) and \( \phi \). It is remarkable that for a wide range of \( \psi \) and \( \phi \) values the probabilities of undetected conversion events are large. This means that for

![Figure 4](image-url)
where

If each conversion event can be seen as point mutation events, and

\( k_T \) events exchanged), by the expression:

\[
\text{populations in the sampled sequences,}
\]

detected conversion tracts

If \( \psi = 0.05 \) and/or between DNA regions with large expected tract lengths (large \( \phi \) values, \( \phi > 0.999 \)). Let us assume that there exist 20 exchanged conversion tracts in the sampled sequences. If \( \psi = 0.995 \) (an expected true tract of 200 bp), then we will observe 16.6 tracts (83%) when \( \psi = 0.05 \), but only an average of 0.5 tracts (3%) for \( \psi = 0.001 \). On the other hand, large tracts can pass undetected for small values of \( \psi \) (Figure 5b). In rp49 data set, \( \psi \) and \( \phi \) are 0.00847 and 0.9918, respectively. \( P(L < 2) = 0.742 \), that is, 74% of conversion tracts occurred among both arrangements were undetected. The expected true length of undetected tracts estimated for this data set is 59.62.

**Gene conversion rate:** If \( \psi \), \( \phi \), and the number of detected conversion tracts (\( k \)) have been estimated, we can also estimate the expected true number (observed plus hidden) of conversion events between two any subpopulations in the sampled sequences, \( E(k_T) \), where \( k_T \) is the random variable true number of conversion events exchanged), by the expression: \( E(k_T) = k / P(L \geq 2) \). The variance of \( k_T \), assuming a binomial distribution for \( k \), is

\[
\text{Var}(k_T) = \frac{\text{Var}(k) - E(k_T) P(L \geq 2) [1 - P(L \geq 2)]}{P(L \geq 2)^2}.
\]

If each conversion event is assumed to be unique, then conversions can be seen as point mutation events, and \( k_T \) can be expressed as

\[
E(k_T) = \Phi E(T) = He E(T),
\]

where \( \Phi = He \) and \( E(T) \) are, respectively, the flux of conversion events per generation between different subpopulations for the analyzed region, and the expected sum of the lengths of the branches of the genealogy of the sample in generations (HUDSON 1990). \( H \) is the inter-subpopulation heterozygosity, that is, the probability that the two alleles of an individual comes from any two subpopulations. \( c \) is the rate of gene conversion events per generation for the analyzed DNA sequence. The number of exchanged conversion events is assumed to be Poisson distributed through generations with parameter \( \Phi = He \). If \( \Phi \) varies slightly among generations (because \( H \), \( c \), or both vary), \( He \) can be substituted by \( \overline{H} \tau \), the average heterozygosity and average conversion event rate per generation on the course of the history of the sample. Expression (11) can be used to estimate the parameter \( c \). Let \( c_{bp} \) be the probability of a site being transferred per generation (the conversion rate per base pair per generation). \( c \) and \( c_{bp} \) are related according to the expression \( c_{bp} = cE(N) / m \). Note that while \( c \) refers to the rate of conversion events regardless of their tract lengths, \( c_{bp} \) considers both the rate and the average length of conversion tracts, being therefore, the parameter measuring the potential recombination effect of gene conversion. So, \( H_{ds}c_{bp} \) is the per base pair proportion of new recombinant sites (converted) per generation (where \( ds \) is the average number of nucleotide substitutions per site between subpopulations, Nei 1987, equation 10.20).

Taking the ML estimate of \( \phi \), the expected true number of conversion events in ROZAS and AGUADE's data is \( E(k_T) = \{ 5 / [P(L \geq 2) = 0.258] \} = 19.4 \). So, the probability of a transferred site in the 34 analyzed sequences is \( [19.4 \times 121.95 / (34 \times 1518)] = 0.046 \); that is, 4.6% of nucleotide sites have been transferred by conversion between rearrangements, although only a \( ds \) fraction of them will produce recombining sites. Assuming that this region is neutral evolving, \( E(T) \), the sum of the lengths of the branches of the genealogy of

![Figure 5](image-url)

**(a)** Probability of an undetected conversion event as a function of \( \phi \) and \( \psi \). **(b)** Expected true length associated with unobserved events.
the sample, is equal to $2.78 \times 10^7$ generations (estimated from $S_0/\mu$, where $S_0$ is the total number of segregating sites and $\mu$ is the mutation rate per DNA sequence per generation, from ROZAS and AGUADÉ's data). Substituting this value in expression (11) we obtain a flux estimate of $\Phi = He = 6.98 \times 10^{-7}$. If we suppose that the current value of heterozygosity $H = 0.2054$ (from ROZAS et al. 1995) has remained approximately constant along the history of the sample, then the conversion rate per the sequenced region per generation would be equal to $\varepsilon = 3.40 \times 10^{-8}$ (here we are assuming that the conversion rate remains unaltered in heterokaryotypes, CHOVNICK 1973). The estimate of the conversion rate per base pair per generation is $\varepsilon_{fp} = (3.40 \times 10^{-6} \times 121.95/1518) = 2.73 \times 10^{-7}$.

