DOI: 10.1534/genetics.106.065433

Variance of the Parental Genome Contribution to Inbred Lines Derived From Biparental Crosses

Matthias Frisch and Albrecht E. Melchinger[1]

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

Manuscript received August 29, 2006
Accepted for publication February 26, 2007

ABSTRACT

The expectation of the parental genome contribution to inbred lines derived from biparental crosses or backcrosses is well known, but no theoretical results exist for its variance. Our objective was to derive the variance of the parental genome contribution to inbred lines developed by the single-seed descent or double haploid method from biparental crosses or backcrosses. We derived formulas and tabulated results for the variance of the parental genome contribution depending on the chromosome lengths and the mating scheme used for inbred line development. A normal approximation of the probability distribution function of the parental genome contribution fitted well the exact distribution obtained from computer simulations. We determined upper and lower quantiles of the parental genome contribution for model genomes of sugar beet, maize, and wheat using normal approximations. These can be employed to detect essentially derived varieties in the context of plant variety protection. Furthermore, we outlined the application of our results to predict the response to selection. Our results on the variance of the parental genome contribution can assist breeders and geneticists in the design of experiments or breeding programs by assessing the variation around the mean parental genome contribution for alternative crossing schemes.

THE expected contribution of a parental line to the genome of an inbred line derived from a biparental cross is $\frac{1}{2}$. For inbred lines derived from a backcross, the expected genome contribution of the nonrecurrent parent is $\frac{1}{2^t}$, where t is the number of backcross generations. Experimental studies showed a considerable variation in the parental genome contribution around these mean values (Heckenberger *et al.* 2006) but until now no theoretical concept for describing the variance of the parental genome contribution to homozygous inbred lines existed.

Inbred lines are developed for various purposes in genetic research and applied plant breeding programs, *e.g.*, for direct use as line cultivars or as parents of hybrid and synthetic varieties. A theoretical concept for calculating the variance of the parental genome contribution to inbred lines can be used (1) in plant variety protection to test hypotheses on the mating scheme that was employed for inbred line development and (2) to assess and compare the variability in experimental and breeding populations generated with a certain mating scheme depending on the number and length of the chromosomes of the species under consideration.

Hill (1993) derived the variance of the parental genome contribution to heterozygous backcross individuals under the assumption of no interference in crossover formation. Employing his formula for the variance, he found that a normal approximation fitted well the probability distribution of the parental genome contribution obtained from computer simulations. Using the cattle genome as an example, he demonstrated that his results can be employed to determine approximate upper bounds for the parental genome contribution of the nonrecurrent stock.

Our objectives were to (1) derive the variance of the parental genome contribution to inbred lines developed by the single-seed descent (SSD) or double haploid (DH) method from biparental crosses or backcrosses adopting the approach of Hill (1993), (2) investigate with computer simulations the fit of a normal approximation to the probability distribution of the parental genome contribution, and (3) demonstrate the application of the formulas in the context of plant variety protection.

THEORY

Assumptions: We assume that the offspring are completely homozygous lines, derived without selection from a biparental cross of completely homozygous parents P_1 and P_2. For all derivations, we assume absence of interference (Stam 1979) in crossover formation such that the recombination frequency r_{uv} between two loci on a chromosome with map positions u and v is calculated by Haldane's (1919) mapping function

$$r_{uv} = (1 - e^{-2|v-u|})/2. \tag{1}$$

[1] *Corresponding author:* Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. E-mail: melchinger@uni-hohenheim.de

TABLE 1

Formulas for the expected gametic disequilibrium *D*(*u*, *v*) between two loci at map positions *u* and *v* in populations of infinite size under four mating schemes

	D (u, v)	
Mating system	General form	After inserting Haldane's mapping function
$(F_2)^t$-SSD	$\frac{1-2r_{uv}}{4+8r_{uv}}(1-r_{uv})^t$	$=\frac{1}{2^{t+2}}\sum_{n=1}^{t+1}\binom{t}{n-1}\frac{e^{-2n\lvert v-u\rvert}}{2-e^{-2\lvert v-u\rvert}}$
$(F_1)^t$-DH	$\frac{1-2r_{uv}}{4}(1-r_{uv})^t$	$=\frac{1}{4}\left(\frac{1+e^{-2\lvert v-u\rvert}}{2}\right)^t e^{-2\lvert v-u\rvert}$
BC_t-SSD	$\frac{1}{2^{t+1}}\frac{(1-r_{uv})^t}{1+2r_{uv}}-\frac{1}{4^{t+1}}$	$=\frac{1}{4^{t+1}}\left[2\sum_{n=1}^{t}\binom{t}{n}\frac{e^{-2n\lvert v-u\rvert}}{2-e^{-2\lvert v-u\rvert}}+\frac{e^{-2\lvert v-u\rvert}}{2-e^{-2\lvert v-u\rvert}}\right]$
BC_t-DH	$\frac{1}{2^{t+1}}(1-r_{uv})^{t+1}-\frac{1}{4^{t+1}}$	$=\frac{1}{4^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}e^{-2n\lvert v-u\rvert}$

$D(u, v)$ depends on the recombination frequency r_{uv} between the two loci and the number t of intermating or backcrossing generations.

Variance of the parental genome contribution: Meiosis on different chromosomes is stochastically independent. Hence, the variance of the genome contribution Z of parent P_1 to the genome of a derived line can be written in terms of the variances $\text{Var}(Z_i)$ for individual chromosomes as

$$\text{Var}(Z)=\sum_{i=1}^{c}\left(\frac{l_i}{l}\right)^2\text{Var}(Z_i), \tag{2}$$

where c is the number of chromosomes, l_i the length of the ith chromosome, and $l=\sum_{i=1}^{c}l_i$ the total length of the genome in Morgan units.

