Theor Appl Genet (2014) 127:2585–2597
DOI 10.1007/s00122-014-2401-0

ORIGINAL PAPER

Computationally efficient map construction in the presence of segregation distortion

Rohan Shah · Colin R. Cavanagh · B. Emma Huang

Received: 7 March 2014 / Accepted: 11 September 2014 / Published online: 27 September 2014

Abstract

Key message **We present a novel estimator for map construction in the presence of segregation distortion which is highly computationally efficient. For multi-parental designs this estimator outperforms methods that do not account for segregation distortion, at no extra computational cost.**

Abstract Inclusion of genetic markers exhibiting segregation distortion in a linkage map can result in biased estimates of genetic distance and distortion of map positions. Removal of distorted markers is hence a typical filtering criterion; however, this may result in exclusion of biologically interesting regions of the genome such as introgressions and translocations. Estimation of additional parameters characterizing the distortion is computationally slow, as it relies on estimation via the Expectation Maximization algorithm or a higher dimensional numerical optimisation. We propose a robust M-estimator (RM) capable of handling tens of thousands of distorted markers from a single linkage group. We show via simulation that for multi-parental designs the RM estimator can perform much better than uncorrected estimation, at no extra computational cost. We then apply the RM estimator to chromosome 2B in wheat in a multi-parent population segregating for the Sr36 introgression, a known transmission distorter. The resulting map contains over 700 markers, and is consistent with maps constructed from crosses which do not exhibit segregation distortion.

List of symbols

M_1, M_2, M_3	Genetic markers
M_s	Segregation distortion locus (SDL)
r_{12}, r_{23}, r_{13}	Recombination fractions between markers M_1, M_2 and M_3
r_{1s}, r_{s3}	Recombination fractions between M_s and markers M_1, M_3
$g_y(t)$	Expected proportion of allele y at M_2
n_{xyz}	Number of lines with multilocus genotype x, y, z at markers M_1, M_2, M_3
$n_{x.z}$	Number of lines with multilocus genotype x, z at markers M_1, M_3
$\mathbb{P}_d$	Distorted probability model
$\mathbb{P}_{d,f}$	Distorted probability model for MAGIC8 population, assuming funnel f
$\mathbb{P}_{u,f}$	Undistorted probability model for MAGIC8 population, assuming funnel f
$\mathbb{P}_u$	Undistorted probability model
p_f	Proportion of lines from funnel f in a MAGIC8 population
p_{xyzf}	Proportion of lines from funnel f having multi-locus genotype x, y, z at marker M_1, M_2, M_3
$p_{.y.f}$	Proportion of lines from funnel f having genotype y at marker M_2
$\hat{p}_{.y.}$	Empirical proportion of lines having genotype y at M_2
$\hat{p}_{x.z}$	Empirical proportion of lines having multi-locus genotype x, z at M_1, M_3

Communicated by Jiankang Wang.

R. Shah · B. E. Huang (✉)
CSIRO Computational Informatics and Food Futures National Research Flagship, Brisbane, Australia
e-mail: emma.huang@csiro.au

R. Shah
University of Queensland School of Mathematics and Physics, Brisbane, Australia

C. R. Cavanagh
CSIRO Plant Industry and Food Futures National Research Flagship, Acton, Australia

$\hat{p}_{xyz}$	Empirical proportion of lines having multi-locus genotype x, y, z at M_1, M_2, M_3
$p_{xyz}(r_{12}, r_{23})$	Probability of multi-locus genotype x, y, z at markers M_1, M_2, M_3, with given recombination fractions and no distortion
$p_{.y.}$	Proportion of lines having genotype y at M_2
G	Number of underlying genotypes at each marker
F	Set of all funnels for the MAGIC8 population
$\|F\|$	Number of funnels for the MAGIC8 population

Introduction

Segregation distortion refers to the inheritance of alleles among progeny in a different proportion from that implied by Mendelian inheritance. Such distortion is commonly detected in major crop species, with a variety of underlying causes, including natural and artificial selection (for a review of common factors see Liu et al. 2010). While regions of distortion may thus be of biological interest, they are also problematic in genetic analysis, as they do not conform to standard assumptions about inheritance. Hence, regions of the genome displaying segregation distortion are often discarded during preprocessing of genotypes (Xu and Hu 2009) to avoid bias in genetic map construction and QTL mapping.

Discarding these regions not only removes biologically interesting segments of the genome (Wang et al. 2005; Xu 2008), but also may drastically reduce marker density. Previous work has attempted to account for segregation distortion, either by robust estimation of recombination fraction or parametrization of the process itself (Lorieux et al. 1995a, b; Cheng et al. 1996, 1998). However, the parameterization approach has a significant computational burden and robust estimation is difficult for designs where individual markers are highly non-informative.

