DOI: 10.1534/genetics.106.057273

# Marker-Based Prediction 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 February 17, 2006
Accepted for publication July 25, 2006

ABSTRACT

Molecular markers can be employed to predict the parental genome contribution to inbred lines. The proportion $\alpha$ of alleles originating from parent $P_1$ at markers polymorphic between the parental lines $P_1$ and $P_2$ is commonly used as a predictor for the genome contribution of parent $P_1$ to an offspring line. Our objectives were to develop a new marker-based predictor $\xi$ for the parental genome contribution, which takes into account not only the alleles at marker loci but also their map distance, and to compare the prediction precision of $\xi$ with that of alternative methods. We derived formulas for $\xi$ for inbreds derived from biparental crosses ($F_1$ and backcrosses) with the single-seed descent or double-haploid method and presented an extension $\xi^*$ possessing statistical optimum properties. In a simulation study, $\alpha$ showed a systematic overestimation of large parental genome contribution that was not observed for $\xi$. The mean squared prediction error of $\xi$ was at least 50% smaller than that of $\alpha$ for linkage maps with unequal distances between adjacent markers. A data set from a study on plant variety protection in maize was used to illustrate the application of $\xi$. We conclude that $\xi$ provides substantially greater prediction precision than the commonly used predictor $\alpha$ in a broad range of applications in genetics and breeding.

GENETIC fingerprinting of inbred lines and their crossing parents with molecular markers provides a means to assess the parental origin of the genome of a line. It is carried out routinely in basic genetic research and applied breeding programs. Applications include, for example, the prediction of the donor genome proportion in inbred lines derived from backcross individuals of a gene introgression program or in near-isogenic lines of an introgression library developed either for fine mapping of QTL or for identification of favorable chromosome segments in genetic resources. In plant variety protection, prediction of the parental genome contribution is employed to decide whether or not a line is derived essentially from a progenitor line. In estimation of the breeding value of a line using phenotypic information from its crossing parents, marker-based prediction of the parental genome contribution can replace the assumption that each parent of a biparental cross contributes one-half to the genome of an offspring line.

In these applications, the proportion of marker alleles that are identical with the alleles of a parental line is commonly used to predict the contribution of the parental line to the genome of the derived inbred line (*cf.* Bernardo *et al.* 2000; Heckenberger *et al.* 2005b). The major shortcoming of this unweighted prediction is that neither linkage between markers nor the stochastic dependence between the parental origin of the marker alleles and the parental origin of the adjacent genomic regions is taken into account.

In the context of recurrent backcrossing, Visscher (1996) suggested to predict the contribution of the donor parent to the genome of a backcross individual by assigning different weights to the markers. He treated prediction of the parental genome contribution based on linked markers analogously to prediction of the breeding value of an individual based on different sources of phenotypic information. Extending the previous work of Hill (1993), he applied selection index theory (Hazel 1943) to derive weights depending on the recombination frequency between markers. However, for inbred lines no advanced theory has been elaborated for molecular marker-based prediction of the parental genome contribution.

We focused on inbred lines developed from biparental crosses ($F_1$ or backcrosses) with the single-seed descent or double-haploid method. The objectives of our research were to (1) develop a new marker-based predictor $\xi$ for the parental genome contribution, which takes into account not only the alleles at marker loci but also their map distance, (2) present an extension $\xi^*$, which possesses statistical optimum properties, and (3) compare the prediction precision of $\xi$ with that of alternative methods. Furthermore, various examples for applications of the predictor $\xi$ in genetics and breeding are discussed.

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

## THEORY

**Outline of the prediction approach:** The parental origin of the genome in a derived inbred line can be traced with molecular markers, which are polymorphic in the parental lines $P_1$ and $P_2$. The markers can be regarded as a sample of all loci in the genome and, therefore, the parental genome contribution to marker loci can be used as a predictor for the parental genome contribution to the entire genome. However, typically marker maps are not equally spaced and the different lengths of marker intervals are ignored in such a prediction. We suggest a predictor $\xi$, which takes into account not only the genotype at the marker loci, but also the map distance between adjacent markers. The principle of $\xi$ is to determine for each locus in the genome the conditional expectation that it carries the allele of parent $P_1$ under the condition of the observed genotype at flanking markers. The genome is subdivided into nonoverlapping chromosome intervals, of which the borders are defined by the markers, and the conditional expectations are integrated along the chromosome intervals. This yields a prediction of the parental genome contribution of $P_1$ to each chromosome interval. Subsequently, the predictions for the chromosome intervals are weighted with the interval lengths and averaged to obtain a predictor for the genome contribution of parent $P_1$ to the entire genome.

**Notation and assumptions:** Map positions, measuring the distance of a locus from a telomere in morgan units, are denoted by $x$. Indicator variables $G$ take the value 1 if the allele at the corresponding locus originates from parent $P_1$ and 0 otherwise. Realizations of $G$ are denoted with $g$. We subdivide the genome into $n$ nonoverlapping chromosome intervals. A chromosome interval $i$ is delimited by either (1) two markers with map positions $x_{a_i} < x_{b_i}$ or (2) a marker and a telomere. For case 2 we assume without loss of generality that the telomere has map position 0 and the distance between the marker and the telomere is $x_{a_i}$. The length of a chromosome interval is $d_i = x_{b_i} - x_{a_i}$ (case 1) and $d_i = x_{a_i}$ (case 2). The genome length is $l = \sum_i d_i$.

We assume that the offspring are completely homozygous lines, derived without selection from a biparental cross or a backcross of completely homozygous parents $P_1$ and $P_2$ that are polymorphic in at least one marker per chromosome. We further assume absence of interference (Stam 1979) in crossover formation such that the recombination frequency $r_{uv}$ between two loci with map positions $x_u \leq x_v$ is calculated by Haldane's (1919) mapping function:

$$r_{uv} = (1 - e^{-2(x_v - x_u)})/2. \tag{1}$$

**The predictor $\xi$:** The predictor $\xi$ of the genome contribution of parent $P_1$ to a derived line is defined as

$$\xi = \sum_{i=1}^{n} \frac{d_i}{l} \xi_i, \tag{2}$$

where $\xi_i$ is the prediction of the genome contribution of parent $P_1$ to the $i$th chromosome interval.