Following the approach introduced by Hill (1993) in the context of backcross populations, the variance of the parental genome contribution to a chromosome equals the expected covariance between two randomly sampled loci on the chromosome,

$$\begin{aligned}\text{Var}(Z_i)&=E[\text{Cov}(G_u,G_v)]\\&=E[E(G_uG_v)-E(G_u)E(G_v)]\\&=E[D_{uv}]\\&=\frac{1}{l_i^2}\int_0^{l_i}\int_0^{l_i}D(u,v)\,du\,dv,\end{aligned} \tag{3}$$

where G_u and G_v are random variables taking the value 1 if the loci at map positions u and v carry the allele of parent P_2 and 0 otherwise, and D_{uv} is a random variable describing the linkage disequilibrium between two loci on the chromosome with probability density

$$D(u,v)=P(G_v=1,G_u=1)-P(G_u=1)P(G_v=1). \tag{4}$$

Using the formulas for

$$p=P(G_u=1)\quad\text{and}\quad q_{uv}=P(G_v=1\mid G_u=1) \tag{5}$$

given in Frisch and Melchinger (2006, Table 1 therein), $D(u, v)$ can be calculated as

$$D(u,v)=pq_{uv}-p^2. \tag{6}$$

We present formulas for $D(u, v)$ for the following four mating systems (Table 1): (1) $(F_2)^t$-SSD lines, developed by t ($t \geq 0$) generations of random mating of a F_2 population and subsequent application of the SSD method for line development; (2) $(F_1)^t$-DH lines, developed by t ($t \geq 0$) generations of random mating of a F_1 cross and subsequent inbred line development with the DH method; and (3) BC_t-SSD and (4) BC_t-DH lines, developed from a F_1 cross backcrossed t ($t \geq 1$) times to parent P_1, with subsequent line development by the SSD or DH method.

Inserting $D(u, v)$ (Table 1) into Equation 3 yields $\text{Var}(Z_i)$. Analytical results for $\text{Var}(Z_i)$ are derived in the appendix and summarized in Table 2. Numerical results for $\text{Var}(Z_i)$ are given in Table 3. To check our derivations, we determined the results in Table 3 also with computer simulations using Plabsoft (Maurer *et al.* 2004). The differences between simulated and analytically determined variances were < 0.001 if one million chromosomes were simulated.

Probability distribution of the parental genome contribution: The probability distribution of the parental genome contribution is determined by the number and location of crossover events occuring during the meioses in inbred line development. We investigated the probability distribution assuming no interference in crossover formation (Stam 1979), employing properties of the Poisson process (*cf.* Karlin 1968).

For an individual chromosome, the probability that exactly k crossovers occur during all meioses in inbred line development can be obtained from the probability function of the Poisson distribution. If no crossover

TABLE 2

Formulas for the variance Var(Z_i) of the parental genome contribution to a chromosome of length l_i under four mating schemes

Mating system	Var(Z_i)
$(F_2)^t$-SSD	$\frac{1}{l_i^2}\frac{1}{2^{t+2}}\sum_{n=1}^{t+1}\binom{t}{n-1}\left[2^{n-1}\xi_6+2\sum_{k=1}^{n-1}\xi_3(\xi_7-l_i)\right]^a$
$(F_1)^t$-DH	$\frac{1}{l_i^2}\frac{1}{4(t+1)}\left[l_i\left(2-\frac{1}{2^t}\right)-\frac{1}{2^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{n}(1-e^{-2l_i n})\right]$
BC_t-SSD	$\frac{1}{l_i^2}\frac{1}{4^{t+1}}\left\{\sum_{n=1}^{t}\binom{t}{n}\left[2^n\xi_6+4\sum_{k=1}^{n-1}\xi_3(\xi_7-l_i)\right]+\xi_6\right\}$
BC_t-DH	$\frac{1}{l_i^2}\frac{1}{4^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{2n^2}(2nl_i-1+e^{-2nl_i})$

[a] $\xi_3=\frac{2^{k-2}}{n-k}$

$\xi_6=l_i\ln 2-\frac{1}{2}\mathrm{dilog}\left(\frac{1}{2}\right)+\frac{1}{2}\mathrm{dilog}\left(1-\frac{1}{2}e^{-2l_i}\right)$

$\xi_7=\frac{1-e^{-2l_i(n-k)}}{2(n-k)}$

occurs ($k = 0$), then the genome contribution of parent P_1 is either 0 or 1. In consequence, the probabilities $P(Z_i = 0)$ and $P(Z_i = 1)$ do exist and the random variable Z_i is discrete for $Z_i = 0$ and $Z_i = 1$. If $k > 0$ crossovers occur, then the length of chromosome segments between crossovers is exponentially distributed and the sum of lengths of chromosome segments is gamma distributed. In consequence, Z_i is in the interval (0, 1) a mixture of linear transformations of the gamma distributions for different values of k. For the entire genome, the distribution of the parental genome contribution is a convolution of the distributions for the individual chromosomes.

Analytical results for the exact probability distribution of the parental genome contribution could be derived by employing the above considerations. However, the resulting equations would be rather unwieldy and using them to derive important parameters such as quantiles directly from the density functions would require a heavy use of high quality numerical mathematics. Alternatively, we suggest employing our relatively simple equations for the variance (Table 2) and a normal approximation instead.

DISCUSSION

Genetic model: For all derivations we used the assumption of no interference (Stam 1979) underlying Haldane's (1919) mapping function. This is a simplified mathematical model and there exist more sophisticated models of crossover formation in meiosis, which fit experimental data better (McPeek and Speed 1995). Briefly, the advantages of the assumption of no interference are

TABLE 3

Variance Var(Z_i) of the parental genome contribution to a chromosome of length l_i under four mating schemes

