Theor Appl Genet (2005) 111: 100–109
DOI 10.1007/s00122-005-1998-4

ORIGINAL PAPER

E. Esch

Estimation of gametic frequencies from F_2 populations using the EM algorithm and its application in the analysis of crossover interference in rice

Received: 13 July 2004 / Accepted: 7 March 2005 / Published online: 20 April 2005

Abstract The gametes produced in meiosis provide information on the frequency of recombination and also on the interdependence of recombination events, i.e. interference. Using F_2 individuals, it is not possible in all cases to derive the gametes, which have fused, and which provide the information about interference unequivocally when three or more segregating markers are considered simultaneously. Therefore, a method was developed to estimate the gametic frequencies using a maximum likelihood approach together with the expectation maximisation algorithm. This estimation procedure was applied to F_2 mapping data from rice (*Oryza sativa* L.) to carry out a genome-wide analysis of crossover interference. The distribution of the coefficient of coincidence in dependence on the recombination fraction revealed for all chromosomes increasing positive interference with decreasing interval size. For some chromosomes this mutual inhibition of recombination was not so strong in small intervals. The centromere had a significant effect on interference. The positive interference found in the chromosome arms were reduced significantly when the intervals considered spanned the centromere. Two chromosomes even demonstrated independent recombination and slightly negative interference for small intervals including the centromere. Different marker densities had no effect on the results. In general, interference depended on the frequency of recombination events in relation to the physical length. The strength of the centromere effect on interference seemed to depend on the strength of recombination suppression around the centromere.

Electronic Supplementary Material Supplementary material is available for this article at 10.1007/s00122-005-1998-4

Communicated by H.C. Becker

E. Esch
Abteilung Angewandte Genetik, Universität Hannover,
Herrenhäuser Straße 2, 30419 Hannover, Germany
E-mail: esch@genetik.uni-hannover.de
Tel.: +49-511-7623603
Fax: +49-511-7623608

Keywords Crossover interference · *Oryza sativa* · Gametic frequencies · Coefficient of coincidence · Centromere

Introduction

Meiosis, the nuclear division resulting in gametes, is one of the most important processes to generate genetic variability. Especially the crossover and recombination, respectively, which take place during meiosis contribute to genetic variation by breaking up linkage (i.e. the association of genes located on the same chromosome) and thereby creating new combinations of genes in the gametes. The frequency of crossovers between two loci and thereby, the strength of linkage, depends on the distance of these two loci on the chromosome. The incidence of recombination can be used to measure the distance between loci by analysing the segregation of genes and markers.

The degree of linkage and thus, the distance, between loci is estimated with the recombination fraction. It is the probability for recombinant (i.e. non-parental) gametes to be generated during meiosis. To determine the frequency of recombinant gametes and linkage, the populations used have to be in linkage disequilibrium. Therefore, in plant genetics usually certain types of populations with a high degree of linkage disequilibrium are used, e.g. F_2 populations, BC_1 populations or double haploid (DH) lines. In these populations (except the DH lines) the gametes cannot be observed directly. Consequently, the genotypes, which are the result of the random fusion of the gametes, and their frequencies, respectively, are used to estimate the recombination fraction (reviewed, e.g. by Weber and Wricke 1994 and Liu 1998).

The gametic frequencies do not only provide information concerning the frequency of recombination but also on the interdependence of recombination events. In particular, the gametes, which are produced from individuals heterozygous at multiple loci, give this additional information when three or more markers are considered simultaneously. In most species crossovers do not occur independently from each other. This is called interference, i.e. the non-random distribution of crossovers and recombination events, respectively (Muller 1916). In case of interactions between recombination events two situations can be distinguished: with positive interference a recombination event inhibits further ones in its vicinity, and with negative interference additional recombination events are supported. So far known exceptions where a random distribution of crossovers was found are *Schizosaccharomyces pombe* (Snow 1979; Munz 1994), *Aspergillus nidulans* (Egel-Mitani et al. 1982) and *Ascobolus immersus* (Hastings 1988).

Usually, positive interference is assumed in eukaryotes. Recently, Esch and Weber (2002) showed in barley, using data sets from seven mapping populations, that in addition to positive interference strong negative interference could be found. The relationship between recombination fraction and interference could not be described with a uniform function, which is assumed by most of the mapping functions used in genetic mapping. Interference seems not to act in the same way in the whole genome. In barley positive interference was found in the chromosome arms and no or negative interference in the genetically small but physically large centromeric region. Esch and Weber (2002) used data from DH lines where the gametic frequencies used for the analysis of interference can be determined directly from the genotypes. In contrast to DH populations, in F_2 populations—which are widely used in genetic mapping in plants—it is not possible to derive the F_1 gametes which have fused to form an F_2 individual unequivocally from all F_2 genotypes when considering three or more markers simultaneously.

The aim of the present study was to develop a method to estimate the gametic frequencies in F_2 populations in order to extend the approach used by Esch and Weber (2002) to the F_2 population type and thus make more data accessible for interference analysis. For this purpose a maximum likelihood approach together with the expectation maximisation (EM) algorithm (Dempster et al. 1977; Liu 1998) was used.

The estimation procedure developed was applied to a data set underlying one of the largest genetic maps in plants, the rice genetic map published by Harushima et al. (1998). Because in rice large integrated genetic and physical maps exist (Wu et al. 2002, 2003; Chen et al. 2002), a detailed comparison of the results on interference with data on the distribution and frequency of recombination in the physical map could be done.

Due to the high marker density of that map it was possible to study as an additional methodical aspect the influence of the marker density on the methods proposed by Esch and Weber (2002). Using their methods, regions in the genome with different marker density may contribute unequally to the analysis. Thus, derived data sets with evenly distributed markers were analysed.