The effect of segregation distortion varies depending on experimental design and genotyping platform. Lorieux et al. (1995a, b) showed that distortion can bias estimates of recombination fraction in F2 populations, although estimates between pairs of codominant markers are less affected than dominant markers. In contrast, Hackett and Broadfoot (2003) examined linkage map construction via simulation of doubled-haploid (DH) populations and found the effect of distortion to be small. However, they note that their experimental design is completely robust against distortion introduced by a single locus, a feature not shared by F2 populations with dominant markers.

Multi-parental designs were introduced as alternatives to biparental crosses for mapping populations (Cavanagh et al. 2008). Multiparent populations have been created in a number of crops including wheat and rice (Huang et al. 2012; Bandillo et al. 2013), as well as the model plant *Arabidopsis thaliana* (Kover et al. 2009). While these have numerous advantages for map construction including higher genetic diversity and resolution, they are also more vulnerable to the effects of segregation distortion. Biallelic markers are unable to distinguish between all of the parents, and hence these populations are subject to a more extreme version of the bias observed in F2 populations with dominant markers. Additionally, the inclusion of more parents increases the potential number of segregation distortion loci (SDLs) segregating in the population.

We present here a computationally efficient estimator of recombination fractions in the presence of segregation distortion. The proposed estimator is based on the theory of M-estimators (Huber 1964), an important class of robust estimators. Robust estimators are intended to be reliable and reasonably efficient even when modelling assumptions are incorrect, as in the case of segregation distortion.

We focus on estimation for the F2 population as it is extremely common, and for the Multiparent Advanced Generation InterCross (MAGIC) population as it is extremely vulnerable to the effects of segregation distortion. We model the segregation distortion by dividing the population on the basis of the genotype at the SDL. The resulting subpopulations are genetically different (Farr et al. 2011), but can be weighted and recombined to produce an estimator that incorporates data from a marker closely linked to the SDL. We characterize its behavior through simulations, and apply it to a MAGIC population with a known SDL to construct a map for a chromosome containing several hundred markers.

Methods and materials

Statistical model

Many map construction algorithms are based on estimates of recombination fractions between all markers. Any bias in these estimates, such as that due to segregation distortion, will lead to bias in the final genetic map. Hence we focus our attention on the estimation of recombination fractions in the hopes of reducing or eliminating this bias. For two markers M_1 and M_3, we would like to estimate the recombination fraction r_{13} between them, where bias may be introduced by the presence of an additional SDL M_s lying between M_1 and M_3.

We parametrize our model by the recombination fractions r_{1s} and r_{s3}, as well as the parameter t which measures the strength of distortion due to the SDL. While M_s is not necessarily genotyped, we assume that there is a genotyped

marker M_2 which lies between M_1 and M_3, and is tightly linked to M_s. As our model of segregation distortion will depend on the genetic distances between the flanking markers M_1 and M_3 and the SDL, the fact that M_2 is close to M_s means that we should be able to replace r_{1s} by r_{12} and r_{s3} by r_{23} with negligible error.

We use the term genotype to refer to the pair of alleles inherited at each locus. For all three loci, and indeed for all markers genotyped in the given population, we assume that there are G genotypes observed, and that without segregation distortion each progeny has probability $\frac{1}{G}$ of carrying each genotype at each marker.

We assume a lack of crossover interference, so that crossovers follow a Poisson process and the Haldane map function can be applied. We refer to the undistorted probability model as $\mathbb{P}_u$, and in particular, the probability of a multi-locus genotype x, y, z is

$$p_{xyz}(r_{12}, r_{23}) \stackrel{\text{def}}{=} \mathbb{P}_u(M_1 = x, M_2 = y, M_3 = z).$$

Without crossover interference these probabilities depend only on the recombination fractions, and they are well documented for common experimental designs (Wu et al. 2007; Teuscher and Broman 2007). For convenience, we will abbreviate $p_{xyz}(r_{12}, r_{23})$ as p_{xyz}.