	Chromosome length l_i							
t	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
				$(F_2)^t$-SSD				
0	0.1410	0.1231	0.1091	0.0978	0.08860	0.0809	0.0743	0.0687
1	0.1246	0.1064	0.0927	0.0821	0.07356	0.0666	0.0608	0.0559
2	0.1110	0.0931	0.0800	0.0701	0.06232	0.0561	0.0509	0.0467
3	0.0997	0.0823	0.0700	0.0608	0.05374	0.0481	0.0435	0.0398
				$(F_1)^t$-DH				
0	0.1740	0.1566	0.1419	0.1294	0.11867	0.1094	0.1014	0.0943
1	0.1517	0.1330	0.1181	0.1060	0.09604	0.0877	0.0806	0.0745
2	0.1335	0.1145	0.1000	0.0886	0.07948	0.0720	0.0657	0.0605
3	0.1186	0.0998	0.0860	0.0754	0.06711	0.0604	0.0549	0.0503
				BC_t-SSD				
1	0.1058	0.0923	0.0818	0.0734	0.06654	0.0607	0.0558	0.0516
2	0.0576	0.0497	0.0436	0.0389	0.03500	0.0318	0.0291	0.0269
3	0.0283	0.0241	0.0209	0.0185	0.01654	0.0150	0.0137	0.0126
4	0.0133	0.0112	0.0096	0.0084	0.00749	0.0067	0.0061	0.0056
				BC_t-DH				
1	0.1194	0.1057	0.0945	0.0854	0.07769	0.0712	0.0656	0.0608
2	0.0632	0.0550	0.0486	0.0435	0.03929	0.0358	0.0328	0.0303
3	0.0306	0.0262	0.0229	0.0203	0.01821	0.0165	0.0151	0.0139
4	0.0143	0.0121	0.0104	0.0092	0.00816	0.0074	0.0067	0.0061

(1) mathematical simplicity, yielding equations that can be easily evaluated, and (2) that the results can be applied without knowing the exact amount of interference in the chromosome region under consideration. For a more detailed discussion concerning the use of the assumption of no interference see Frisch and Melchinger (2001).

Equation 3, defining the variance of the parental genome contribution in terms of the linkage disequilibrium $D(u, v)$, and the formulas for $D(u, v)$, in terms of the recombination frequency r_{uv} presented in Table 1, hold true irrespectively of the amount of interference. These formulas can be used with arbitrary mapping functions to derive the variance of the parental genome contribution under the assumption of interference. Presumably, analytical solutions as presented in the appendix cannot be derived for some mapping functions. In such cases, approximative solutions of Equation 3 can be obtained with numerical integration routines of mathematical software packages.

Compared with no interference, negative interference results in a greater number of chromosome segments with intermediate length and a smaller number of very long or short chromosome segments. Therefore, negative interference will result in smaller variances of the parental genome contribution than those presented in our results. The opposite is the case for positive interference.

Comparison with previous studies: Hill (1993) derived the variance of the parental genome contribution to backcross individuals. Each backcross individual receives from the recurrent backcross parent one set of homologous chromosomes, for which the variance of the parental genome contribution is zero. Hence, the variance of the parental genome contribution to backcross individuals is entirely determined by the variance $\text{Var}(\bar{Z}_{(n)})$ (following Hill's 1993 notation) of the parental genome contribution to the homologous chromosome set originating from the nonrecurrent parent. These homologous chromosomes are genetically identical to the chromosomes of DH lines derived from a backcross individual. In consequence, $\text{Var}(\bar{Z}_{(n)})$ derived by Hill (1993) for backcross individuals equals $\text{Var}(Z_i)$ for BC_t-DH lines.

Wang and Bernardo (2000) derived the variance $V({}_kX)$ of marker estimates of parental genome contribution to F_2- and BC_1-SSD lines. They considered a finite number k of marker loci per chromosome and employed Kosambi's (1944) mapping function $r_{uv} = \frac{1}{2}\tanh(2v - 2u)$. The major difference to our approach is that Wang and Bernardo (2000) obtain $V({}_kX)$ by summing over a discrete number of marker loci, whereas we obtain $\text{Var}(Z_i)$ by integrating over an infinite number of genomic loci (Equation 3). The results on $V({}_kX)$ and $\text{Var}(Z_i)$ can be related as follows. Inserting $D(u, v)$ (Table 1) in Equation 3, but employing Kosambi's instead of Haldane's mapping function, yields $\lim_{k\to\infty} V({}_kX) = \text{Var}(Z_i)$. In consequence, $V({}_kX)$ of Wang and Bernardo (2000) converges to $\text{Var}(Z_i)$ for large numbers of markers on a chromosome (assuming that the same mapping function is employed).

TABLE 4

Standard deviation Var$\sqrt{\text{Var(Z)}}$ of the parental genome contribution to F_2-SSD and BC_1-SSD maize lines for experimental and simulated data (Heckenberger *et al.* 2006), the model of Wang and Bernardo (2000), and the model developed in this study

	Mating system ($\sqrt{\text{Var}(Z)}$)	
	F_2-SSD	BC_1-SSD
Experimental data of Heckenberger *et al.* (2006)[a]		
100 SSR markes	0.10	0.09
1017 AFLP markers	0.10	0.11
Simulations of Heckenberger *et al.* (2006)[b]		
Entire genome	0.09	0.07
Model of Wang and Bernardo (2000)[c]		
100 markers	0.088	0.076
1020 markers	0.090	0.078
∞ markers	0.090	0.078
Model of this study		
Entire genome	0.088	0.076

[a] The linkage map consisted of 10 chromosomes of 1.70, 1.30, 1.06, 1.48, 1.28, 1.15, 1.14, 1.21, 0.99, and 0.91 M length.

[b] The linkage map of the experimental data and a noninterference model was used for the simulations.

[c] The "nonterminal marker model" of the authors was employed with 10 chromosomes of 1.22 M length and 10 SSRs or 102 AFLPs equally spaced on each chromosome.

Heckenberger *et al.* (2006) estimated the parental genome contribution to 102 F_2-SSD and 11 BC_1-SSD maize lines with 100 SSR and 1017 AFLP markers. They determined the standard deviations of the parental genome contribution (Table 4) and compared their results with computer simulations. The observed standard deviations were not significantly different (χ^2 test with $\alpha = 0.05$) from the simulated values. The standard deviations determined with Equation 2 as well as those obtained with the model of Wang and Bernardo (2000) were in good agreement with the experimental and simulated values (Table 4). In conclusion, both theoretical models fit the data set of Heckenberger *et al.* (2006) well.