**DISCUSSION**

The enhancing effect of gene conversion on the extant haplotypic diversity depends on two main parameters: (1) the rate of gene conversion events per generation and (2) the average length of a conversion tract. Our model allows estimation of both parameters for population DNA sequences granted we have subsets of very divergent haplotypes or those with a high value of $\psi$. Our model can, therefore, be applied to a broad data set.

We distinguish between observed and true tract lengths. Observed lengths, if considered as true lengths, will considerably underestimate the true distribution, especially when the $\psi$ values are lower than 0.01, which is probably frequently the case. Because of the positive correlation between expected lengths and $\psi$ values, a spurious relationship (for example, that the true tract lengths are larger for more divergent sequences) could be inferred between these variables if observed lengths are not corrected.

We show that many conversion events are undetected for a wide range of $\psi$ and $\phi$ values. This result is in agreement with the theoretical works of HUDSON and KAPLAN (1985) and STEPHENS (1986), which show that conventional recombination events (single crossovers) are usually undetectable in homologous DNA sequences. Consequently, there will be a tendency to detect only those conversion events occurring between very divergent haplotypes or those with a high value of $\phi$. This explains why gene conversion is recognized when it occurs between paralogous sequences (SLIGHTOM et al. 1980), but often overlooked when occurring among homologous sequences. This bias, that can be corrected by the factor $1/P(S \geq 2)$ from expression $2$, can result in an underestimation of the importance of gene conversion as a recombination factor.

The parameter $\psi$ is a measure related with the extent of differentiation among sequences from different subpopulations. Nevertheless, unlike other measures of genetic differentiation between populations $\left(d_\alpha, d_s; \text{NEI} 1987\right)$, $\psi$ takes into account both the pair base nucleotide differences within and between populations, and the direction of a conversion event (one population is donor and the other is receiver). NEI's distance measures, for example, do not distinguish between donor and receiver populations. In addition, we introduce a criterion that considers the type of polymorphism at a site to decide whether or not a converted nucleotide is detectable. The criterion is, however, restrictive and can be modified depending on the amount of differentiation between subpopulations. So, when the value of $E(S) = \psi E(N)$ (the expected number of informative sites included in a conversion tract) is $>2$, conditions (1) and (2) of our criterion given in APPENDIX A can be relaxed.

Our estimators work if, and only if, each observed tract comes from one single conversion event. The older a conversion tract, the higher the probability of being part of a recurrent conversion event. Likewise, the higher the density of conversion tracts in a region, the higher the probability of recombination in a previously converted tract. If the per site probability of a converted site between subpopulations is lower than 0.05, then the probability of an informative site being involved in a recombination event more than once is negligible. Otherwise, the per site probability of a converted site being involved in a subsequent recombination within a subpopulation is assumed to be of the same order of magnitude than between subpopulations. This assumption could not hold in inversions because of the suppression of crossover. However, at the molecular level, the probability of a crossover seems to be 20% that of a conversion event (HILLIKER et al. 1991), thereby inversions would not change significantly the probability of a subsequent recombination in a tract. Another implicit assumption in our model is that the observed tracts are a random sample of all new occurring tracts, which implies, for example, the absence of selective differences associated with tract lengths.

The effect of recombination on nucleotide variation in samples of DNA sequences is typically summarized using the parameter $4Nr$, which captures the distribution of linkage disequilibria among sites and the variance of the number of segregating sites in the sampled sequences (HUDSON 1987). Yet the parameter $4Nr$ does not distinguish between the two basic recombination processes: crossing over and gene conversion. It fails to take advantage of each process's specific footprint on patterns of nucleotide variation. By using our method to identify conversion events in the sampled sequences, we can estimate the number and the length distribution of gene conversion tracts, assessing, therefore, the relative importance of gene conversion as a factor of haplotypic differentiation.
From ROZAS and AGUADE’s data it could be estimated that around 10 of the 19 (19 × P(S > 0)) estimated conversion events result in the transfer of one or more different nucleotides between arrangements, thereby generating new haplotypes. The estimate of the conversion rate per base pair per generation is $\hat{c}_m = 2.73 \cdot 10^{-7}$. This value is about two orders of magnitude higher than the estimated value of the mutation rate $2.73 \cdot 10^{-9}$. This value is about two orders of magnitude higher than the estimated value of the mutation rate per base pair per generation is $2.73 \cdot 10^{-9}$ (ROZAS and AGUADE 1994). Given that only 34 sequences were analyzed, this indicates that gene conversion is indeed a main factor generating intragenic sequence diversity.