We consider at first a finite number $w$ of loci equidistantly distributed at positions $x_1, \ldots, x_w$ on a chromosome interval delimited by markers at positions $x_{a_i}$ and $x_{b_i}$. We then have

$$\xi_i = \frac{1}{w} \sum_{s=1}^{w} E(G_s \mid g_{a_i}, g_{b_i}), \tag{3}$$

where $E(G_s \mid g_{a_i}, g_{b_i})$ is the conditional expectation that the locus at map position $x_s$ carries the allele of parent $P_1$ under the condition that the genotypes $g_{a_i}$ and $g_{b_i}$ were observed at the two flanking markers with map positions $x_{a_i}$ and $x_{b_i}$. Following the principle used by Franklin (1977) and Hill (1993), Equation 3 can be extended to an infinite number of loci at positions $x_s$:

$$\xi_i = \frac{1}{d_i} \int_{x_{a_i}}^{x_{b_i}} E(G_s \mid g_{a_i}, g_{b_i}) dx_s. \tag{4}$$

For telomere chromosome intervals we have in analogy

$$\xi_i = \frac{1}{d_i} \int_{0}^{x_{a_i}} E(G_s \mid g_{a_i}) dx_s. \tag{5}$$

The conditional expectation of $G_s$ is (omitting the subscript $i$ for the chromosome interval)

$$\begin{aligned} E(G_s \mid g_a, g_b) &= P(G_s = 1 \mid g_a, g_b) \\ &= \frac{P(G_a = g_a, G_s = 1, G_b = g_b)}{P(G_a = g_a, G_b = g_b)} \end{aligned} \tag{6}$$

for $x_s$ in chromosome intervals flanked by two markers and

$$E(G_s \mid g_a) = \frac{P(G_a = g_a, G_s = 1)}{P(G_a = g_a)} \tag{7}$$

for $x_s$ in chromosome intervals flanked by a marker and the telomere.

**Mating systems:** We express the one-, two-, and three-locus genotype frequencies required for Equations 6 and 7 in terms of

$$p = P(G_u = 1) \quad \text{and} \quad q_{uv} = P(G_v = 1 \mid G_u = 1), \tag{8}$$

where $x_u, x_v \in \{x_a, x_b, x_s\}$. The values of $p$ and $q_{uv}$ depend on the mating system used for deriving the inbred line and the map distance between the markers at positions $x_u$ and $x_v$. In this study, we consider four mating systems: (1) $(F_2)_t$-single-seed descent (SSD) lines are developed by $t$ ($t \geq 0$) generations of random mating of an $F_2$ population and subsequent application of the single-seed descent method for line development; (2) $(F_1)_t$-double-haploid (DH) lines are developed by $t$ ($t \geq 0$) generations of random mating of an $F_1$ cross and subsequent inbred line development with double haploids; and (3) backcross $(BC)_t$-SSD and (4) $BC_t$-DH lines are developed from an $F_1$ cross backcrossed $t$ ($t \geq 1$) times to parent $P_2$, with subsequent line development by the

**TABLE 1**

**Definition of parameters $p$ and $q_{uv}$ for four mating systems**

| Mating system | $p$ | $q_{uv}$ |
|---|---|---|
| $(F_1)^t$-DH | $\frac{1}{2}$ | $\frac{1}{2}+\frac{1-2r_{uv}}{2}(1-r_{uv})^t$ |
| $(F_2)^t$-SSD | $\frac{1}{2}$ | $\frac{1}{2}+\frac{1-2r_{uv}}{2+4r_{uv}}(1-r_{uv})^t$ |
| $BC_t$-DH | $(1/2)^{t+1}$ | $(1-r_{uv})^{t+1}$ |
| $BC_t$-SSD | $(1/2)^{t+1}$ | $\frac{(1-r_{uv})^t}{1+2r_{uv}}$ |

single-seed descent or double-haploid method, respectively. Expressions for $p$ and $q_{uv}$ under these mating systems are given in Table 1, and the corresponding derivations are presented in the APPENDIX.

**Genotype frequencies:** For the derivations a shorthand notation is used. We omit the names of random variables in definitions of multilocus genotype frequencies and use only the value of the realizations. For example, $P(G_a = 1, G_s = 1)$ is abbreviated as $P(11)$, and $P(G_a = 1, G_s = 1, G_b = 1)$ as $P(111)$. For the derivation of three-locus genotype frequencies, two-locus genotype frequencies referring to subsets of the three loci are required. In this case, the realization of the third (not considered locus) is denoted with a "–," *e.g.*, $P(G_a = 1, G_b = 1)$ is abbreviated as P(1–1).

The single-locus genotype frequencies follow directly from the definition of $p$ in Equation 8:

$$\begin{aligned} P(1) &= p \\ P(0) &= 1 - p. \end{aligned} \tag{9}$$

The two-locus genotype frequencies for two loci at map positions $x_u, x_v \in \{x_a, x_b, x_s\}$ can be written as

$$\begin{aligned} P(11) &= pq_{uv} \\ P(01) &= p(1 - q_{uv}) \\ P(10) &= p(1 - q_{uv}) \\ P(00) &= 1 - 2p + pq_{uv}. \end{aligned} \tag{10}$$

For deriving the three-locus genotype frequencies with respect to three loci at map positions $x_a < x_s < x_b$ for $(F_1)^t$-DH and $(F_2)^t$-SSD lines, we follow an approach outlined by HALDANE and WADDINGTON (1931) and recently developed by BROMAN (2005) for $F_2$-SSD lines (named two-way RILs in his article). We have

$$\begin{aligned} P(111) + P(011) &= P(\text{–}11) = pq_{sb} \\ P(111) + P(101) &= P(1\text{–}1) = pq_{ab} \\ P(111) + P(110) &= P(11\text{–}) = pq_{as} \\ P(010) + P(110) &= P(\text{–}10) = p(1 - q_{sb}) \end{aligned} \tag{11}$$

and because of symmetry

$$P(101) = P(010). \tag{12}$$

Solving this system of linear equations and using $p = \frac{1}{2}$ yields

$$\begin{aligned} P(111) &= (q_{ab} + q_{as} + q_{sb} - 1)/4 \\ P(011) &= (1 + q_{sb} - q_{as} - q_{ab})/4 \\ P(110) &= (1 + q_{as} - q_{ab} - q_{sb})/4 \\ P(010) &= (1 + q_{ab} - q_{sb} - q_{as})/4. \end{aligned} \tag{13}$$