Numerical results: The variance of the parental genome contribution to a chromosome depends on the expected number of crossovers occurring on the chromosome during inbred line development. A large number of expected crossovers results in many small chromosome segments, whereas few crossovers result in few long chromosome segments. With few long segments, the probability that chromosomes with very large

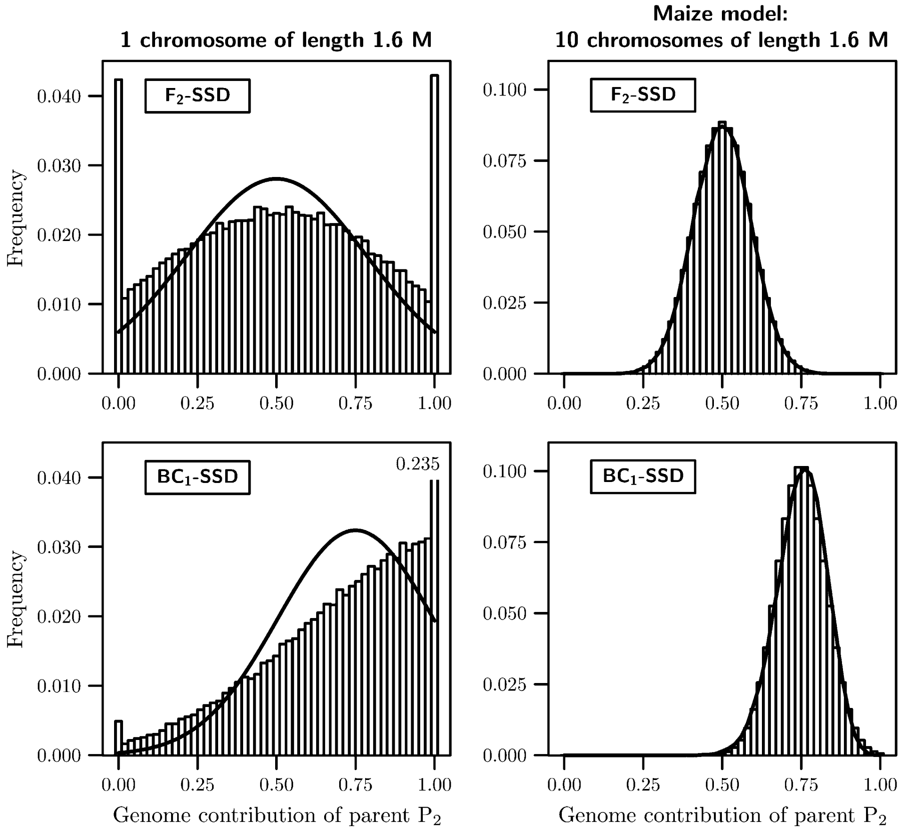

FIGURE 1.—Simulated distribution (histogram) and normal approximation (solid line) of the parental genome contribution to one chromosome of 1.6 M length (left) and to a model of the maize genome with 10 chromosomes of 1.6 M length (right) for F_2-SSD lines (top) and BC_1-SSD lines (bottom).

or very small parental genome contributions do occur is greater and, therefore, the variance of the parental genome contribution is greater than for many small segments.

The number of crossovers expected per meiosis on a chromosome equals its length in Morgan units. Therefore, the variance of the parental genome contribution is smaller for long chromosomes than for short chromosomes. This trend can be observed irrespective of the employed breeding scheme for inbred line development (Table 3).

The total number of crossovers occurring on a chromosome during inbred line development depends on the total number of meioses and, hence, the employed breeding scheme. Intermating or backcrossing prior to employing the SSD or DH method results in an increased total number of meioses and, therefore, in a smaller variance of the parental genome contribution (Table 3). Generating DH lines comprises only one meiosis, whereas in the SSD scheme one meiosis occurs in each selfing generation. Therefore the variances of the parental genome contribution is greater for DH than for SSD lines.

Normal approximation: A normal approximation is not expected to fit the distribution of the parental genome contribution for individual chromosomes well, because $Z_i = 0$ and $Z_i = 1$ can occur with rather high probabilities, especially for short chromosomes or when inbreds are generated by the DH method. However, the genomes of important crops consist of many chromosomes (9 in sugar beet, 10 in maize, and 21 in wheat). Therefore, the random variable describing the parental genome contribution to the entire genome is a sum of independent random variables for the individual chromosomes. According to the central limit theorem (SHAO 1999) the probability distribution of a sum of a large number of random variables converges to a normal distribution, irrespective of the type of distributions of random variables that are summed up. As a consequence, theory suggests that a normal approximation of the probability distribution of the parental genome contribution to the entire genome should fit the true distribution well.

To investigate the fit of the normal approximation, we used the software Plabsoft (MAURER *et al.* 2004) to simulate the parental genome contribution to (a) one chromosome of 1.6 M length and (b) a model of the maize genome consisting of 10 chromosomes each of 1.6 M length for the F_2-SSD and BC_1-SSD mating schemes. The normal approximations fit the simulated distributions of the parental genome contribution for individual chromosomes only poorly (Figure 1). In contrast, the fit was very good for the simulated distribution of the entire genomes for both F_2-SSD and BC_1-SSD

TABLE 5

Quantiles of the parental genome distribution for models of the genomes of sugar beet (9 chromosomes of 1.0 M length), maize (1.0 chromosomes of 1.6 M length), and wheat (21 chromosomes of 1.8 M length)