We appreciate very much the helpful discussions and/or comments of A. BERRY, M. AGUADE, J. BRANTZMAN, A. RUEZ, W. STEPHAN and two anonymous referees. This work has been supported by grants PB93-0607 and PB94-0923 awarded, respectively, to A. RUEZ and M. AGUADE by the Dirección General de la Investigación Científica y Técnica, Spain. E.B. and A.N. have been supported by doctoral F.P.I. fellowships from the DGU (Generalitat de Catalunya, Catalonia, Spain).

LITERATURE CITED


SCHUTTM, J. L., A. E. BLEicher, and O. SMITHIES, 1980 Human fetal $\alpha$- and $\gamma$-globin genes: complete nucleotide sequences. DNA can be exchanged between these duplicated genes. Cell 21: 627–638.


Communicating editor: B. S. WEIR

APPENDIX A

Algorithms for conversion tract detection and $\psi$ estimation: We introduce first an operational criterion with two frequency-dependent conditions to decide if a given nucleotide variant in a particular site is informative of a conversion event. Consider a conversion tract in a chromosome of subpopulation 2 (receiver) coming from a chromosome of subpopulation 1 (donor). Nucleotide variant $i$ (at a particular site) is assumed to be informative of a conversion tract if the following conditions are met.

Condition 1: Its relative frequency in subpopulation 2 is 20% or less (the same applies when the subpopulation 1 is the receiver subpopulation). The reasoning for this condition is as follows: an observed tract is defined by at least two co-converted sites. An apparent tract with two co-converted sites can also be originated by two independent parallel mutations. Given two parallel mutations in subpopulation 2, the probability that they appear on the same chromosome, according to our criterion and assuming independent association (linkage equilibrium) among sites within subpopulation 2, is $P = 0.2^2 = 0.04$. This probability is less than the conventional signification level of 0.05 (of course, if the number of informative co-converted sites is $>2$, this probability will be much lower). This condition implies that five or more sequences must be determined for the receiver (converted) subpopulation.

Condition 2: The relative frequency of nucleotide variant $i$ in subpopulation 1 (donor) is three or more
times higher than in subpopulation 2 (receiver). Here we make the plain assumption that a transferred nucleotide is quite more frequent in the donor than in the receiver subpopulation. This condition implies that at least three sequences must be determined for the donor subpopulation. Following these conditions, we propose both an algorithm for tract detection and for the estimation of parameter \( \psi \).

Algorithm for tract detection: Detection of gene conversion tracts from subpopulation 1 (donor) to subpopulation 2 (receiver). Let \( s_1 \) and \( s_2 \) be the number of analyzed sequences in subpopulation 1 and 2, respectively. Let \( I_{jk2} \) be the informative state of nucleotide variant \( i \) (\( i = 1, 2, 3 \) or 4 corresponding to nucleotides A, T, C or G) at site \( j \) of the sequence \( k \) in subpopulation 2. There are two possible states at this site: informative when \( I_{jk2} = 1 \), and noninformative, when \( I_{jk2} = 0 \). \( I_{jk2} \) is defined by the above criterion as follows:

\[
I_{jk2} = \begin{cases} 
1 & \text{if } (x_{j1} \approx x_{j2}) \text{ and } (x_{j2} \leq 0.2) \\
0 & \text{otherwise},
\end{cases}
\]  

(A1)

where \( x_{j1} \) is the relative frequency of the nucleotide variant present at site \( j \) in subpopulation 1. \( x_{j2} \) is the relative frequency of nucleotide variant \( i \) at site \( j \) in subpopulation 2. A putative tract must not contain incongruent nucleotides. We consider that a nucleotide within a putative tract in population 2 is incongruent with an origin in population 1 when \( x_{j1} = 0 \) and \( x_{j2} > 0.2 \).

If the outermost informative nucleotide sites of a congruent tract are \( T_L \) (left) and \( T_R \) (right), \( T_R > T_L \), then the observed tract length is simply the length spanned by these nucleotides: \( L = T_R - T_L + 1 \).