For $BC_t$-DH and $BC_t$-SSD lines we employ the system of equations

$$\begin{aligned} P(111) + P(011) &= P(\text{–}11) = pq_{sb} \\ P(011) + P(010) &= P(01\text{–}) = p(1 - q_{as}) \\ P(010) + P(110) &= P(\text{–}10) = p(1 - q_{sb}) \\ P(111) + P(110) &= P(11\text{–}) = pq_{as}. \end{aligned} \tag{14}$$

For $BC_t$-DH lines we use

$$P(111) = [(1 - r_{as})(1 - r_{sb})/2]^{t+1} = pq_{as}q_{sb} \tag{15}$$

to solve the system of equations in Equation 14 and obtain

$$\begin{aligned} P(111) &= pq_{as}q_{sb} \\ P(011) &= p(q_{sb} - q_{as}q_{sb}) \\ P(110) &= p(q_{as} - q_{as}q_{sb}) \\ P(010) &= p(1 - q_{as} - q_{sb} + q_{as}q_{sb}). \end{aligned} \tag{16}$$

For $BC_t$-SSD lines we have

$$P(111) = p^b q_{as}^b q_{sb}^b (q_{ab}^s + q_{as}^s + q_{sb}^s - 1)/4, \tag{17}$$

where the superscript $b$ refers to the parameters of a $BC_t$ individual, which can be obtained from Equation 16 by replacing $t + 1$ with $t$, and superscript $s$ refers to the parameters of $F_2$-SSD lines. Solving Equation 14 yields

$$\begin{aligned} P(011) &= pq_{sb} - P(111) \\ P(110) &= pq_{as} - P(111) \\ P(010) &= P(111) + p(1 - q_{as} - q_{sb}). \end{aligned} \tag{18}$$

Note that (1) the genotype frequencies obtained with Equation 13 after inserting $p$ and $q_{ab}$ for $F_2$-SSD lines are identical to those of BROMAN (2005), and (2) the genotype frequencies for $BC_t$-DH lines are identical with those obtained with the formulas of VISSCHER and THOMPSON (1995) for $BC_{t+1}$ individuals.

**Conditional expectation ξ*:** The predictor ξ can be extended by replacing $E(G_s \mid g_a, g_b)$ (Equation 4) and $E(G_s \mid g_a)$ (Equation 5) with

$$E(G_s \mid \mathbf{g}_i) = \frac{P(G_s = 1, \mathbf{g}_i)}{P(\mathbf{g}_i)}, \tag{19}$$

where $\mathbf{g}_i$ is a vector consisting of the marker genotype of the markers on the $i$th chromosome. The resulting predictor ξ* is the conditional expectation of the parental genome contribution to an inbred line under the condition of the observed marker genotype. For

calculation of the multilocus genotype frequencies in Equation 19, the recursion equations of Hospital *et al.* (1996) can be employed. Further, the closed-form equations derived by Visscher and Thompson (1995) for $BC_t$ individuals can be applied to $BC_{t-1}$-DH lines.

$$V_{uv} = \left(\tfrac{1}{2}\right)^{t+1}\left\{[1 - r_{uv}]^{t+1} - \left(\tfrac{1}{2}\right)^{t+1}\right\}$$
$$y_u = \frac{1}{4^{t+1}l}\sum_{n=1}^{t+1}\binom{t+1}{n}\frac{1}{2n}[2 - e^{-2nx_u} - e^{-2n(l-x_u)}]. \quad (23)$$

## DISCUSSION

**Other predictors for the parental genome contribution:** A commonly used predictor (*cf.* Bernardo *et al.* 2000; Heckenberger *et al.* 2005b) of the genome contribution of parent $P_1$ to an inbred line is the proportion of marker alleles from $P_1$ in the set of polymorphic markers between $P_1$ and $P_2$,

$$\alpha = \frac{1}{m}\sum_{j=1}^{m} g_j, \quad (20)$$

where $m$ is the number of markers and $g_j$ refers to the genotypes at the marker loci. Major shortcomings of the unweighted predictor α are that (i) the correlation between markers due to linkage and (ii) the stochastic dependence between the markers and the adjacent genomic regions are not taken into account. The advantage of α is that no prior information about the mating system used to develop the line is required.

No previous studies exist about more efficient predictors for the parental genome contribution to inbred lines. However, Visscher (1996) developed an approach for predicting the proportion of the genome originating from the donor parent in backcross individuals, borrowing ideas from selection index theory (Hazel 1943),

$$\beta = \sum_{i=1}^{c}\frac{l_i}{l}(\mathbf{V}_i^{-1}\mathbf{y}_i)'\mathbf{g}_i, \quad (21)$$

where $c$ is the number of chromosomes, $l_i$ is the length of the $i$th chromosome, $\mathbf{g}_i$ is a vector consisting of the marker genotype of the markers on the $i$th chromosome, $\mathbf{V}_i$ is the covariance matrix of $\mathbf{g}_i$, and $\mathbf{y}_i$ is a vector consisting of the covariances between the donor genome at the markers and the donor genome on the carrier chromosome of the markers.

Visscher's (1996) approach can be extended to inbred lines derived from arbitrary mating systems by defining for each chromosome (we omit the index for the chromosome)

$$V_{uv} = D_{uv}$$
$$y_u = \frac{1}{l}\int_0^l D_{us}\, dx_s, \quad (22)$$

where $D_{us} = (q_{us} - p)p$ is the expected gametic disequilibrium between loci at map positions $x_u$ and $x_s$ under the considered mating system, with expressions for $p$ and $q_{us}$ given in Table 1. For example, for $BC_t$-DH lines we have

In comparison with the predictor α, the weighted predictor β has the advantage that the markers contribute with different weights to β, depending on their linkage; *i.e.*, β takes into account the correlation between the markers on a chromosome. However, it ignores the stochastic dependence between the markers and the adjacent genomic regions on a chromosome.

In contrast to α and β, the predictor ξ takes into account both the correlation between markers and the stochastic dependence between markers and the adjacent genomic regions. The former is considered by weighting $\xi_i$ with the distance $d_i$ between adjacent markers (Equation 2) and the latter by the integration of $E(G_s|\, g_a, g_b)$ along the chromosome (Equations 4 and 5).