Materials and methods

Estimation of gametic frequencies

Considering recombination in two adjacent intervals between the markers A, B, and C four different types of gametes exist:

1. Gametes showing recombination between marker A and B, and simultaneously between marker B and C (double recombination), having the expected frequency a_{12}.
2. Gametes showing recombination between marker A and B only, having the expected frequency a_1.
3. Gametes showing recombination between marker B and C only, having the expected frequency a_2.
4. Gametes showing no recombination in either interval, having the expected frequency a_0.

Considering the markers A, B and C, the F_1 gametes, which have fused to form an F_2 individual, cannot be derived unequivocally from each genotype in an F_2 population. This is only possible for the threefold homozygous (e.g. AABBCC from two ABC gametes) and the ‘onefold’ heterozygous genotypes (e.g. AaBBCC from the gametes ABC and aBC).

Nevertheless, it is possible for all observed genotype frequencies, n_i, to give the expected genotype frequencies, e_i, in terms of the expected gametic frequencies, a_j. Typical examples are given in Table 1 for the case of all three markers being codominantly inherited. (The complete table is provided as Electronic Supplementary Material, Table 1). Because the data set used to apply the estimation procedure contained only few dominant markers, which are less informative, only codominant markers were considered. However, from Table 1 and Electronic Supplementary Material Table 1, respectively, all possible situations considering dominant markers and linkage phase can be deduced by combining the appropriate genotype classes.

Maximum likelihood (ML) estimates of the expected gametic frequencies a_{12}, a_1, a_2 and a_0 were obtained using the EM algorithm (Dempster et al. 1977; Everitt 1987; Liu 1998). This is an iterative approach, which is powerful when the observations have incomplete data. Considering three codominantly inherited loci with two alleles each in an F_2 progeny, only 27 categories can be observed. However, there are 36 categories for complete information. Concerning the four different types of gametes, the data are incomplete in the sense that we do not know for each observed genotype, i.e. the two- and threefold heterozygous genotypes, to what proportion it consists of the respective categories. For example, the genotype AABbCc could have emerged from the fusion either of the gametes ABC and Abc, or of the gametes AbC and ABc.

Each iteration of the EM algorithm comprises two steps—the expectation step E and the maximisation step M—which are performed repeatedly. In the first iteration, $a_j = 0.25$ was used as initial guess for all four unknown gametic frequencies.

In the E step the a_j were used to estimate the complete data. For this, a distribution of the complete data conditional on the observed data is needed, i.e. the probability that an individual arose from a specific type of gamete conditional on the observed genotype. These conditional probabilities P_i $(E_j|G)$ are given in Table 1 and Electronic Supplementary Material Table 1.

In the following M step the new ML estimates of the gametic frequencies were computed as the number of the respective recombinants divided by the total number of observations:

$$a'_j = \frac{\sum n_i P_i(E_j|G)}{\sum n_i}$$

The E and M steps were repeated until the convergence criterion $|a_0 - a'_0| \leqslant 1 \times 10^{-9}$ was satisfied.

The EM algorithm was extremely easy to implement on a computer, and was realized with a simple Pascal programme on a Sparc workstation (Sun Microsystems, Palo Alto, Calif., USA). Despite the EM algorithm often shows a slow convergence (Everitt 1987), the estimates for the gametic frequencies converged fast, and for all marker triples estimates could be obtained.

Estimation of recombination frequencies and the coefficient of coincidence

The estimates of the gametic frequencies were used to calculate the recombination fractions $(\hat{r})$ between the three markers considered (A, B and C):

$$\hat{r}_{\mathrm{AB}} = a_1 + a_{12},$$

$$\hat{r}_{\mathrm{BC}} = a_2 + a_{12},$$

$$\hat{r}_{\mathrm{AC}} = a_1 + a_2.$$

The relative position of the marker B between A and C was given by

$$\hat{q} = \frac{a_1 + a_{12}}{a_1 + a_2 + 2a_{12}} = \frac{\hat{r}_{\mathrm{AB}}}{\hat{r}_{\mathrm{AB}} + \hat{r}_{\mathrm{BC}}}.$$

For the analysis of interference the coefficient of coincidence was used (Muller 1916). It is defined as the quotient of the observed frequency of double recombinations and the expected frequency of double recombinations with independence of recombination in the adjacent intervals AB and BC. It was estimated by

$$\hat{C} = \frac{a_{12}}{(a_1 + a_{12})(a_2 + a_{12})} = \frac{a_{12}}{\hat{r}_{\mathrm{AB}} \cdot \hat{r}_{\mathrm{BC}}}.$$

The expectation is $\mathrm{E}(\hat{C}) = 1$ without interference, $\mathrm{E}(\hat{C}) < 1$ in case of positive interference, and $\mathrm{E}(\hat{C}) > 1$ in case of negative interference. In the case of $\hat{r}_{\mathrm{AB}} = 0$ or $\hat{r}_{\mathrm{BC}} = 0$ the coefficient of coincidence $\hat{C}$ is not defined, because only one interval exists.

Application of the estimation methods to the analysis of interference in experimental F_2 mapping data

The estimation procedure developed above was applied to mapping data from rice (Harushima et al. 1998). The mapping population consisted of 186 F_2 plants. The data of 2,277 markers were reduced to 1,175 distinct map positions, because cosegregating markers do not provide additional information about interference. From the cosegregating markers those with the most data points were retained in the data set. Dominant markers were also not used. Due to limitations in computer capacity the number of markers for the chromosomes 1, 2 and 3 was restricted to 120 by removing markers evenly according to their order. The markers left after this selection procedure were called 'all markers', and their numbers per chromosome are shown in Table 2. For these markers the percentage of missing values was 1.7%.