	Quantile								
Mating system	0.005	0.025	0.050	0.900	0.950	0.975	0.990	0.995	0.999
Sugar beet model: 9 chromosomes of length 1 M									
F_2-SSD	0.216	0.284	0.319	0.641	0.681	0.716	0.756	0.784	0.840
$(F_2)^2$-SSD	0.257	0.315	0.345	0.621	0.655	0.685	0.719	0.743	0.791
F_1-SSD	0.177	0.254	0.293	0.661	0.707	0.746	0.792	0.823	0.888
$(F_1)^2$-SSD	0.228	0.293	0.327	0.635	0.673	0.707	0.745	0.772	0.826
BC_1-SSD	0.504	0.563	0.593	0.872	0.907	0.937	0.972	0.996	1.000
BC_1-DH	0.486	0.549	0.581	0.881	0.919	0.951	0.988	1.000	1.000
Maize model: 10 chromosomes of length 1.6 M									
F_2-SSD	0.268	0.324	0.352	0.615	0.648	0.676	0.709	0.732	0.778
$(F_2)^2$-SSD	0.307	0.353	0.377	0.596	0.623	0.647	0.674	0.693	0.731
F_1-DH	0.231	0.295	0.328	0.634	0.672	0.705	0.743	0.769	0.823
$(F_1)^2$-DH	0.281	0.334	0.360	0.609	0.640	0.666	0.697	0.719	0.762
BC_1-SSD	0.549	0.597	0.622	0.850	0.878	0.903	0.931	0.951	0.991
BC_1-DH	0.533	0.585	0.611	0.858	0.889	0.915	0.946	0.967	1.000
Wheat model: 21 chromosomes of length 1.8 M									
F_2-SSD	0.347	0.383	0.402	0.576	0.598	0.617	0.638	0.653	0.684
$(F_2)^2$-SSD	0.373	0.403	0.419	0.563	0.581	0.597	0.615	0.627	0.652
F_1-DH	0.321	0.364	0.386	0.589	0.614	0.636	0.662	0.679	0.715
$(F_1)^2$-DH	0.356	0.390	0.408	0.572	0.592	0.610	0.630	0.644	0.673
BC_1-SSD	0.617	0.649	0.665	0.816	0.835	0.851	0.870	0.883	0.909
BC_1-DH	0.606	0.640	0.658	0.822	0.842	0.860	0.880	0.894	0.923

lines. Hence, our formulas for the variances, together with a normal approximation, provide a good means by which to investigate the distribution of the parental genome contribution in many applications in genetics and breeding.

Application in plant variety protection: An essentially derived variety is a cultivar or an inbred line, which is for the most part identical to one of its ancestors. Essentially derived varieties can be detected by comparing predictions of the parental genome contribution to inbred lines with threshold values. The variances of the parental genome contribution derived here can be employed together with the prediction method described in a companion article (Frisch and Melchinger 2006) to establish a test for detecting essentially derived varieties.

The first step of the test is to identify breeding schemes that are generally considered acceptable for inbred line development. For example, in wheat breeding in Europe, it is an accepted breeding scheme to cross a proprietary inbred line with a registered line cultivar of a competitor and to select a new line cultivar from the resulting population of F_2-SSD lines. In contrast, deriving inbred lines from backcross populations is not accepted.

Then the null hypothesis, "An inbred line was derived using an accepted breeding scheme," is tested. The critical value for the test is determined from the quantiles of a normal approximation of the distribution of the parental genome contribution under the null hypothesis. For example, in wheat, the 0.99 quantile of the parental genome contribution to F_2-SSD lines is 0.638 (Table 5). As test statistic, the genome contribution of the parental line that is assumed to be plagiarized to the putative essentially derived variety is determined by using the prediction method of Frisch and Melchinger (2006). If the test statistic is greater than the critical value, then the null hypothesis is rejected and plagiarism is assumed. (Of course, the accused breeder always has the possibility to prove that an accepted method was employed, *e.g.*, by disclosing the breeding records.)

For use as threshold values, we determined quantiles of the parental genome contribution for model genomes of sugar beet (9 chromosomes of 1.0 M length), maize (10 chromosomes of 1.6 M length), and wheat (21 chromosomes of 1.8 M length) by employing normal approximations (Table 5). The upper quantiles were considerably lower for long genomes than for short ones, *e.g.*, the 0.95 quantile for F_2-SSD lines was 0.598 for wheat and 0.681 for sugar beet. Breeding schemes with intermating before inbred line development had slightly smaller 0.95 quantiles than the corresponding breeding schemes without intermating.

The upper quantiles for F_1-DH lines were considerably greater than those for F_2-SSD lines. For example, the 0.95 quantile for F_2-SSD lines of maize was 0.648, whereas for F_1-DH lines it was 0.672 (Table 5). Typically, the expectation of the parental genome contribution is the criterion that determines acceptance or nonacceptance of a certain breeding scheme for inbred line development. The F_2-SSD scheme is often suggested as an accepted breeding method for determining critical threshold values (*cf.* Heckenberger *et al.* 2006). If F_2-SSD lines are considered acceptable, then F_1-DH lines should also be considered acceptable, because both have an expected parental genome contribution of one-half. However, F_1-DH lines have a considerably greater variance of the parental genome contribution (Table 3) and, consequently, greater upper quantiles (Table 5). Therefore, the F_1-DH mating scheme seems in general more appropriate than the F_2-SSD scheme for determining threshold values.

The test described above can be modified by using alternative test statistics or/and alternative methods to determine critical values. Alternative predictors of the parental genome contribution for use as test statistics were discussed by Frisch and Melchinger (2006), and alternative methods to determine critical threshold values were proposed by Smith *et al.* (1995), Wang and Bernardo (2000), and Heckenberger *et al.* (2005).

Smith *et al.* (1995) suggested employing fixed threshold values and proposed a parental genome contribution of 0.9 as threshold for maize lines. Compared with using fixed values as thresholds, our method has the advantage that it is genetically justified. For F_2 and F_1 derived lines of maize, the 0.999 quantiles of the parental genome contribution ranged between 0.73 and 0.82 (Table 5). In consequence, employing 0.9 as threshold value results in a low power of detecting backcross-derived inbreds.

Wang and Bernardo (2000) suggested determining threshold values using the variance of marker estimates of the parental genome contribution. Compared with the method of Wang and Bernardo (2000), our method has the advantage that the threshold values (Table 5) are independent of the employed set of molecular markers.

Heckenberger *et al.* (2005) suggested determining threshold values with computer simulations. Our results on the quantiles of the parental genome contribution for F_2-SSD lines of maize were in good agreement with the corresponding results of Heckenberger *et al.* (2005). However, our method has the advantage that no computer simulations are required.