The detection of gene conversion tracts from subpopulation 2 (donor) to subpopulation 1 (receiver) follows the same reasoning.

ROZAS and AGUADE’s (1994) data were used to illustrate the algorithm for detection. In the line 111, for example, they found five informative nucleotide variants (G in site 48, C in site 62, G in site 64, G in site 68, and G in site 112) without incongruent variants among them (Figure 1). From these sites, a conversion tract was identified of length 65 bp. In total, five different conversion tracts were detected with lengths 12, 17, 39, 65 and 122 bp (Figure 1). The average observed length of the five tracts being 51 bp.

Algorithm for the estimation of \( \psi \): To make the whole model internally consistent, the algorithm for the estimation of \( \psi \) must be in agreement with that of detecting a conversion tract. The probability that a site is informative is the sum of the information of its constituent nucleotides. More specifically, \( \psi_j \), the probability of detecting a conversion event at site \( j \), is defined as

\[
\psi_j = \left[ \sum_{i=1}^{4} x'_{j1} p_{j1k} + \sum_{i=1}^{4} x'_{j2} p_{j2k} \right] / 2. 
\]  

(A2)

\( x'_{j1} \) is the relative frequency of the nucleotide variant \( i \) present at site \( j \) in subpopulation \( k \). For computing \( x'_{j1} \), each informative nucleotide within a detected tract is replaced by the most frequent nucleotide at that site in its respective subpopulation. \( p_{j1k} \) is the probability of observing the transfer of nucleotide variant \( i \), from subpopulation 1 to 2, at site \( j \), and takes the following values:

\[
p_{j1k} = \begin{cases} 
1 & \text{if } (x'_{j1} \approx 3 x'_{j2}) \text{ and } (x'_{j2} \leq 0.2) \\
0 & \text{otherwise}.
\end{cases}
\]  

(A3)

\( x'_{j1} \) and \( x'_{j2} \) refers to the relative frequency of nucleotide variant \( i \) before and after a hypothetical conversion tract (number of sequences in population 2). The probability of observing the transfer of the nucleotide variant \( i \) at site \( j \) from subpopulation 2 to 1, \( p_{j2k} \), is defined in the same way.

The average value of \( \psi \) per site is

\[
\bar{\psi} = \frac{\sum_{j=1}^{n} \psi_j}{m},
\]  

(A4)

where \( m \) is the total number of nucleotide sites compared.

As an example, the estimation of \( \psi \) (using Equation A2) at sites 48, 377, 442 and 694 of Figure 1 is \( \psi = 1.000, 0.250, 0.156 \) and 0.006, respectively. The average value of \( \psi \) for the whole region (1518 bp) is, from Equation A4, 0.00847.

Computer simulations: We have carried out different computer simulations to check the robustness of the parameter \( \psi \) in different sequence configurations. The simulation steps were as follows:

Step 1: Forty monomorphic sequences of DNA of 12,000 nucleotides each were generated (20 sequences from each subpopulation).

Step 2: Nucleotide polymorphisms are thrown down at random in both subpopulations until the sequences have reached a predetermined value of \( \psi \). We employed two types of polymorphic sites: (1) sites fixed in both subpopulations but with different nucleotide variant in each subpopulation (as site 3 in Figure 2), (2) sites with two different nucleotides segregating in a subpopulation (sites 8–11 in Figure 2). We performed simulations with different proportions of both types of polymorphic sites.

Step 3: In any simulation (10,000 replicates with a fixed value of \( \psi \) and a fixed proportion of polymorphic sites), a tract of fixed length is converted from one subpopulation to the other. For each replicate, we determined the observed gene conversion tract length. Finally, estimates were computed as the average values of replicates.