The gametic frequencies, the recombination fractions and the coefficient of coincidence were estimated as described above for all combinations of three markers (triple) possible when considering each chromosome separately. When building triples the marker order given by Harushima et al. (1998) was considered as fixed. Triples with $\hat{r}_{\mathrm{AB}} = 0$ or $\hat{r}_{\mathrm{BC}} = 0$ were omitted as in such cases $\hat{C}$ was not defined (see above).

Influence of marker density

In order to analyse the influence of the marker density different data sets with markers distributed evenly along the chromosomes were generated from the rice data set 'all markers' (Table 2). For this, markers were retained in a framework of 2 cM (±1 cM deviation), 5 cM (±2 cM) and 10 cM (±3 cM), respectively. All other

Table 1 Genotypes in the F_2 population: frequency (n_i), expectation (e_i) and probability that an individual arose from a gamete produced at a specific recombination event conditional on the observed genotype [P_i $(E_j|G)$]

Genotype	Frequency	Expectation	P_i $(E_0\|G)$	P_i $(E_1\|G)$	P_i $(E_2\|G)$	P_i $(E_{12}\|G)$
Threefold homozygous, e.g. AABBCC	n_1	$e_1 = \frac{a_0^2}{4}$	1	0	0	0
Twofold homozygous, e.g. AABBCc	n_2	$e_2 = \frac{a_0 a_2}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0
'Onefold' homozygous e.g. AABbCc	n_3	$e_3 = \frac{a_0 a_1 + a_{12} a_2}{2}$	$\frac{a_0 a_1}{2(a_0 a_1 + a_2 a_{12})}$	$\frac{a_0 a_1}{2(a_0 a_1 + a_2 a_{12})}$	$\frac{a_2 a_{12}}{2(a_0 a_1 + a_2 a_{12})}$	$\frac{a_2 a_{12}}{2(a_0 a_1 + a_2 a_{12})}$
'Zerofold' homozygous, e.g. AaBbCc	n_4	$e_4 = \frac{a_{12}^2 + a_1^2 + a_2^2 + a_0^2}{2}$	$\frac{a_0^2}{a_{12}^2 + a_1^2 + a_2^2 + a_0^2}$	$\frac{a_1^2}{a_{12}^2 + a_1^2 + a_2^2 + a_0^2}$	$\frac{a_2^2}{a_{12}^2 + a_1^2 + a_2^2 + a_0^2}$	$\frac{a_{12}^2}{a_{12}^2 + a_1^2 + a_2^2 + a_0^2}$

In the examples, the recombination events are E_0, resulting in the gametes ABC or abc, E_1 with Abc or aBC, E_2 with ABc or abC and E_{12} with AbC or aBc (see text as well for explanation)

Table 2 Number of markers and mean map distance of the data sets used

Chromosome	Data set			
	All markers	2-cM Framework	5-cM Framework	10-cM Framework
1	120	69	32	18
2	120	63	30	15
3	120	61	32	17
4	86	42	22	12
5	102	50	24	13
6	93	46	23	13
7	86	41	22	11
8	79	44	25	13
9	49	27	13	7
10	60	30	13	9
11	69	38	23	12
12	53	27	16	9
Total number of markers	1,037	538	275	149
Mean distance between markers	1.5 cM	2.9 cM	5.7 cM	10.7 cM

markers were removed. The number of this framework markers and their mean distance according to the centiMorgan distance given by Harushima et al. (1998) are shown in Table 2.

Statistical test of the experimental coefficients of coincidence

To test the coefficients of coincidence $\hat{C}$ in the experimental data against the null hypothesis of independent recombination, information was needed about the distribution under this null hypothesis. Following Esch and Weber (2002), simulations where used to obtain the distribution in case of independent recombination considering the present situation of small population size, and to derive a test statistic from that distribution. The simulation of the distribution of the coefficient of coincidence in small F_2 populations was analogous to the simulations used by Esch and Weber (2002) for DH lines:

The probabilities of the four types of gametes a_j, and the expected frequencies of the different F_2 genotypes were expressed as functions of r_{AC} and q. In the simulations under no interference, values for r_{AC} and q were given, and the expected frequencies of the 27 genotype classes were calculated. Using these frequencies samples of F_2 genotypes were simulated. Corresponding to the actual mean number of plants per triple in the experimental data (taking into account missing data) the number of simulated genotypes per sample was 180. From the simulated genotypes the gametic frequencies were estimated using the EM algorithm as described above. Using the gametic frequencies the recombination frequencies, the coefficient of coincidence and the ratio between the two subintervals were calculated.

To evaluate the unbiasedness of the simulated distribution without interference 10^5 samples were performed for each combination of r_{AC} and q. The estimated $\hat{r}_{AC}$ and $\hat{q}$ were unbiased, even for very small given values for r_{AC} and extreme given values for q (data not shown). Table 3 shows the estimated values for the coefficient of coincidence $\hat{C}$. The deviation of the coefficient of coincidence from the expected value $E(\hat{C}) = 1$ (no interference) increased with decreasing recombination frequency. Due to this bias the simulated distribution was restricted to r_{AC}≥0.1. The estimation of the coefficient of coincidence was irrespective of r_{AC} hardly influenced by q for $0.35 \leq q \leq 0.65$.