Application in selection theory: Selection for parental marker alleles in backcross populations was investigated and a comprehensive selection theory was developed by Frisch and Melchinger (2005). That approach takes into account (a) the exact distribution of the parental genome contribution and (b) that selection for the parental alleles at marker loci is actually an indirect selection for the parental alleles at all loci of the entire genome. However, such theory is not available for inbred lines developed with the mating schemes considered in this study. Using a simpler mathematical model that neglects (a) and (b), the variances of the parental genome contribution can be employed to estimate the response to selection.

We consider a population of inbred lines analyzed for a large number of polymorphic molecular markers, which are covering the entire genome without larger gaps (*e.g.*, one marker per centimorgan). Selection is carried out for the alleles of one parental line and the marker score is regarded as the target trait for selection. Under these assumptions, an approximate pre-test estimate of the response to selection R can be obtained adopting from standard selection theory (Falconer and Mackay 1996, p. 189, Equation 11.3),

$$R = ih^2\sigma_p, \tag{7}$$

where i is the selection intensity, h^2 the heritability, and σ_p the square root of the phenotypic variance. Assigning a heritability of $h^2 = 1$ for the markers and using the variance of the parental genome contribution as phenotypic variance we obtain

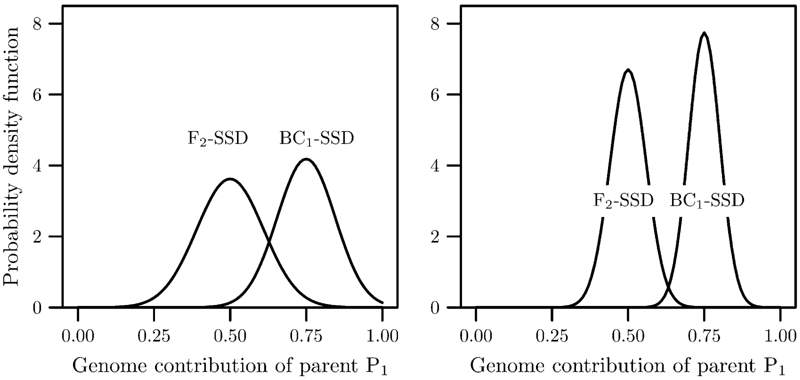

Figure 2.—Probability density functions of the normal approximation of the parental genome contribution to F_2-SSD and BC_1-SSD lines in (left) sugar beet (9 chromosomes of 1 M length) and (right) wheat (21 chromosomes of 1.8 M length).

$$R = i\sqrt{\mathrm{Var}(Z)}. \tag{8}$$

Further applications: In addition to the above applications, the results presented are of general interest for breeders and geneticists because they allow comparison of the distribution of the parental genome contribution for alternative mating schemes.

For example, an important goal in second-cycle breeding is the development of inbred lines that share the general characteristics with one parental line and are improved by specific characteristics of a second crossing partner. Such derived lines are then used as a replacement for parental lines in a breeding program. As a rule of thumb, the breeder may attempt to derive lines with a parental genome contribution of 3/4 from the parental line, which should be replaced by the derived line. The probability distribution of the parental genome contribution can help to assess the suitability of mating schemes to deliver such inbred lines. For sugar beet, the overlap of the probability density functions of the parental genome contribution to F_2-SSD and BC_1-SSD lines is considerable (Figure 2) and it is possible to select lines with a parental genome contribution of 70–75% from an F_2-derived population. In contrast, for wheat, F_2-SSD lines with parental genome contributions of 3/4 or more from one crossing partner do occur only with an extremely small probability (Figure 2). Therefore, in wheat a BC_1-derived population must be generated to be able to select lines with the desired parental genome contribution.

These examples demonstrate that our results can be used to assess the expected variation of the parental genome contribution in populations derived from planned crosses of parental lines, depending on the number and length of the chromosomes of the species. This information can help breeders and geneticists in the design of breeding programs and experiments.

We thank Frank M. Gumpert for checking the derivations in the APPENDIX and an anonymous reviewer for helpful comments and suggestions.

LITERATURE CITED

Falconer, D. S., and T. C. F. Mackay, 1996 *Introduction to Quantitative Genetics,* Ed. 4. Longman Group, Harlow, UK.

Frisch, M., and A. E. Melchinger, 2001 The length of the intact chromosome segment around a target gene in marker-assisted backcrossing. Genetics **157:** 1343–1356.

Frisch, M., and A. E. Melchinger, 2005 Selection theory for marker-assisted backcrossing. Genetics **170:** 909–917.

Frisch, M., and A. E. Melchinger, 2006 Marker-based prediction of the parental genome contribution to inbred lines derived from biparental crosses. Genetics **174:** 795–803.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman *et al.*, 2006 *GNU Scientific Library Reference Manual,* Ed. 2, V1.8. Network Theory, Bristol, UK.

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distance between the loci of linkage factors. J. Genet. **8:** 299–309.

Heckenberger, M., M. Bohn, M. Frisch, H. P. Maurer and A. E. Melchinger, 2005 Identification of essentially derived varieties with molecular markers: an approach based on statistical test theory and computer simulations. Theor. Appl. Genet. **111:** 598–608.

Heckenberger, M. J. Muminovic, J. R. van der Voort, J. Peleman, M. Bohn, and A. E. Melchinger, 2006 Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. III. AFLP data from maize inbreds and comparison with SSR data. Mol. Breeding **17:** 111–125.

Hill, W. G., 1993 Variation in genetic composition in backcrossing programs. J. Hered. **84:** 212–213.

Karlin, S., 1968 *A First Course in Stochastic Processes.* Academic Press, New York.

Kosambi, D. D., 1944 The estimation of the map distance from recombination values. Ann. Eugen. **12:** 172–175.

Maurer, H. P., A. E. Melchinger and M. Frisch, 2004 Plabsoft: Software for simulation and data analysis in plant breeding. Proceedings of the 17th Eucarpia General Congress, September 8–11, 2004, Tulln, Austria, pp. 359–362.

McPeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics **139:** 1031–1044.

Shao, J., 1999 *Mathematical Statistics.* Springer-Verlag, New York.