**Conditional expectation ξ*:** The predictor ξ* is unbiased, and this can be shown using $E(G_u) = p$ and $E(E[Z|\,\mathbf{G}]) = E(Z)$ (Shao 1999, p. 33, Proposition 1.12.iv), where the random variable $Z$ denotes the parental genome contribution to an inbred line and the random vector $\mathbf{G}$ its multilocus marker genotype. From $\xi^* = E(Z|\,\mathbf{g})$ it follows that ξ* is also unbiased in the sample space $\Omega_{\mathbf{g}}$, which comprises the parental genome contribution to all possible inbred lines having a certain marker genotype $\mathbf{g}$. From an applied point of view, this means that ξ* is neither systematically overestimating nor underestimating the parental genome contribution for any given marker genotype $\mathbf{g}$. It can be further shown that the conditional expectation of a random variable has minimum variance among all unbiased predictors (Shao 1999, p. 33, Equation 1.40). In consequence, the conditional expectation ξ* can be regarded as an optimum predictor of the parental genome contribution.

For the $F_1$-DH mating system (which is, for example, often employed for the development of inbred lines in hybrid maize breeding programs) the predictors ξ and ξ* are identical under the assumption of no interference in crossover formation. For other mating systems, calculation of ξ* requires calculation of multilocus genotype frequencies. In contrast to the relatively simple calculations of two- and three-locus genotype frequencies, which can easily be carried out with standard software such as R (Ihaka and Gentleman 1996), calculation of multilocus genotype frequencies requires extensive programming (*cf.* Servin *et al.* 2002).

We compared ξ* and ξ for several special cases and found only small numerical differences in the results. We therefore conclude that the simple calculations required for ξ may outweigh the theoretical optimum properties of ξ* in many practical applications. It could be the subject of further research to investigate whether

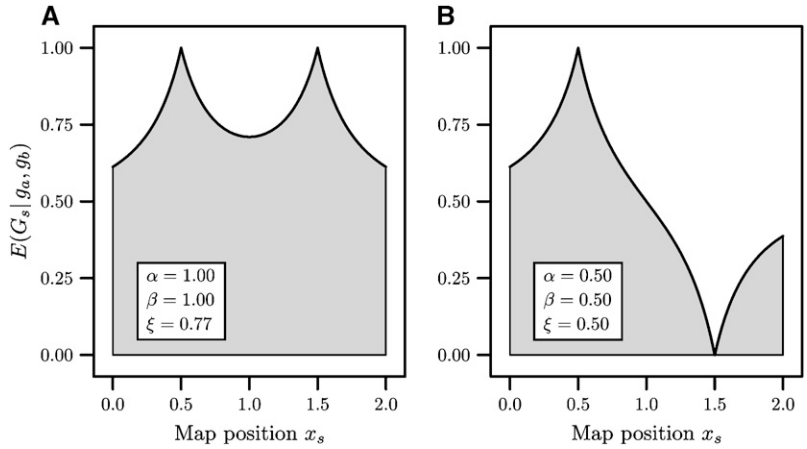


FIGURE 1.—Predictions α, β, and ξ for a 2-M chromosome of an $F_2$-SSD line on which two markers are located 0.5 and 1.5 M from the telomere. (A) Both markers carry the allele of parent $P_1$. (B) The first marker carries the allele of parent $P_1$ and the second marker that of parent $P_2$. The solid line denotes the conditional expectation $E(G_s \mid g_a, g_b)$ that the locus at position $x_s$ carries the allele of parent $P_1$.

the substantially greater programming and computational effort, which is required for calculation of ξ* in the general case, results in a significant improvement of the prediction accuracy compared with ξ.

**Systematic prediction error of α and β:** Consider the example of a 2-M chromosome of an $F_2$-SSD line, on which two markers located 0.5 and 1.5 M from the telomere carry the allele of parent $P_1$ (Figure 1A). Loci in the genome region between the two markers are up to 0.5 M distant from the nearest adjacent marker. Owing to the large recombination frequency between distant loci, the markers predict only poorly the genotype at these loci. The predictors α and β do not take this low correlation into account and the genotype of all loci on the chromosome is predicted to be the same as the genotype of the markers: $\alpha = \beta = 1$. We now focus on a large number of chromosomes carrying the alleles of parent $P_1$ at the two markers. Only ~70% of these chromosomes carry the allele of parent $P_1$ at a locus in the center between the markers [$E(G_s \mid g_a, g_b) \approx 0.7$, see Figure 1]. Hence, with respect to all possible chromosomes having the considered marker genotype, the predictors α and β are systematically overestimating the genome proportion originating from parent $P_1$.

For symmetry reasons, the genome contribution of parent $P_1$ to chromosomes carrying at both markers the allele of parent $P_2$ is systematically underestimated by the predictors α and β. In contrast, α and β show no systematic prediction error for chromosomes on which recombination occurred (Figure 1B). Individuals having no recombination between two markers with map distance 1 M occur in an $F_2$-SSD population with probability 0.54 (*cf.* HALDANE and WADDINGTON 1931). Hence, a considerable systematic prediction error of α and β is observed for more than half of the chromosomes of an $F_2$-SSD population.

**Systematic prediction error for the entire genome:** The above theoretical example illustrates that in principle systematic prediction errors can occur when employing predictors α and β. To investigate whether the extent of such systematic prediction errors is of relevance in practical applications, we conducted a simulation study with Plabsoft (MAURER *et al.* 2004). Simulated data were used because they provide the "true" parental genome contributions $z$ of parent $P_1$ as well as the predictions $\hat{z} = \vartheta$ ($\vartheta \in \{\alpha, \xi\}$) for each simulated inbred line. This allows us to generate a large number of inbred lines and determine the prediction errors $e = \vartheta - z$.

For the simulation we employed a model of the maize genome based on the study of HECKENBERGER *et al.* (2005a). It consists of 10 chromosomes of length 1.70, 1.30, 1.06, 1.48, 1.28, 1.15, 1.14, 1.21, 0.99, and 0.91 M and 100 SSR markers, which were chosen for good coverage of the entire genome. We simulated 1000 $F_2$-SSD lines, for which we determined the prediction errors of α and ξ. The correlation $\rho_{\alpha,e} = 0.36$ between the predicted genome proportions α and the corresponding prediction errors $e$ was highly significant (type I error rate 0.001), whereas no significant correlation was observed for the predictor ξ (Figure 2).

We conclude that the extent of systematic overestimation of large parental genome contributions and systematic underestimation of small ones by the predictor α can cause serious problems with linkages maps commonly used in practical applications.