Table 3 Estimated mean and standard deviation (*SD*) of the coefficient of coincidence $\hat{C}$ from simulation (10^5 per parameter combination) with different values given for the recombination frequency r_{AC} and the proportion of the involved intervals q for 180 genotypes in case of no interference ($C=1$)

r_{AC}	$q=0.5$		$q=0.35$	
	Mean	SD	Mean	SD
0.05	1.37	2.66	1.39	2.85
0.075	1.16	1.56	1.20	1.68
0.1	1.03	1.15	1.05	1.21

The comparison of the distribution of the experimental data with the simulated data under the null hypothesis of no interference was performed as described by Esch and Weber (2002). Confidence intervals of the means were determined with a significance level of 5% corrected for multiple comparisons according to Bonferroni. In the simulations r_{AC} was varied between 0.1 and 0.5 in steps of 0.05, with $q=0.5$, and 10^6 simulations were performed for each combination of r_{AC} and q. Like in the experimental data triples with $\hat{q} = 0$ or $\hat{q} = 1$, this means $\hat{r}_{AB} = 0$ and $\hat{r}_{BC} = 0$ respectively, were omitted. According to the simulations the experimental data were restricted to $0.075\leqslant\hat{r}_{AC}<0.525$ (class means 0.1–0.5 in steps of 0.05) and $0.35\leqslant\hat{q}\leqslant0.65$.

Results

Distribution of the coefficient of coincidence in dependence on the recombination frequency in the experimental data

The number of triples for which the coefficient of coincidence can be calculated increases very quickly with the number of markers. One possibility to describe the distribution of this huge amount of values is to classify them according to the recombination frequency and to consider means, median, quartiles and percentiles of the different classes. This can also be done for the coefficients of coincidence from the simulation under the null hypothesis of independent recombination. For each class the null hypothesis is then tested by comparing the means of the experimental and the simulated distribution. For this test confidence intervals of the means were calculated from the simulation according to the number of values in the corresponding class in the experimental distribution.

In Fig. 1 the means of the experimental distributions for the 12 rice chromosomes and the result of the test of the null hypothesis are shown. For all chromosomes positive interference was found often similar to the Kosambi function (Kosambi 1944). With this function strong positive interference in small intervals is assumed that linearly decreases with increasing interval size until independence is reached. In Fig. 1 the Kosambi function would result in a straight line from $\hat{C} = 0$ at $\hat{r}_{\mathrm{AC}} = 0$ to $\hat{C} = 1$ at $\hat{r}_{\mathrm{AC}} = 0.5$. This was obvious particularly for chromosome 3. Negative interference $\left(\hat{C} > 1\right)$ was observed on chromosome 1 for intervals between 0.375 and 0.475. For chromosomes 2, 4, 7 and 10 the mutual inhibition of recombination events was not so strong in small intervals. For chromosome 7 in very small intervals no interference was found. Positive interference was not so strong for intermediate intervals in chromosomes 11 and 12. Even recombination in large intervals $(0.475 \leqslant \hat{r}_{\mathrm{AC}} < 0.525)$ was not independent. This observation was also made in barley (Esch and Weber 2002).

Variation of interference within the genome and influence of the centromere

In barley the interference level around the centromere was different from that in the chromosome arms, and it seemed that interference was dependent on the frequency of recombination events in relation to the physical length (Esch and Weber 2002). The influence of the centromere on interference was also investigated in the rice data set.

The positions of the centromeres in the linkage map of the data set used were determined from Harushima et al. (1998), and confirmed by Wang et al. (2000) with the exception of chromosome 10. For chromosome 10 the centromere position determined by Cheng et al. (2001) was used (between 15.4 cM and 15.9 cM in the genetic map of Harushima et al. 1998). The centromere positions of chromosomes 1, 2 and 6–9 were confirmed recently by Wu et al. (2003).

To analyse the influence of the centromere the triples were classified to whether they spanned the centromere or not. To consider the interval size the triples were also classified into five classes according to the recombination frequency of the entire interval. The analysis was done for each chromosome separately and for the whole genome. The corresponding classes were compared using the Wilcoxon rank sum test (Wilcoxon 1946). In Table 4 the influence of the centromere on interference is demonstrated for the whole genome and for individual chromosomes showing some differing characteristics.

From the results of the whole genome a clear effect of the centromere on the mutual influence of recombination events could be realized. Positive interference was found in the chromosome arms, increasing with decreasing interval size similar to the Kosambi function. This positive interference was reduced when the centromere was included in the interval. For large intervals $(\hat{r}_{\mathrm{AC}} \geqslant 0.375)$ spanning the centromere recombination was almost independent.

The analysis of each chromosome separately revealed some differences from that general pattern for chromosomes 5, 7, 10, 11 and 12. In chromosomes 7, 11 and 12 the mean coefficient of coincidence of intervals without the centromere first decreased or was constant with decreasing interval size and then increased for smaller intervals. Compared to the genome-wide analysis the