Regardless of the nucleotide composition of the sub-
TABLE A1

Results obtained in the computer simulations

<table>
<thead>
<tr>
<th>n</th>
<th>$\phi$</th>
<th>$L_s \pm $SD</th>
<th>$L_e \pm $SD</th>
<th>$U_s$</th>
<th>$U_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.002</td>
<td>52.86 ± 34.62</td>
<td>53.87 ± 35.90</td>
<td>0.962</td>
<td>0.963</td>
</tr>
<tr>
<td>150</td>
<td>0.005</td>
<td>57.72 ± 36.43</td>
<td>57.85 ± 36.76</td>
<td>0.817</td>
<td>0.827</td>
</tr>
<tr>
<td>150</td>
<td>0.010</td>
<td>64.76 ± 37.79</td>
<td>64.85 ± 37.71</td>
<td>0.548</td>
<td>0.557</td>
</tr>
<tr>
<td>150</td>
<td>0.020</td>
<td>79.65 ± 37.39</td>
<td>79.40 ± 37.53</td>
<td>0.193</td>
<td>0.196</td>
</tr>
<tr>
<td>150</td>
<td>0.050</td>
<td>112.60 ± 26.13</td>
<td>112.54 ± 26.26</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>250</td>
<td>0.002</td>
<td>90.04 ± 59.65</td>
<td>91.81 ± 60.60</td>
<td>0.910</td>
<td>0.910</td>
</tr>
<tr>
<td>250</td>
<td>0.005</td>
<td>102.36 ± 63.72</td>
<td>103.25 ± 65.23</td>
<td>0.636</td>
<td>0.644</td>
</tr>
<tr>
<td>250</td>
<td>0.010</td>
<td>124.94 ± 63.40</td>
<td>123.36 ± 63.28</td>
<td>0.282</td>
<td>0.286</td>
</tr>
<tr>
<td>250</td>
<td>0.020</td>
<td>160.75 ± 56.78</td>
<td>160.47 ± 55.99</td>
<td>0.036</td>
<td>0.039</td>
</tr>
<tr>
<td>250</td>
<td>0.050</td>
<td>212.95 ± 27.96</td>
<td>212.01 ± 27.53</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>400</td>
<td>0.002</td>
<td>149.93 ± 98.79</td>
<td>153.29 ± 98.46</td>
<td>0.807</td>
<td>0.809</td>
</tr>
<tr>
<td>400</td>
<td>0.005</td>
<td>183.38 ± 101.27</td>
<td>183.61 ± 101.47</td>
<td>0.390</td>
<td>0.405</td>
</tr>
<tr>
<td>400</td>
<td>0.010</td>
<td>235.00 ± 95.44</td>
<td>233.82 ± 96.14</td>
<td>0.086</td>
<td>0.090</td>
</tr>
<tr>
<td>400</td>
<td>0.020</td>
<td>304.40 ± 67.28</td>
<td>303.01 ± 67.42</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>400</td>
<td>0.050</td>
<td>362.29 ± 27.51</td>
<td>362.00 ± 27.57</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Only results based in simulations including a 50% of each type of polymorphic sites are shown. $n$, true gene conversion tract length; $L_s$, observed tract length in the simulations; $L_e$, expected tract length from equation (5); SD, standard deviation; $U_s$, frequency of undetected tracts in simulations; $U_e$, expected frequency of undetected tracts (from Equation 2).

populations with respect to the different classes of polymorphisms, the observed values of the tract lengths in the simulation are in good agreement with the expected theoretical ones (Equation 5). Table A1 summarizes some of our simulation results.

**APPENDIX B**

**Length and number of informative sites in a conversion tract:** Our model only takes into account the information provided by the two outermost informative sites of a conversion tract. It does not consider the potential information of the number of informative sites $s$ spanning a tract. Instead of using $P(L^* = l)$ for the ML estimation of $\phi$ (expression 8), we could use $P(L^* = l; S^* = s)$. However, $S^*$ (the variable $S$ redefined for $S \geq 2$) is a function dependent completely on $\phi$ (expression 1), and therefore this variable is redundant when considering $\phi$, that is, it does not add any additional information for the estimation of $\phi$, the model purpose. Let us show it.

The probability for $L^* = l$ and $S^* = s$ given that $N = n$ is

$$P(L^* = l \text{ and } S^* = s | N = n) = P(L^* = l | N = n) P(S^* = s | N = n).$$

The probability of the second factor is derived from expression (1), and that of the first factor can be derived from Stephens (1985), as follows:

$$P(L^* = l | N = n \text{ and } S^* = s) = \frac{(n - l + 1) \binom{s-2}{l-1}}{(s-1)^{s-1}}.$$  

Note that Stepmen’s $d$ variable is $l - 1$ here.

Consider now $k$ observed conversion events of lengths $l_1, l_2, \ldots, l_k$, with a respective number of informative sites $s_1, s_2, \ldots, s_k$. For the likelihood estimation of $\phi$, the likelihood function of an observed data set is given by

$$L(\phi) = \prod_{i=1}^{k} P(L^* = l_i \text{ and } S^* = s_i).$$  

(B2)

This expression differs from that of Equation 8 by factor

$$\binom{s-2}{l-1} \psi^{s-2} (1 - \psi)^{(s-1)},$$

which multiplies each product. Because this factor is not a function of $\phi$, the $\phi$ value that maximizes expression (B2) is the same than that of Equation 8. Consequently, to consider the number of informative sites spanning an observed tract does not change the $\phi$ estimate.