**Precision of prediction:** To assess systematically the precision of prediction of α, β, and ξ in the four mating systems under consideration, we conducted a simulation study. We employed a model of the maize genome with 10 chromosomes of length 1.6 M. Twenty to 200 markers were assumed to be (a) randomly distributed and (b) equally spaced in the genome. In practice, the marker distribution ranges between these two extremes, which can be regarded as a "worst-case" scenario (a) and a "best-case" scenario (b). For each combination of marker density and spacing we simulated 500 $F_2$-SSD, $F_1$-DH, $BC_1$-SSD, and $BC_1$-DH populations of size 100. (For random spacing of markers, different maps were used for each of the 500 populations.) The correlations $\rho_{\alpha,e}$, $\rho_{\beta,e}$, and $\rho_{\xi,e}$ between predicted values and prediction errors as well as the mean squared prediction errors

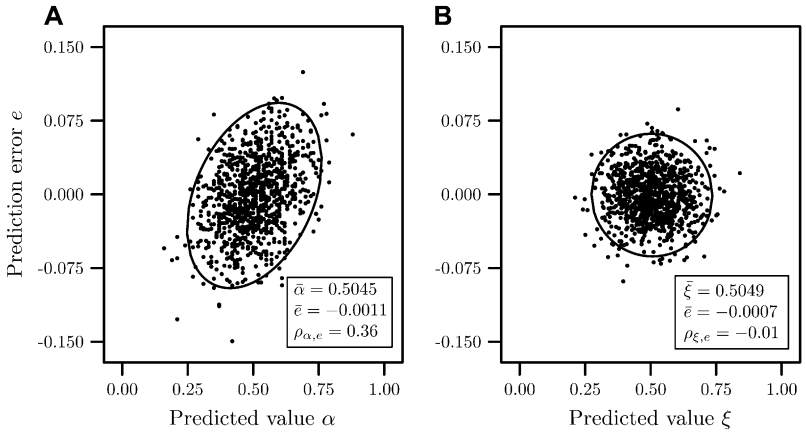


FIGURE 2.—Prediction error $e$ of $\alpha$ and $\xi$ in a simulated maize data set. $\bar{\alpha}$, $\bar{\xi}$, and $\bar{e}$ are mean values, and $\rho_{\alpha,e}$ and $\rho_{\xi,e}$ are the correlations between the predicted value and prediction error.

$$M_\vartheta = \frac{1}{n}\sum_n [\vartheta - z]^2 \quad \vartheta \in \{\alpha, \beta, \xi\} \tag{24}$$

were determined for each simulated population. The results were then averaged over the 500 populations.

The correlations $\rho_{\alpha,e}$, $\rho_{\beta,e}$ were highly significant (type I error rate 0.001) for all combinations of the investigated parameters, while $\rho_{\xi,e}$ was not significantly different from zero for any combination. The largest correlations, amounting to 0.75, were observed for predictor $\alpha$ with sparse maps and random marker positions (Table 2). However, even with 200 equally spaced markers $\rho_{\alpha,e} \geq 0.25$ and $\rho_{\beta,e} \geq 0.15$. The mean squared prediction error $M_\xi$ was at least 50% smaller than $M_\alpha$ for randomly distributed markers, and $M_\beta$ ranged in between and approached the values of $M_\xi$ for ≥100 markers. With equally spaced maps, the differences between $M_\alpha$, $M_\beta$, and $M_\xi$ were negligible for >80 markers.

**TABLE 2**

**Correlations $\rho_{\alpha,e}$ and $\rho_{\xi,e}$ and mean squared prediction errors $M_\alpha$, $M_\beta$, and $M_\xi$ for simulated maize lines depending on marker density and spacing for four mating systems**

| | No. of randomly distributed markers | | | | | | No. of equally spaced markers | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20 | 40 | 60 | 80 | 100 | 200 | 20 | 40 | 60 | 80 | 100 | 200 |
| | | | | | | $F_2$-SSD lines | | | | | | |
| $\rho_{\alpha,e}$ | 0.75 | 0.64 | 0.58 | 0.51 | 0.48 | 0.36 | 0.69 | 0.53 | 0.43 | 0.37 | 0.34 | 0.26 |
| $\rho_{\beta,e}$ | 0.73 | 0.58 | 0.50 | 0.42 | 0.38 | 0.23 | 0.69 | 0.52 | 0.41 | 0.34 | 0.30 | 0.18 |
| $M_\alpha$ | 0.0105 | 0.0059 | 0.0041 | 0.0030 | 0.0024 | 0.0012 | 0.0054 | 0.0017 | 0.0009 | 0.0005 | 0.0003 | 0.0001 |
| $M_\beta$ | 0.0091 | 0.0042 | 0.0025 | 0.0016 | 0.0011 | 0.0004 | 0.0054 | 0.0017 | 0.0008 | 0.0005 | 0.0003 | 0.0001 |
| $M_\xi$ | 0.0042 | 0.0027 | 0.0019 | 0.0013 | 0.0010 | 0.0003 | 0.0028 | 0.0013 | 0.0007 | 0.0004 | 0.0003 | 0.0001 |
| | | | | | | $F_1$-DH lines | | | | | | |
| $\rho_{\alpha,e}$ | 0.68 | 0.59 | 0.51 | 0.47 | 0.42 | 0.32 | 0.61 | 0.46 | 0.40 | 0.35 | 0.34 | 0.28 |
| $\rho_{\beta,e}$ | 0.64 | 0.51 | 0.41 | 0.35 | 0.30 | 0.17 | 0.61 | 0.44 | 0.35 | 0.29 | 0.25 | 0.15 |
| $M_\alpha$ | 0.0099 | 0.0059 | 0.0040 | 0.0031 | 0.0024 | 0.0012 | 0.0038 | 0.0011 | 0.0005 | 0.0003 | 0.0002 | 0.0001 |
| $M_\beta$ | 0.0074 | 0.0035 | 0.0019 | 0.0012 | 0.0008 | 0.0002 | 0.0038 | 0.0011 | 0.0005 | 0.0003 | 0.0002 | 0.0000 |
| $M_\xi$ | 0.0043 | 0.0025 | 0.0016 | 0.0011 | 0.0007 | 0.0002 | 0.0024 | 0.0008 | 0.0004 | 0.0002 | 0.0002 | 0.0000 |
| | | | | | | $BC_1$-SSD lines | | | | | | |
| $\rho_{\alpha,e}$ | 0.75 | 0.64 | 0.57 | 0.52 | 0.47 | 0.36 | 0.69 | 0.52 | 0.43 | 0.36 | 0.34 | 0.25 |
| $\rho_{\beta,e}$ | 0.73 | 0.59 | 0.50 | 0.43 | 0.37 | 0.23 | 0.69 | 0.51 | 0.41 | 0.33 | 0.30 | 0.17 |
| $M_\alpha$ | 0.0078 | 0.0044 | 0.0030 | 0.0023 | 0.0018 | 0.0009 | 0.0041 | 0.0013 | 0.0006 | 0.0004 | 0.0002 | 0.0001 |
| $M_\beta$ | 0.0067 | 0.0031 | 0.0018 | 0.0012 | 0.0008 | 0.0003 | 0.0041 | 0.0013 | 0.0006 | 0.0004 | 0.0002 | 0.0001 |
| $M_\xi$ | 0.0031 | 0.0020 | 0.0014 | 0.0010 | 0.0007 | 0.0002 | 0.0021 | 0.0010 | 0.0005 | 0.0003 | 0.0002 | 0.0001 |
| | | | | | | $BC_1$-DH lines | | | | | | |
| $\rho_{\alpha,e}$ | 0.72 | 0.61 | 0.53 | 0.49 | 0.44 | 0.34 | 0.64 | 0.48 | 0.41 | 0.36 | 0.33 | 0.26 |
| $\rho_{\beta,e}$ | 0.69 | 0.54 | 0.44 | 0.38 | 0.32 | 0.20 | 0.64 | 0.47 | 0.38 | 0.31 | 0.26 | 0.16 |
| $M_\alpha$ | 0.0078 | 0.0044 | 0.0029 | 0.0022 | 0.0018 | 0.0009 | 0.0034 | 0.0010 | 0.0005 | 0.0003 | 0.0002 | 0.0000 |
| $M_\beta$ | 0.0064 | 0.0028 | 0.0015 | 0.0010 | 0.0007 | 0.0002 | 0.0034 | 0.0010 | 0.0004 | 0.0003 | 0.0002 | 0.0000 |
| $M_\xi$ | 0.0033 | 0.0020 | 0.0012 | 0.0009 | 0.0006 | 0.0002 | 0.0020 | 0.0008 | 0.0004 | 0.0002 | 0.0002 | 0.0000 |

