A Simple Formula Useful for Positional Cloning

Richard T. Durrett,* Kai-Yi Chen† and Steven D. Tanksley†,[1]

**Department of Mathematics, Cornell University, Ithaca, New York 14853-4201 and †Department of Plant Breeding and Department of Plant Biology, Cornell University, Ithaca, New York 14853-1902*

Manuscript received July 19, 2001
Accepted for publication October 15, 2001

ABSTRACT

We derive a formula for the distribution of the length T of the recombination interval containing a target gene and using N gametes in a region where R kilobases correspond to 1 cM. The formula can be used to calculate the number of meiotic events required to narrow a target gene down to a specific interval size and hence should be useful for planning positional cloning experiments. The predictions of this formula agree well with the results from a number of published experiments in Arabidopsis.

POSITIONAL cloning has been widely used in both plants and animals to isolate genes known only by their phenotypic effects. Underlying positional cloning is the assumption that a gene can be pinpointed with sufficient precision to narrow its location to a DNA segment small enough to be sequenced and/or subjected to transformation/complementation experiments (Tanksley *et al.* 1995; Lukowitz *et al.* 2000).

The chromosomal position of a gene targeted for positional cloning is typically defined by the closest flanking crossover events. So, if a gene is to be pinpointed to a defined segment size, a minimum of two crossovers are required, one on either side of the target gene (Figure 1). Theoretically, by having an estimate of the kilobase to centimorgan ratio for a particular genome or genomic region, one can estimate the number of meiotic gametes that one must sample to narrow the position of the gene to a prescribed physical segment of DNA. Surprisingly, despite the large number of genes that have been isolated by positional cloning and the popularity of this technique, we have been unable to find a published formula for making this calculation. It is for this reason we herein describe the derivation of such a formula and its application to positional cloning.

THE FORMULA AND ITS APPLICATIONS

These calculations assume that one can observe crossovers in gametes derived from an F_1 hybrid, which is heterozygous for the target locus, that the genotype of the target locus can be unambiguously determined, and that recombination rates can be assumed to be constant near the target locus. Let T represent the distance (in kilobases) between two crossovers that bracket a target gene; R the kb/cM ratio for the genomic region in question; N the number of gametes to sample (N is equivalent to the number of testcross progeny or two times the number of F_2 progeny); and P the probability of finding in N gametes a minimum of two crossovers (one on each side of the target gene) at a physical distance $< T$.

Then:

$$P = 1 - (1 + NT/(100R))e^{-NT/(100R)}.$$

This formula assumes that the kilobase/centimorgan ratio (R) is constant in the region of width $2T$ centered at the target gene. The formula should be applicable to plants, animals, and any organism in which screens can be made for meiotic recombination. The greatest efficiency will be obtained in populations where meiotic recombination can be deduced simultaneously for both male and female gametes (*e.g.*, F_2 or recombinant inbred populations).

To illustrate the use of this formula, suppose that we are working in a region where 250 kb corresponds to 1 cM and we are interested in a target size of 100 kb, the size of a bacterial artificial chromosome. In our example with $T = 100$ kb and $R = 250$ kb/cM, we show the results in Table 1. We have included the middle column to emphasize the fact that the answer depends only on the variables through the ratio $NT/(100R)$. One consequence of this is that the sample size needed is proportional to the estimate of the recombination rate R. If one believes that there are 750 kb/cM then the needed sample size for a given success probability will be three times as large.

Table 2 allows one to pick the sample size that will produce a given probability. For example, if we want the recombination interval to be smaller than the target with probability $P = 0.95$, then we should take $NT/(100R) = 4.744$. In our example $T = 100$ kb and $R = 250$ kb/cM, so this translates into $N = 1186$.

As a final check on our formula we compare its predic-

[1]*Corresponding author:* Department of Plant Breeding and Department of Plant Biology, 252 Emerson Hall, Cornell University, Ithaca, NY 14853-1902. E-mail: sdt4@cornell.edu

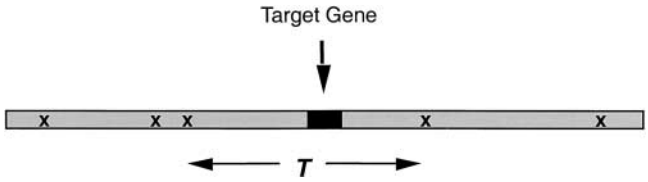

FIGURE 1.—Independent meiotic crossovers (x) observed in a segment of chromosome containing a gene targeted for positional cloning. T is the size of the genomic segment in kilobases between the two closest flanking crossovers.

tions with the results of positional cloning experiments in Arabidopsis (Table 2 in LUKOWITZ *et al.* 2000). In Table 3, N is the number of gametes (two times the number of individuals screened), T gives the estimated mapping resolution in kilobases, and P is the probability as computed by our formula of getting a result better than the one observed assuming that $R = 250$ kb/cM (the value they suggest in their article). Note that most of the P values are in the range 0.25–0.75, indicating that these experimental results are typical of what we expect. (If the assumptions underlying our formula are correct then the observed P values will be uniformly distributed between 0 and 1.) One group got lucky in localizing their target to a 20-kb interval using only a sample of size 972, while another was somewhat unlucky when their sample of size 1914 resulted in a 50-kb interval. These results and the fact that the target size is proportional to R suggest that $R = 250$ kb/cM is a good guess for Arabidopsis.