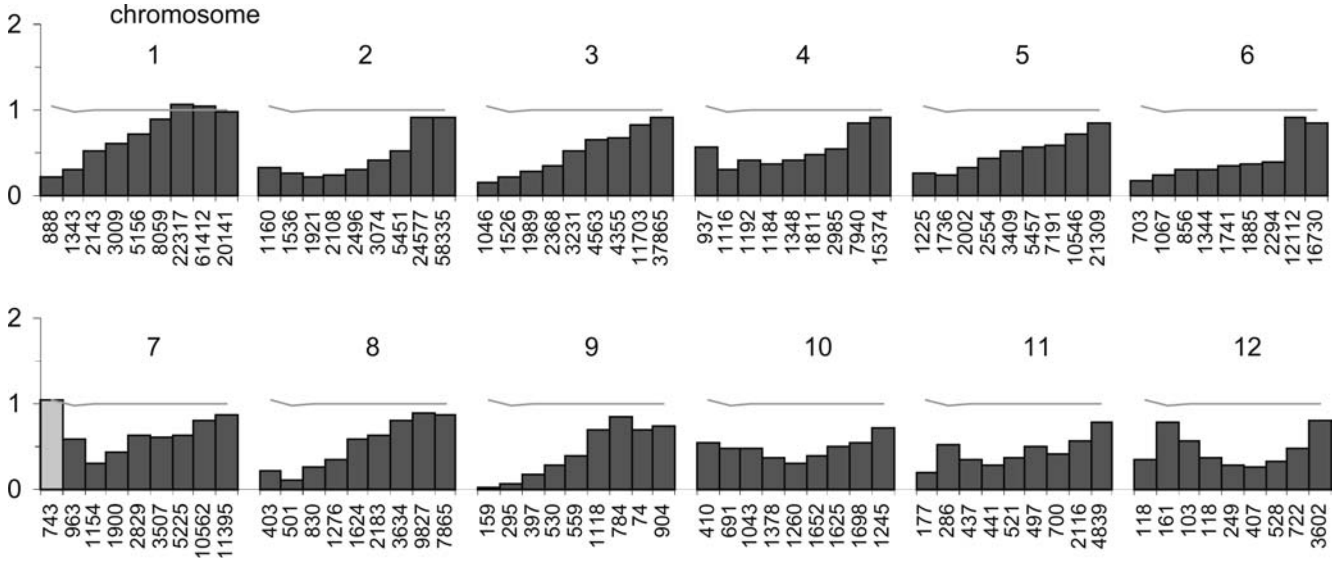

Fig. 1 Mean coefficient of coincidence (*y-axis*) for classes of recombination frequencies (*x-axis*) for the 12 rice chromosomes. Classes range from 0.075 to 0.525 in steps of 0.05 (from *left* to *right*). Along the *x*-axis the number of triples for each class is given. The *lines* indicate the mean values of the simulation without interference. Means, which significantly deviate from the simulated mean, are shown in *dark grey*

Table 4 Influence of the centromere on the coefficient of coincidence $\hat{C}$. Triples were classified according to the recombination frequency $\hat{r}_{AC}$ and whether they spanned the centromere or not. Mean and SD of $\hat{C}$ were calculated for the whole genome (1–12) and for each chromosome separately (not all results shown). *P*-values from the Wilcoxon rank sum test

Chromosome	$\hat{r}_{AC}$	With centromere			Without centromere			*P*-value
		Number of triples	Mean	SD	Number of triples	Mean	SD	
1–12	0.075–0.175	1,855	0.60	0.80	16,625	0.28	0.52	< 0.0001
	0.175–0.275	6,142	0.51	0.38	25,501	0.34	0.35	< 0.0001
	0.275–0.375	17,598	0.67	0.34	40,219	0.54	0.34	< 0.0001
	0.375–0.475	165,481	0.95	0.24	43,348	0.63	0.30	< 0.0001
	0.475–0.525	174,315	0.91	0.14	23,962	0.78	0.18	< 0.0001
5	0.075–0.175	279	1.01	0.68	2,555	0.15	0.42	< 0.0001
	0.175–0.275	1,373	0.73	0.42	2,956	0.24	0.43	< 0.0001
	0.275–0.375	4,365	0.76	0.24	4,048	0.30	0.24	< 0.0001
	0.375–0.475	13,646	0.73	0.20	3,383	0.43	0.13	< 0.0001
	0.475–0.525	21,280	0.85	0.13	17	0.59	0.14	–[a]
7	0.075–0.175	377	1.30	1.12	1,329	0.63	0.70	< 0.0001
	0.175–0.275	937	0.63	0.43	2,117	0.27	0.32	< 0.0001
	0.275–0.375	3,354	0.83	0.28	2,982	0.38	0.19	< 0.0001
	0.375–0.475	12,254	0.83	0.27	3,533	0.48	0.24	< 0.0001
	0.475–0.525	11,239	0.88	0.16	156	0.58	0.05	< 0.0001
10	0.075–0.175	92	0.60	0.19	512	0.18	0.38	< 0.0001
	0.175–0.275	718	0.49	0.30	1,427	0.35	0.20	< 0.0001
	0.275–0.375	1,250	0.18	0.17	1,662	0.49	0.26	< 0.0001
	0.375–0.475	2,355	0.48	0.27	968	0.64	0.13	< 0.0001
	0.475–0.525	1,034	0.74	0.06	211	0.62	0.08	< 0.0001
11	0.075–0.175	14	0.10	0.37	449	0.40	0.50	–[a]
	0.175–0.275	104	0.38	0.20	774	0.31	0.20	< 0.0001
	0.275–0.375	436	0.68	0.32	582	0.24	0.12	< 0.0001
	0.375–0.475	2,063	0.63	0.22	753	0.24	0.10	< 0.0001
	0.475–0.525	4,430	0.82	0.19	409	0.27	0.11	< 0.0001
12	0.075–0.175	45	0.57	0.53	231	0.59	0.79	0.8972
	0.175–0.275	125	0.56	0.27	73	0.36	0.48	< 0.0001
	0.275–0.375	173	0.55	0.19	459	0.16	0.16	< 0.0001
	0.375–0.475	996	0.48	0.21	188	0.15	0.11	< 0.0001
	0.475–0.525	3,482	0.81	0.19	74	0.27	0.06	< 0.0001

[a] Differences were not tested if the number of triples in a single class were too small

mutual inhibition of recombination events in the chromosome arms of these chromosomes was stronger in larger intervals and weaker in small intervals.

For chromosomes 5 and 7 the influence of the centromere was so strong that in very small intervals spanning the centromere, independent recombination and slightly negative interference was found, respectively. Chromosome 11 showed also for intervals with the centromere a relationship between interference and recombination frequency comparable to the Kosambi function. For intermediate intervals spanning the centromere of chromosome 10, a stronger inhibition of recombination could be observed compared to intervals within the chromosome arms.