**TABLE 3**

**Predictors α, β, and ξ for the experimental data from maize**

| Line no. | *m* | α (%) | β (%) | ξ (%) | α − ξ (%) |
|---|---|---|---|---|---|
| 1 | 64 | 95.3[a] | 91.3[a] | 84.2[a] | 11.1 |
| 2 | 62 | 85.5[a] | 83.6[a] | 78.2[a] | 7.3 |
| 3 | 64 | 82.8[a] | 81.1[a] | 77.2[a] | 5.6 |
| 4 | 48 | 68.8[a] | 67.4[a] | 65.8 | 3.0 |
| 5 | 48 | 68.8[a] | 68.7[a] | 65.4 | 3.4 |
| 6 | 48 | 68.8[a] | 68.7[a] | 65.4 | 3.4 |
| 7 | 58 | 58.6 | 64.9 | 63.7 | −5.0 |
| 8 | 55 | 69.1[a] | 63.4 | 62.5 | 6.6 |
| 9 | 51 | 62.7 | 64.0 | 61.6 | 1.1 |
| 10 | 48 | 58.3 | 63.1 | 60.8 | −2.4 |
| 11 | 44 | 56.8 | 62.0 | 59.6 | −2.8 |
| 12 | 38 | 57.9 | 57.4 | 59.5 | −1.6 |
| | | | | | |
| 45 | 44 | 38.6 | 41.4 | 44.6 | −5.9 |
| 46 | 54 | 40.7 | 42.5 | 44.5 | −3.8 |
| 47 | 50 | 48.0 | 45.6 | 44.0 | 4.0 |
| 48 | 48 | 41.7 | 40.3 | 42.4 | −0.7 |
| 49 | 39 | 35.9 | 39.1 | 41.9 | −6.0 |
| 50 | 41 | 39.0 | 36.7 | 41.3 | −2.3 |
| 51 | 58 | 36.2 | 38.3 | 41.0 | −4.8 |
| 52 | 45 | 31.1 | 32.0 | 38.3 | −7.2 |
| 53 | 57 | 36.8 | 34.5 | 37.3 | −0.5 |
| 54 | 51 | 35.3 | 33.2 | 36.4 | −1.1 |
| 55 | 48 | 27.1 | 25.1 | 32.5 | −5.4 |
| 56 | 58 | 20.7 | 24.7 | 28.0 | −7.3 |

*m*, the number of markers polymorphic between the parental lines.

[a] Predictor is greater than the threshold value of 66.2.

We conclude that the superiority of ξ compared to α and β with respect to the mean squared prediction error reduces with increasing numbers of equally spaced markers. However, even for dense maps with equally spaced markers, the correlations $\rho_{\alpha,e}$ and $\rho_{\beta,e}$ between predicted value and prediction error indicate that systematic prediction errors of α and β are to be expected, with negative effects for practical applications.

**Application to experimental data:** Prediction of the parental genome contribution is illustrated with experimental data from a study on plant variety protection in maize (Heckenberger *et al.* 2005a). The genotype of 100 SSR markers was assessed at 56 $F_2$-SSD lines and their crossing parents. For each inbred line, markers not polymorphic between its crossing parents were discarded. This resulted in different marker sets used for the calculations in each line, with numbers of polymorphic markers *m* ranging between 38 and 67. From the genotype at the polymorphic markers, the predictors α, β, and ξ were calculated (Table 3 lists results for the 12 lines with the largest and smallest values of ξ).

The differences between the predictors α and ξ were mostly negative for small values of ξ and mostly positive for large values, reaching up to 11% (line 1). Comparing these values with simulation results for the same linkage map (Figure 2) leads to the conclusion that the differences observed for large and small parental genome contributions are partially caused by the systematic prediction error of α.