DERIVATION

The formula is obtained by combining the following facts:

1. When distances are measured in morgans, recombinations follow a Poisson process with rate 1; *i.e.*, the distances between successive crossovers are independent and have an exponential density with mean 1.
2. When recombinations from N gametes are combined, the result is a Poisson process with rate N; *i.e.*, the distances between successive crossovers are independent and have an exponential density with mean $1/N$. See, for example, DURRETT (1999, p.141).
3. The distances to the first crossover to the left (X) and right (Y) of the target site are independent exponentials with mean $1/N$, so their sum $X + Y$ has a gamma distribution. See, for example, DURRETT (1999, p. 129). This implies

$$P(X + Y > z) = (1 + Nz)e^{-Nz}.$$

The distances X and Y are measured in morgans. To convert our target from morgans to kilobases, we set $z = T/(100R)$.

The reader should note that until the last step all of our computations are exact; *i.e.*, the size of the recombination interval is measured in morgans and has a gamma distribution. In the last step we use the assumption of a constant recombination rate per unit distance near (*i.e.*, within $2T$ kb) the target to convert from morgans to kilobases.

TABLE 1

Probability that the target interval <100 kb for various sample sizes N

N	$NT/(100R)$	P
200	0.8	0.1912
400	1.6	0.4751
600	2.4	0.6916
800	3.2	0.8288
1000	4.0	0.9084
1200	4.8	0.9523
1400	5.6	0.9756
1600	6.4	0.9877
1800	7.2	0.9939

TABLE 2

Values of the design ratio $NT/(100R)$ that are needed to achieve success with probability P

P	$NT/(100R)$
0.5	1.678
0.75	2.693
0.90	3.890
0.95	4.744
0.975	5.572
0.99	6.638

DISCUSSION

Two cases cause severe reduction or suppression of crossover events: centromeric regions (heterochromatic regions) or a chromosomal inversion between two par-

TABLE 3

Results of positional cloning experiments for various Arabidopsis genes (reported in LUKOWITZ *et al.* 2000)

Gene	N	T	P
KNF	972	20	0.1832
GN	1058	50	0.6245
IFL 1	1304	30	0.4636
MP	1796	20	0.4208
BRI 1	1914	50	0.8950
IXR 1	2112	10	0.2074
NIM 1	2276	30	0.7569
ZLL	2500	20	0.5940
KN	2696	30	0.8333
WUS	3150	10	0.3589
SUP	5026	10	0.5968
CYT 1	5696	10	0.6641

ents that are used to create a mapping population that suppresses recombination. In both cases, positional cloning is not a proper strategy to isolate the target gene.

Chromosomal interference also serves to reduce crossover events in some chromosomal regions and results in inconstant R values across the whole genome. However, chromosomal interference guarantees some minimum crossovers (Copenhaver *et al.* 1998), so the formula is still useful in this situation. As noted above, we do not need to assume a constant R value across the whole genome. If the target chromosomal region has much bigger R than the estimated average of a certain species, the formula can calculate the required numbers of gametes on the basis of the new R value. For example, for the positional cloning of *HY2*, two rounds of recombinant screening were performed (Kohchi *et al.* 2001). In the first screening, *HY2* was mapped in an interval of ~360 kb equal to 0.51 cM and R was estimated to be ~700 kb/cM, which is much bigger than the average 250 kb/cM of Arabidopsis. Then the second screening used 3818 gametes to achieve the goal to delimit *HY2* in a 66-kb contig.

Richard T. Durrett was partially supported by a grant from the program in probability at the National Science Foundation (no. DMS 9877066). Steven D. Tanksley was supported by grants from the National Science Foundation (no. DBI-9872617) and U.S. Department of Agriculture Plant Genome Program (no. 97-35300-4384).

LITERATURE CITED

Copenhaver, G. P., W. E. Browne and D. Preuss, 1998 Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. USA **95:** 247–252.

Durrett, R., 1999 *The Essentials of Stochastic Processes.* Springer, New York.

Kohchi, T., K. Mukougawa, N. Frankenberg, M. Munehisa, A. Yokota *et al.*, 2001 The Arabidopsis *HY2* gene encodes phytochromobilin synthase, a ferredoxin-dependent biliverdin reductase. Plant Cell **13:** 425–436.

Lukowitz, W., C. S. Gillmor and W. R. Schiebe, 2000 Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you. Plant Physiol. **123:** 795–806.

Tanksley, S. D., M. W. Ganal and G. B. Martin, 1995 Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. Trends Genet. **11:** 63–68.

Communicating editor: J. A. Birchler