Smith, J. S. C., D. S. Ertl and B. A. Orman, 1995 Identification of maize varieties, pp. 253–264 in *Identification of Food-Grain Varieties,* edited by C. W. Wrigley. American Association of Cereal Chemists, St. Paul, MN.

Stam, P., 1979 Interference in genetic crossing over and chromosome mapping. Genetics **92:** 573–594.

Wang, J., and R. Bernardo, 2000 Variance of marker estimates of parental contribution to F2 and BC1-derived inbreds. Crop Sci. **40:** 659–665.

Communicating editor: R. W. Doerge

APPENDIX

We derive the variance of the parental genome contribution to a chromosome according to Equation 3 for four mating systems.

BC_t-DH lines:

Inserting $D(u, v)$ for BC_t-DH lines (Table 1) into Equation 3 yields

$$\mathrm{Var}(Z_i) = \frac{1}{l_i^2}\frac{1}{4^{t+1}}\sum_{n=1}^{t+1}\left[\begin{pmatrix} t+1 \\ n \end{pmatrix}\int_0^{l_i}\int_0^{l_i} e^{-2n|v-u|}du\,dv\right]. \tag{A1}$$

With

$$\begin{aligned}\int_0^{l_i} e^{-2n|v-u|}du &= \int_0^{v} e^{-2n(v-u)}du + \int_v^{l_i} e^{2n(v-u)}du\\ &= \left[\frac{1}{2n}e^{-2n(v-u)}\right]\Bigg|_0^v + \left[\frac{1}{-2n}e^{2n(v-u)}\right]\Bigg|_v^{l_i}\\ &= \frac{1}{2n}[1 - e^{-2nv} - e^{-2n(l_i-v)} + 1] \end{aligned} \tag{A2}$$

and

$$\frac{1}{2n}\int_0^{l_i} 2 - e^{-2nv} - e^{-2n(l_i-v)}dv = \frac{1}{2n^2}\left(2nl_i - 1 + e^{-2nl_i}\right) \tag{A3}$$

we get

$$\mathrm{Var}(Z_i) = \frac{1}{l_i^2}\frac{1}{4^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{2n^2}\left(2nl_i - 1 + e^{-2nl_i}\right). \tag{A4}$$

BC$_t$-SSD lines:

For BC$_t$-SSD lines we have (Table 1)

$$\int D(u,v)du = \frac{1}{4^{t+1}}\left[2\sum_{n=1}^{t}\binom{t}{n}\int\frac{e^{-2|v-u|n}}{2 - e^{-2|v-u|}}du + \int\frac{e^{-2|v-u|}}{2 - e^{-2|v-u|}}du\right]. \tag{A5}$$

We consider the second indefinite integral in Equation A5 for the case $u \leq v$ and set

$$f(u) = 2 - e^{-2(v-u)} \quad \text{and} \quad f'(u) = -2e^{-2(v-u)}, \tag{A6}$$

and with logarithmic integration we get

$$\int\frac{e^{-2(v-u)}}{2 - e^{-2(v-u)}}du = -\frac{1}{2}\int\frac{f'(u)}{f(u)}du = -\frac{1}{2}\ln|f(u)|. \tag{A7}$$

Applying the same principle to the case $u > v$ we get

$$\int\frac{e^{-2|v-u|}}{2 - e^{-2|v-u|}}du = \begin{cases} -\frac{1}{2}\ln[2 - e^{-2(v-u)}] & \text{for } u \leq v\\ \frac{1}{2}\ln[2 - e^{2(v-u)}] & \text{for } u > v. \end{cases} \tag{A8}$$

We now consider the first indefinite integral in Equation A5. Adding to the numerator

$$0 = \sum_{k=1}^{n-1}\left[-2^k e^{-2|v-u|(n-k)} + 2^k e^{-2|v-u|(n-k)}\right] \tag{A9}$$

and applying

$$2^{k-1}e^{-2|v-u|(n-k+1)} - 2^k e^{-2|v-u|(n-k)} = -(2 - e^{-2|v-u|})(2^{k-1}e^{-2|v-u|(n-k)}) \tag{A10}$$

we get

$$\begin{aligned}\int\frac{e^{-2|v-u|n}}{2 - e^{-2|v-u|}}du &= \int\left[\frac{2^{n-1}e^{-2|v-u|}}{2 - e^{-2|v-u|}} - \sum_{k=1}^{n-1}2^{k-1}e^{-2|v-u|(n-k)}\right]du\\ &= \begin{cases} -2^{n-2}\ln[2 - e^{-2(v-u)}] - \sum_{k=1}^{n-1}\frac{2^{k-2}}{n-k}e^{-2(v-u)(n-k)} & \text{for } u \leq v\\ 2^{n-2}\ln[2 - e^{2(v-u)}] + \sum_{k=1}^{n-1}\frac{2^{k-2}}{n-k}e^{2(v-u)(n-k)} & \text{for } u > v. \end{cases}\end{aligned} \tag{A11}$$

Hence, we get for a fixed value of v,

$$\int_0^{l_i} D(u,v)\,du = \frac{1}{4^{t+1}}\left\{2\sum_{n=1}^{t}\binom{t}{n}\left[2^{n-2}(\xi_1+\xi_2)+\sum_{k=1}^{n-1}\xi_3(\xi_4+\xi_5-2)\right]+\frac{1}{2}(\xi_1+\xi_2)\right\}, \tag{A12}$$

where

$$\begin{aligned}
\xi_1 &= \ln(2-e^{-2v}), & \xi_2 &= \ln(2-e^{2(v-l_i)})\\
\xi_3 &= \frac{2^{k-2}}{n-k}, & \xi_4 &= e^{-2v(n-k)}\\
\xi_5 &= e^{2(v-l_i)(n-k)}.
\end{aligned} \tag{A13}$$

For symmetry reasons

$$\int_0^{l_i}\xi_1\,dv = \int_0^{l_i}\xi_2\,dv \quad \text{and} \quad \int_0^{l_i}\xi_4\,dv = \int_0^{l_i}\xi_5\,dv. \tag{A14}$$