A method to detect essentially derived varieties is to compare a prediction of the parental genome contribution to an inbred line with a threshold value. Heckenberger *et al.* (2005b) suggested to use as thresholds the quantiles of the probability distribution of the parental genome contribution to inbred lines under an accepted breeding method. For the chromosome lengths underlying the study of Heckenberger *et al.* (2005a) we investigated this strategy and determined with a simulation the 0.95 quantile of the parental genome contribution to $F_2$-SSD lines as $t = 0.662$. When comparing the predictions of the parental genome contribution to an $F_2$-SSD population with this threshold value, then it is expected that 5% of the $F_2$-SSD lines are classified incorrectly as essentially derived varieties. In our experimental $F_2$-SSD population $\alpha > t$ for 7 lines (12.5% of the 56 lines) and $\beta > t$ for 6 lines (10.7%), but $\xi > t$ only for 3 lines (5.4%) (Table 3).

Consequently, the systematic overestimation of large parental genome contributions by α and β can result in a greater error rate of incorrectly classifying a line as essentially derived than is nominally associated with a chosen threshold value. However, due to the stochastic nature of meiosis, using α does not necessarily result in a greater error rate in every experimental population. This can be seen, for example, when comparing the lower tail of the distribution of the experimental data with $1 - t$.

Summarizing, ξ allows prediction of the parental genome contribution with a much higher precision than the unweighted predictor α commonly employed in practice, in particular when extreme values of the parental genome contribution are of interest. Thus, using ξ instead of α is clearly advantageous for obtaining reliable conclusions on the true parental genome contribution to an inbred line.

**Deviations from the assumptions:** As applies to most mathematical models of biological systems, the presented prediction method is not capable of capturing every detail of the underlying biological process, and the results should be interpreted with this in mind. Among the assumptions made in our derivations, the following seem of particular importance:

1. We assumed absence of interference in crossover formation, although it is well known that interference occurs (for a discussion on using noninterference models, see Frisch and Melchinger 2001).
2. We assumed known map positions of the markers. However, in practice, the linear order and map distances are estimated from mapping experiments with one or several segregating populations. Depending on the size and type of the mapping population(s), the estimated map may deviate from the true map due to sampling error or other causes.

3. We assumed absence of selection during backcrossing and inbred line development. If selection is carried out, the probability that a certain locus carries the allele of $P_1$ may differ from our derivations.

If, for a certain study, one or several of these assumptions do not hold true, the actual advantage in precision of $\xi$ compared with $\alpha$ and $\beta$ may be smaller than that under the idealized model, where all assumptions are fulfilled.

**Applications in genetics and breeding:** Being aware of the above limitations, the presented results demonstrate that the predictor $\xi$ provides a substantial improvement in the precision of predicting the parental genome contribution to inbred lines compared with the commonly used unweighted predictor $\alpha$. This improved precision can be important in a broad range of practical applications.

In inbred lines developed from backcross individuals of a gene introgression program, exact prediction of the parental contribution of the donor parent can help to assess the risk of negative phenotypic effects caused by the donor genome. The prediction of the parental genome contribution to inbred lines complements the prediction of the donor genome proportion in backcross individuals described by Frisch and Melchinger (2005). The combination of both approaches allows monitoring of the parental genome proportion from the first backcross generation until the converted inbred line is finally developed.

Introgression libraries of near-isogenic lines (Eshed and Zamir 1995) are increasingly developed in various crops, *e.g.*, for fine mapping of QTL or for identification of advantageous chromosome segments in exotic genetic resources or landraces (Tanksley and Nelson 1996). The presented approach can be used to predict the donor genome proportion in chromosome regions where a line of an introgression library carries the marker alleles of the recurrent parent. This can help to assess the risk that the observed phenotypic effect is not caused by the chromosome segment introgressed on purpose but by other donor chromosome segments not detected by the employed marker set.

In plant variety protection, exact prediction of the parental genome contribution to an inbred line is of crucial importance to draw conclusions whether or not the line (1) was developed with a generally accepted breeding method or (2) has a parental genome proportion below a generally accepted threshold. The predictor $\xi$ can be employed, for example, to estimate precisely the parental genome contribution of one parent, assuming a given mating system, which can then be compared with threshold values (Heckenberger *et al.* 2005b). In this context, it is of particular interest that $\alpha$ is overestimating systematically large parental genome contributions.

In quantitative genetic studies, the genome contribution of a parent to a biparental crossing progeny is usually assumed to be one-half. If $\xi$ is employed instead, this allows us to consider not only the expected relation between two lines on the basis of the mating system, but also the actual similarity at the level of the entire genome. A possible application is, for example, the best linear unbiased prediction of the breeding value of a line employing phenotypic information from its crossing parents.

Recurrent full-sib mating is used to generate homozygous strains in animals such as mice. Employing results of Haldane and Waddington (1931), parameters $p$ and $q_{uv}$ for recurrent full-sib mating can be derived analogously to the derivations for recurrent selfing in the appendix. Using this extension, the theory presented here can be used straightforwardly for applications in animal genetics.

Prediction of the parental genome proportion with $\xi$ can be interpreted as a "map-based genetic distance" between an inbred line and its crossing parent. It seems promising for further research to investigate whether the principle used in this study can be extended to provide map-based genetic distances for general pedigrees and/or heterozygous individuals.

We thank the anonymous reviewers for their comments and suggestions, which helped to improve the manuscript. In particular, we are greatly indebted to an anonymous reviewer for pointing out a major mistake in an earlier version of the manuscript.

## LITERATURE CITED

Bernardo, R., J. Romero-Severson, J. Ziegle, J. Hauser, L. Joe *et al.*, 2000 Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. Theor. Appl. Genet. **100:** 552–556.

Broman, K., 2005 The genomes of recombinant inbred lines. Genetics **169:** 1133–1146.

Cockerham, C. C., and B. S. Weir, 1973 Descent measures for two loci with some applications. Theor. Popul. Biol. **4:** 300–330.

Eshed, Y., and D. Zamir, 1995 An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield associated QTL. Genetics **141:** 1147–1162.

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

Franklin, I. R., 1977 The distribution of the proportion of genome which is homozygous by descent in inbred individuals. Theor. Popul. Biol. **11:** 60–80.

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.

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.

Haldane, J. B. S., and C. H. Waddington, 1931 Inbreeding and linkage. Genetics **16:** 357–374.

Hazel, L. N., 1943 The genetic basis for constructing selection indices. Genetics **28:** 476–490.