Because here in this analysis and also in the distribution of the coefficient of coincidence in dependence on the recombination frequency, deviations were found mainly for small intervals, the locations of intervals of the size $0.075 \leqslant \hat{r}_{AC} < 0.225$ and their values for the coefficient of coincidence were studied in more detail. The assignment of the coefficients of coincidence to distinct locations within the genome was done by calculating 'marker means' (Esch and Weber 2002). The coefficient of coincidence of a triple was assigned to the mid-marker, because this marker enabled the observation of double recombinations in the entire interval. The mean of all $\hat{C}$ values assigned to a certain marker was calculated, and this marker mean could then be assigned to a location in the genetic map. Only marker means calculated from at least ten single values were considered. The distribution of the marker means throughout the genome showed some regions with locally increased values. In Table 5 regions comprising at least three markers with marker means larger than 1 are given. These five regions had a size between 1.3 cM and 7.4 cM. One of the regions on chromosomes 5 and 7 each covered the centromere, which is located between 52.3 cM and 53.7 cM on chromosome 5 and at 49.3 cM on chromosome 7 (Harushima et al. 1998). Because these regions had a high marker density, it was investi-

Table 5 Regions with at least three markers showing mean coefficients of coincidence for the markers (*Marker mean*) greater than 1. Regions were found using all markers and were confirmed using the 2-cM framework data set

Chromosome	All markers			2 cM Framework	
	Position in (cM)	Number of markers	Marker mean	Number of markers	Marker mean
4	66.6–69.1	5	1.16–1.76	1	1.51
5	46.0–53.4	8	1.00–1.22	4	1.00–1.32
	71.4–75.7	5	0.92–1.48	2	1.25–1.59
7	46.4–50.7	10	1.43–2.42	2	1.43–2.21
	79.5–80.8	4	1.18–1.44	1	1.38

Table 6 Influence of the centromere on the coefficient of coincidence $\hat{C}$ in data sets with different marker density. Triples were classified according to the recombination frequency $\hat{r}_{\mathrm{AC}}$ and whether they spanned the centromere or not. Mean and SD were calculated for the whole genome. *P*-values from the Wilcoxon rank sum test

Data set	$\hat{r}_{\mathrm{AC}}$	With centromere			Without centromere			*P*-value
		Number of triples	Mean	SD	Number of triples	Mean	SD	
10-cM-Framework	0.075–0.175	246	0.48	0.68	2,253	0.26	0.47	< 0.0001
	0.175–0.275	890	0.49	0.34	3,806	0.34	0.35	< 0.0001
	0.275–0.375	2,580	0.66	0.34	5,943	0.54	0.35	< 0.0001
	0.375–0.475	28,113	0.97	0.23	6,137	0.64	0.31	< 0.0001
	0.475–0.525	25,997	0.92	0.13	2,617	0.77	0.17	< 0.0001
5-cM-Framework	0.075–0.175	44	0.68	0.77	293	0.29	0.52	0.0002
	0.175–0.275	136	0.44	0.34	456	0.32	0.33	0.0001
	0.275–0.375	390	0.62	0.31	660	0.47	0.34	< 0.0001
	0.375–0.475	3,289	0.94	0.25	780	0.61	0.30	< 0.0001
	0.475–0.525	3,566	0.91	0.14	377	0.76	0.18	< 0.0001
2-cMFramework	0.075–0.175	4	1.11	0.39	15	0.62	0.56	−[a]
	0.175–0.275	26	0.50	0.35	87	0.29	0.31	0.0041
	0.275–0.375	48	0.59	0.31	96	0.49	0.37	0.0765
	0.375–0.475	485	0.92	0.26	141	0.64	0.32	< 0.0001
	0.475–0.525	608	0.92	0.14	57	0.75	0.22	< 0.0001

[a] Differences were not tested if the number of triples in a single class were too small

gated if the high mean coefficients of coincidence were caused only by an effect of the marker density and/or an interdependency between the values for the individual triples. For this purpose the analysis of the small intervals was repeated with the 2-cM framework data set. Also in this data set the regions showed increased mean coefficients of coincidence (Table 5).

Regions with locally increased marker means could also be found on chromosomes 10, 11 and 12. For chromosome 10 this region comprised six markers between 11.0 cM and 15.9 cM, with mean $\hat{C}$ values between 0.88 and 1.27, for chromosome 11 seven markers between 7.2 and 10.2 cM, with mean values between 0.77 and 0.97 and for chromosome 12 four markers between 95.9 and 101.4 cM, with mean values between 0.92 and 1.81. The region on chromosome 10 included the centromere (between 15.4 cM and 15.9 cM, Cheng et al. 2001). The increase of the marker means in this regions was not as clear as that of the regions in Table 5, and could not be confirmed in the 2-cM framework data set because here the number of triples per marker mean was too low.

Influence of the marker density

The markers were not evenly distributed on the genetic map, but had different density in different regions. From regions with a high marker density more triples could be built and therefore, these areas outweigh areas with less markers in the analysis. Furthermore, the coefficients of coincidence for the individual triples are not independent from each other, because up to two markers and thus one subinterval could be identical. Therefore, in addition, the different pattern of dependency between the triples gets a different weight in the analysis. In order to investigate these influences framework data sets with evenly distributed markers in 2-, 5- and 10-cM distances, respectively, were used. In these data sets the distribution of the coefficient of coincidence in dependence on the recombination frequency and the influence of the centromere on interference was analysed.

The distribution of the coefficient of coincidence in dependence on the recombination frequency in the 10-cM framework data set could not be analysed for the chromosomes separately, because the numbers of triples per class were too small. The results of the 2-cM and 5-cM framework data set showed if any only minor differences to the results of all markers for intervals with $\hat{r}_{AC} \geqslant 0.225$ In the smaller intervals different effects of the wider marker distances could be observed. For chromosomes 2 and 5 the mean coefficients of coincidence slightly increased, whereas the opposite effect was observed for chromosome 4. For chromosomes 7, 11 and 12 the results varied between the data sets. Overall, in addition, for small intervals the differences between the different data sets were rather small.