Employing the dilogarithm function (*cf.* Galassi *et al.* 2006)

$$\begin{aligned}
\int \ln(2-e^{-2v})\,dv &= \int \ln 2 + \int \ln\left(1-\frac{1}{2}e^{-2v}\right)dv\\
&= \int \ln 2 - \int \sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{1}{2}e^{-2v}\right)^k dv\\
&= v\ln 2 + \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k^2}\left(\frac{1}{2}e^{-2v}\right)^k\\
&= v\ln 2 + \frac{1}{2}\operatorname{dilog}\left(1-\frac{1}{2}e^{-2v}\right)
\end{aligned} \tag{A15}$$

we get

$$\xi_6 = \int_0^{l_i}\ln(2-e^{-2v})\,dv = l_i\ln 2 - \frac{1}{2}\operatorname{dilog}\left(\frac{1}{2}\right)+\frac{1}{2}\operatorname{dilog}\left(1-\frac{1}{2}e^{-2l_i}\right). \tag{A16}$$

Using this and

$$\xi_7 = \int_0^{l_i} e^{-2v(n-k)}\,dv = \frac{1-e^{-2l_i(n-k)}}{2(n-k)} \tag{A17}$$

yields

$$\begin{aligned}
\operatorname{Var}(Z_i) &= \frac{1}{l_i^2}\int_0^{l_i}\int_0^{l_i} D(u,v)\,du\,dv\\
&= \frac{1}{l_i^2 4^{t+1}}\left\{\sum_{n=1}^{t}\binom{t}{n}\left[2^n\xi_6+4\sum_{k=1}^{n-1}\xi_3(\xi_7-l_i)\right]+\xi_6\right\}.
\end{aligned} \tag{A18}$$

$(\mathrm{F}_1)^t$-DH lines:

For $(\mathrm{F}_1)^t$-DH lines we have (Table 2)

$$D(u,v) = \frac{1}{4}\left(\frac{1+e^{-2|v-u|}}{2}\right)^t e^{-2|v-u|}. \tag{A19}$$

Consider $u \le v$ and set

$$g(u) = \frac{1+e^{-2(v-u)}}{2}\; g'(u) = e^{-2(v-u)} f(x) = x^t; \tag{A20}$$

then

$$D(u,v)=\frac{1}{4}f(g(u))g'(u). \tag{A21}$$

With integration by substitution we get

$$\begin{aligned}\int_0^v D(u,v)\,du &= \frac{1}{4}\int_{g(0)}^{g(v)} f(x)\ \ dx\\ &= \frac{1}{4}\frac{x^{t+1}}{(t+1)}\Bigg|_{g(0)}^{g(v)}\\ &= \frac{1}{4(t+1)}\left[1-\left(\frac{1+e^{-2v}}{2}\right)^{t+1}\right].\end{aligned} \tag{A22}$$

In analogy we get for $u > v$

$$\int_v^{l_i} D(u,v)\,du = -\frac{1}{4(t+1)}\left[\left(\frac{1+e^{-2(l_i-v)}}{2}\right)^{t+1}-1\right] \tag{A23}$$

and therefrom for a fixed v

$$\int_0^{l_i} D(u,v)\,du = \frac{1}{4(t+1)}\left[2-\left(\frac{1+e^{-2v}}{2}\right)^{t+1}-\left(\frac{1+e^{-2(l_i-v)}}{2}\right)^{t+1}\right]. \tag{A24}$$

We have

$$\begin{aligned}\int_0^{l_i}\left(\frac{1+e^{-2v}}{2}\right)^{t+1}dv &= \frac{1}{2^{t+1}}\int_0^{l_i}(1+e^{-2v})^{t+1}dv\\ &= \frac{l_i}{2^{t+1}}+\frac{1}{2^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\int_0^{l_i}e^{-2vn}dv\\ &= \frac{l_i}{2^{t+1}}+\frac{1}{2^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{2n}(1-e^{-2l_in}).\end{aligned} \tag{A25}$$

For symmetry reasons

$$\int_0^{l_i}\left(\frac{1+e^{-2v}}{2}\right)^{t+1}dv = \int_0^{l_i}\left(\frac{1+e^{-2(l_i-v)}}{2}\right)^{t+1}dv \tag{A26}$$

and therefrom we get

$$\begin{aligned}\mathrm{Var}(Z_i) &= \frac{1}{l_i^2}\int_0^{l_i}\int_0^{l_i}D(u,v)\,du\,dv\\ &= \frac{1}{4l_i^2(t+1)}\left[l_i\left(2-\frac{1}{2^t}\right)-\frac{1}{2^{t+1}}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{n}(1-e^{-2l_in})\right].\end{aligned} \tag{A27}$$

$(F_2)^t$-SSD lines:

Using the definition of $D(u, v)$ from Table 1 and Equation A11 we get for a fixed value of v

$$\begin{aligned}\int_0^{l_i}D(u,v)\,du &= \int_0^{l_i}\frac{1}{2^{t+2}}\sum_{n=1}^{t+1}\binom{t}{n-1}\frac{e^{-2|v-u|n}}{2-e^{-2|v-u|}}du\\ &= \frac{1}{2^{t+2}}\sum_{n=1}^{t+1}\binom{t}{n-1}\left[2^{n-2}(\xi_1+\xi_2)+\sum_{k=1}^{n-1}\xi_3(\xi_4+\xi_5-2)\right],\end{aligned} \tag{A28}$$

where ξ_1, ξ_2, ξ_3, ξ_4, and ξ_5 are defined in Equation A13, and therefrom

$$\begin{aligned}\mathrm{Var}(Z_i) &= \frac{1}{l_i^2}\int_0^{l_i}\int_0^{l_i} D(u,v)\,du\,dv \\ &= \frac{1}{l_i^2 2^{t+2}}\sum_{n=1}^{t+1}\binom{t}{n-1}\left[2^{n-1}\xi_6 + 2\sum_{k=1}^{n-1}\xi_3(\xi_7 - l_i)\right].\end{aligned} \tag{A29}$$