Heckenberger, M., M. Bohn and A. E. Melchinger, 2005a Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. I. SSR data from maize inbreds. Crop Sci. **45:** 1132–1140.

Heckenberger, M., M. Bohn, M. Frisch, H. P. Maurer and A. E. Melchinger, 2005b Identification of essentially derived varieties with molecular markers: an approach based on statistical

test theory and computer simulations. Theor. Appl. Genet. **111:** 598–608.
Hill, W. G., 1993 Variation in genetic composition in backcrossing programs. J. Hered. **84:** 212–213.
Hospital, F., C. Dillmann and A. E. Melchinger, 1996 A general algorithm to compute multilocus genotype frequencies under various mating systems. Comput. Appl. Biosci. **12:** 455–462.
Ihaka, R., and R. Gentleman, 1996 A language for data analysis and graphics. J. Comput. Graph. Stat. **5:** 299–314.
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.
Servin, B., C. Dillmann, G. Decoux and F. Hospital, 2002 MDM: a program to compute fully informative genotype frequencies in complex breeding schemes J. Hered. **93:** 227–228.
Shao, J., 1999 *Mathematical Statistics*. Springer-Verlag, New York.
Stam, P., 1979 Interference in genetic crossing over and chromosome mapping. Genetics **92:** 573–594.
Tanksley, S. D., and J. C. Nelson, 1996 Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor. Appl. Genet. **92:** 191–203.
Visscher, P. M., 1996 Proportion of the variance in genetic composition in backcrossing programs explained by molecular markers. J. Hered. **87:** 136–138.
Visscher, P. M., and R. Thompson, 1995 Haplotype frequencies of linked loci in backcross populations derived from inbred lines. Heredity **75:** 644–649.

Communicating editor: R. W. Doerge

## APPENDIX

We derive the probabilites $p$ and $q_{uv}$ for $(F_1)^t$-DH, $(F_2)^t$-SSD, $BC_f$DH, and $BC_f$SSD lines. For the derivations we use the relationship

$$\begin{aligned} q_{uv} &= P(G_v = 1 \mid G_u = 1) \\ &= P(G_v = 1, G_u = 1)/P(G_u = 1) \\ &= p + D_{uv}/p, \end{aligned}$$

where

$$D_{uv} = P(G_v = 1, G_u = 1) - P(G_v = 1)P(G_u = 1)$$

is the expected gametic disequilibrium between two loci at positions $x_v$ and $x_u$ in infinite populations.

**$(F_1)^t$-DH lines:** The probability that a locus of an $(F_1)^t$-DH line carries the allele of parent $P_1$ is $p = \frac{1}{2}$. The expected linkage disequilibrium in an $(F_1)^1$ (*i.e.*, an $F_2$) population is

$$\begin{aligned} D_{uv} &= P(G_v = 1, G_u = 1) - P(G_v = 1)P(G_u = 1) \\ &= \frac{1 - r_{uv}}{2} - \frac{1}{4} \\ &= \frac{1 - 2r_{uv}}{4}. \end{aligned}$$

Because (i) the expected gametic disequilibrium in an $(F_1)^{t-1}$-derived DH line equals that of an $(F_1)^t$ population and (ii) in random mating populations, the linkage disequilibrium decreases with ratio $(1 - r_{uv})$ per generation (Falconer and Mackay 1996, p. 18), the expected gametic disequilibrium for $(F_1)^t$-DH lines is

$$D_{uv} = \frac{1 - 2r_{uv}}{4}(1 - r_{uv})^t. \qquad \text{(A1)}$$

In consequence, we have

$$\begin{aligned} q_{uv} &= p + D_{uv}/p \\ &= \frac{1}{2} + \frac{1 - 2r_{uv}}{2}(1 - r_{uv})^t. \end{aligned}$$

**$(F_2)^t$-SSD lines:** The probability that a locus in an $(F_2)^t$-DH line carries the allele of parent $P_1$ is $p = \frac{1}{2}$. The linkage disequilibrium in SSD lines derived from a population in Hardy–Weinberg equilibrium with linkage disequilibrium of $D'_{uv}$ is

$$D_{uv} = \frac{D'_{uv}}{1 + 2r_{uv}}$$

(Cockerham and Weir 1973). Because for an $(F_2)^t$ population (derivation in analogy to Equation A1)

$$D_{uv} = \frac{1 - 2r_{uv}}{4}(1 - r_{uv})^t,$$

we have for $(F_2)^t$-SSD lines

$$D_{uv} = \frac{1}{1 + 2r_{uv}} \frac{1 - 2r_{uv}}{4}(1 - r_{uv})^t$$

and, therefore,

$$\begin{aligned} q_{uv} &= p + D_{uv}/p \\ &= \frac{1}{2} + \frac{1 - 2r_{uv}}{2 + 4r_{uv}}(1 - r_{uv})^t. \end{aligned}$$

**$BC_f$DH lines:** The probability that a locus of $BC_f$ derived DH line carries the allele of parent $P_1$ is $p = (1/2)^{t+1}$ and the probability that a locus at position $x_v$ carries the allele of $P_1$ under the condition that the locus at position $x_u$ carries the allele of parent $P_1$ is

$$\begin{aligned} q_{uv} &= P(G_v = 1 \mid G_u = 1) \\ &= P(G_v = 1, G_u = 1)/P(G_u = 1) \\ &= \left(\frac{1 - r_{uv}}{2}\right)^{t+1} \Big/ \left(\frac{1}{2}\right)^{t+1} \\ &= (1 - r_{uv})^{t+1}. \end{aligned}$$

**$BC_f$SSD lines:** The probability that a locus in a $BC_f$ derived SSD line carries the allele of parent $P_1$ is $p = (1/2)^{t+1}$. The probability that continued selfing of an individual with genotype ABab results in an inbred with one of genotypes AbAb or aBaB is

$$\frac{2r_{uv}}{1 + 2r_{uv}}$$

(Haldane and Waddington 1931). Consequently, for $BC_f$SSD lines,

$$P(G_v = 1, G_u = 1) = \left(\frac{1 - r_{uv}}{2}\right)^t \frac{1}{2}\left(1 - \frac{2r_{uv}}{1 + 2r_{uv}}\right)$$

and, therefore,

$$\begin{aligned} q_{uv} &= P(G_v = 1, G_u = 1)/P(G_u = 1) \\ &= \frac{(1 - r_{uv})^t}{1 + 2r_{uv}}. \end{aligned}$$