The effect of the marker density on the influence of the centromere on interference is shown in Table 6 in comparison to Table 4. No differences were found for intervals larger than 0.175. The only difference worth mentioning concerns the class $0.075 \leqslant \hat{r}_{\mathrm{AC}} < 0.175$ in the 10-cM framework data set. Here the mean values of triples with and without centromere were increased, however with a very small sample size.

For the 5-cM framework, the data set still containing many markers the influence of the centromere on interference was analyses for the chromosomes separately. For chromosomes 5 and 7 the largest centromere effect, i.e. weakening of positive interference, could be observed in the analysis using all markers. This effect was maintained at lower marker density, however with reduced sample size. For the class $0.075 \leqslant \hat{r}_{AC} < 0.175$ of chromosome 5 the means coefficient of coincidence for triples including the centromere was 1.16 (number of triples $n=8$) compared to 0.23 ($n=28$) for triples without the centromere, and 1.53 ($n=9$) compared to 0.47 ($n=27$) for chromosome 7, respectively.

Discussion

The distribution of the coefficient of coincidence in dependence on the recombination frequency revealed for all chromosomes increasing positive interference with decreasing interval size similar to the Kosambi function. In a more general approach analysing the distribution of the number of recombinations per chromosome and the distribution of double crossover interval lengths, Harushima et al. (1998) also detected positive interference. With the detailed interference, analysis demonstrated in the present study it could be shown that in some chromosomes this positive interference was weakened to some extent in small intervals (Fig. 1).

Under the assumption of a limited number of crossovers for the whole genome, positive interference results in an even distribution along the chromosomes. In this way interference can ensure that each chromosome pair will have at least one crossover during meiosis, which is necessary for the regular distribution of the chromosomes in meiosis I (Egel 1995; Roeder 1997). In mutants of *Saccharomyces cerevisiae* showing no interference a higher rate of non-disjunction was observed (Sym and Roeder 1994; Chua and Roeder 1997). *A. nidulans* and *S. pombe*, both lacking interference, exhibit a much higher number of crossovers per chromosome pair and thereby the probability of a chromosome pair to get not any crossover is reduced (Storlazzi et al. 1995). An important role in regulation for the frequency and distribution of crossovers is attributed to the synaptonemal complex [(SC) Egel 1995]. This is mainly based on the observation that *S. pombe* and *A. nidulans* form no SC during meiosis (Olson et al. 1978) and at the same time shows a random distribution of crossovers (Snow 1979; Munz 1994; Egel-Mitani et al. 1982). How the SC is involved in interference is not yet clear (Egel 1995; Hasenkampf 1996; Kleckner 1996). It seems to be necessary for the transmission of a signal about the position of recombination events (Kleckner 1996). Models for positive interference and the involvement of the SC are reviewed by Kaback et al. (1999) and Novak et al. (2001).

Gorlov and Gorlova (2001) proposed a model which explains interference on the basis of a cost–benefit analysis of recombination. Therefore, the advantage of recombination lies in the production of new combinations of alleles of linked loci, which could be selectively advantageous. Because crossover and recombination are complex processes they are susceptible to errors, which will result in deleterious mutations. As a consequence positive interference prevents the occurrence of closely located crossovers. In computer simulations closely located crossovers were less effective in producing recombinant individuals compared to crossovers located more distantly. According to the model of Gorlov and Gorlova (2001) positive interference is especially important in the recombination hot spots. Without interference in those regions, the harmful effects of recombination would accumulate without the compensatory production of recombinant individuals.

The detailed investigation of the variation of interference within the genome presented here revealed an influence of the centromere on interference. This effect was also observed in barley (Esch and Weber 2002). In small intervals spanning the centromere positive interference was weakened. The strength of the effect was different for individual chromosomes. For chromosomes 5 and 7 even negative interference could be observed. Considering the whole genome the centromere effect on interference in rice was not as strong as in barley, and could therefore, with exception of chromosome 7, not be detected in the distribution of the coefficient of coincidence in dependence of the recombination frequency (Fig. 1).

As in barley, in rice a suppression of recombination in the centromeric region is described (Cheng et al. 2001; Chen et al. 2002; Wu et al. 2003). In comparison to the genome-wide average of 0.41 cM/100 kb (Wu et al. 2002) and 0.39–0.42 cM/100 kb (Wu et al. 2003), the relation of genetic/physical distance is reduced to values <0.037 cM/100 kb near the centromere (Wu et al. 2002) and 0 cM/100 kb at the centromere (Wu et al. 2003). In contrast to barley the suppression of recombination is restricted to the region immediate at the centromere (Cheng et al. 2001). From the results of Chen et al. (2002) it can be derived that this region corresponds to 8–25% of the physical length of the individual chromosomes compared to 40–60% in barley (Künzel et al. 2000). Wu et al. (2003) found a mean value of 11.4% of the entire size of the six rice chromosomes they analysed where recombination was completely suppressed.

In the chromosome arms of some rice chromosomes additional regions with weakened positive interference and independence of recombination events, respectively, were found. When comparing the physical and genetic map in the chromosome arms of chromosome 4S and 10S regions with suppressed recombination were identified (Chen et al. 2002; Wu et al. 2002). But these regions did not correspond to the regions without interference described in the present study (Table 5). In the integrated physical and genetic map (Wu et al. 2002) no special characteristics could be found for the regions without interference. Regions in the chromosome arms without interference or even negative interference have also been found in wheat (Peng et al. 2000).

The different marker densities had no influence on the distribution of the coefficient of coincidence and the observation of the centromere effect on interference. Reducing the marker density too much (10-cM framework) resulted in a very small number of triples, disabling the analysis of individual chromosomes. A marker density of on average 2–5 cM was proven to be suitable for the methods applied to investigate interference. A higher marker density resulted in an increased computing time, but led to the same results.

The principal mode of operation of interference in the genome which was found in barley (Esch and Weber 2002) could be confirmed in rice. As in barley the interference depends on the frequency of recombination events in relation to the physical length. Positive interference was observed in the chromosome arms, and this mutual inhibition of recombination events was weakened around the centromere where also the frequency of recombination is decreased. In rice this effect of the centromere was not as pronounced as in barley. Thus, the strength of the centromere effect on interference, i.e. reduction or abolishing of positive interference or even negative interference, seems to depend on the strength of recombination suppression around the centromere. Because recombination suppression in the centromeric region is a widespread observation in plants (Choo 1998) we are just analysing if the effect of the centromere on interference can be observed in more cases.

The comparison of the physical map of rice containing expressed sequence tags (ESTs) with the genetic map revealed that regions with a higher frequency of recombination usually also had a higher EST density (Wu et al. 2002). The EST density was reduced around the centromere, so this region contained relative few genes. According to the model of Gorlov and Gorlova (2001) positive interference is particularly important in gene rich regions to minimise the risk of a possibly faulty crossover. This could be interpreted also in the way that there is no need for positive interference in gene poor regions like the centromeric region. And therefore could be an explanation for the weakening and abolishment, respectively, of positive interference around the centromere which was observed in barley (Esch and Weber 2002) and rice, and which is possibly a general phenomenon in plants.

Acknowledgements The author thanks W.E. Weber (University of Halle, Germany) for stimulating and fruitful discussions about interference and for helpful comments on the manuscript.

References

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, et al (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J (2001) High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics 157:1749–1757

Choo KHA (1998) Why is the centromere so cold? Genome Res 8:81–82

Chua PR, Roeder GS (1997) Tam1, a telomere-associated meiotic protein, functions in chromosome synapsis and crossover interference. Genes Dev 11:1786–1800

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. JR Statist Soc 39B:1–38

Egel R (1995) The synaptonemal complex and the distribution of meiotic recombination events. Trends Genet 11:206–208

Egel-Mitani M, Olson LW, Egel R (1982) Meiosis in *Aspergillus nidulans*: another example for lacking synaptenomal complexes in the absence of crossover interference. Hereditas 97:179–187

Esch E, Weber WE (2002) Investigation of crossover interference in barley (*Hordeum vulgare* L.) using the coefficient of coincidence. Theor Appl Genet 104:786–796

Everitt BS (1987) Introduction to optimization methods and their application in statistics. Chapman & Hall, London

Gorlov IP, Gorlova OY (2001) Cost–benefit analysis of recombination and its application for understanding of chiasma interference. J Theor Biol 213:1–8

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2,275 markers using a single F_2 population. Genetics 148:479–494

Hasenkampf CA (1996) The synaptonemal complex—the chaperone of crossing over. Chromosome Res 4:133–140

Hastings PJ (1988) Conversion events in fugi. In: Kucherlapati R, Smith GR (eds) Genetic recombination. American Society for Microbiology, Washington, pp 397–428

Kaback DB, Barber D, Mahon J, Lamb J, You J (1999) Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae*: the role of crossover interference. Genetics 152:1475–1486

Kleckner N (1996) Meiosis: how could it work? Proc Natl Acad Sci USA 93:8167–8174

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175

Künzel G, Korzun L, Meister A (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. Genetics 154:397–412

Liu BH (1998) Statistical genomics: linkage, mapping and QTL analysis. CRC, Boca Raton

Muller HJ (1916) The mechanism of crossing over. Am Nat 50:193–221, 284–305, 350–366, 421–434

Munz P (1994) An analysis of interference in the fission yeast *Schizosaccharomyces pombe*. Genetics 137:701–707

Novak JE, Ross-Macdonald PB, Roeder GS (2001) The budding yeast Msh4 protein functions in chromosome synapsis and the regulation of crossover distribution. Genetics 158:1013–1025

Olson LW, Edén U, Egel-Mitani M, Egel R (1978) Asynaptic meiosis in fission yeast? Hereditas 89:189–199

Peng JH, Korol AB, Fahima T, Röder MS, Ronin YI, Li YC, Nevo E (2000) Molecular genetic maps in wild emmer wheat, *Triticum dicoccoides*: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Res 10:1509–1531

Roeder GS (1997) Meiotic chromosomes: it takes two to tango. Genes Dev 11:2600–2621

Snow R (1979) Maximum likelihood estimation of linkage and interference from tetrad data. Genetics 92:291–245

Storlazzi A, Xu L, Cao L, Kleckner N (1995) Crossover and noncrossover recombination during meiosis: timing and pathway relationships. Proc Natl Acad Sci USA 92:8512–8516

Sym M, Roeder GS (1994) Crossover interference is abolished in the absence of a synaptonemal complex protein. Cell 79:283–292

Wang S, Wang J, Jiang J, Zhang Q (2000) Mapping of centromeric regions on the molecular linkage map of rice (*Oryza sativa* L.) using centromere-associated sequences. Mol Gen Genet 263:165–172

Weber WE, Wricke G (1994) Genetic markers in plant breeding. Parey Scientific, Berlin

Wilcoxon F (1946) Individual comparison of grouped data by ranking methods. J Econ Entomol 39:269–270

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6,591 expressed sequence tag sites. Plant Cell 14:525–535

Wu JZ, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, Katagiri S, Saji S, Yoshiki S, Karasawa W, Yoshihara R, Hayashi A, Kobayashi H, Ito K, Hamada M, Okamoto M, Ikeno M, Ichikawa Y, Katayose Y, Yano M, Matsumoto T, Sasaki T (2003) Physical maps and recombination frequency of six rice chromosomes. Plant J 36:720–730