In order to specify the distorted model $\mathbb{P}_d$, we introduce a parametrized distribution $g_y(t)$ characterizing the distortion. Here y is one of G different genotypes, and t is the parameter measuring the strength of distortion. Typically $t = 1$ corresponds to the case of no distortion. Under this model, the probability of a multi-locus genotype x, y, z is

$$\mathbb{P}_d(M_1 = x, M_2 = y, M_3 = z) = G p_{xyz} g_y(t). \tag{1}$$

Cheng et al. (1996) give examples of g_y for gametic and zygotic segregation distortion which we will describe in more detail later. Our model has the property that

$$\mathbb{P}_d(M_2 = y) = g_y(t) G \sum_{x,z} p_{xyz} \tag{2}$$

$$= g_y(t) G \mathbb{P}_u(M_2 = y) = g_y(t). \tag{3}$$

Further, for all genotypes x, y and z,

$$\mathbb{P}_d(M_1 = x, M_3 = z | M_2 = y) = \mathbb{P}_u(M_1 = x, M_3 = z | M_2 = y). \tag{4}$$

Applying Eqs. 2 and 4 together gives

$$\mathbb{P}_d(M_1 = x, M_3 = z) = \sum_y \mathbb{P}_u(M_1 = x, M_3 = z | M_2 = y) g_y(t). \tag{5}$$

The genotypes at M_1 and M_3 are modelled by a mixture distribution with G components and weights $g_y(t)$. The only difference between the distorted and undistorted models is these weights, which are all fixed in the undistorted model and vary according to the unknown parameter t in the distorted model.

In practice, for a given triplet of markers we will observe some empirical proportions of lines with multi-locus genotypes x, y and z. We will denote these proportions as $\hat{p}_{xyz}$, calculated by dividing the number of observed genotypes n_{xyz} by the total number of lines n. Similarly, the marginal proportions at individual markers can be denoted with dots in place of the genotypes at other markers, so that $\hat{p}_{x..}$ is the proportion observed to have genotype x for M_1. These empirical proportions will always be assumed to include the effects of segregation distortion.

Our choice of Eq. 1 as statistical model makes minimal changes to the standard model, attempting to incorporate the distortion parameters in a manner which is easily interpretable in terms of the inflation of certain genotype frequencies. Conveniently, this allows a ready interpretation of recombination fractions which may be difficult with more complex models.

F2 population

Consider an F2 population of n individuals with three codominant markers M_1, M_2 and M_3 genotyped. At each marker we denote the two homozygotes by a and b, and the heterozygote by h. Note that according to the notation defined above, $G = 4$ even though we only consider three genotypes, as we are collapsing the two equally probable phased heterozygotes *ab* and *ba* into a single category which is twice as probable. We assume that these three markers are in the correct order. We are interested in estimating the recombination fraction between M_1 and M_3, assuming that M_2 is the SDL and that any apparent segregation distortion at M_1 or M_3 is due only to their linkage with M_2.

In order to estimate recombination fractions under the distorted model, we focus on the empirical values of the quantities in the weighted sum in Eq. 5. We decompose $\hat{p}_{x.z}$ into three parts and weight each part by the ratio of the marginal probabilities of the SDL genotype under the undistorted and distorted models. This defines the nine values

$$s_{xz} = \hat{p}_{xaz} \frac{\frac{1}{4}}{\hat{p}_{.a.}} + \hat{p}_{xhz} \frac{\frac{1}{2}}{\hat{p}_{.h.}} + \hat{p}_{xbz} \frac{\frac{1}{4}}{\hat{p}_{.b.}}. \tag{6}$$

Next we replace the denominators by their expectations, and collect the error from doing so in the remainder term ϵ_{xz}.

$$s_{xz} = \frac{\hat{p}_{xaz}}{4p_{.a.}} + \frac{\hat{p}_{xhz}}{2p_{.h.}} + \frac{\hat{p}_{xbz}}{4p_{.b.}} + \epsilon_{xz} \tag{7}$$

$$\stackrel{\text{def}}{=} t_{xz} + \epsilon_{xz}. \tag{8}$$

The values $\mathbf{s} = \{s_{xz}\}$ are functions of the 27 random variables $\mathbf{n} = \{n_{xyz}\}$ which are counts of the observed multi-locus genotypes. These have a joint multinomial distribution, making the distribution of $\mathbf{s}$ difficult to determine. However, as shown in "Proof of Eq. 9" in Appendix their expectations are

$$\mathbb{E}(s_{xz}) = \mathbb{P}_u(M_1 = x, M_3 = z) + \mathbb{E}(\epsilon_{xz}). \tag{9}$$

The first term is a function only of r_{13}, deriving from the standard undistorted model for an F2 population, and fully detailed in Wu et al. (2007). The second term is an additional error whose distribution is analytically intractable. However, we examine its distribution through simulation in "Simulation of error terms" in Appendix. Assuming the term ϵ_{xz} in Eq. 7 is small, we can estimate r_{13} based on the empirical weighted sums in Eq. 6. The probability mass function $f(\mathbf{n}; r_{13})$ under the undistorted model is a multinomial and straightforward to characterize (Wu et al. 2007). From Eq. 9, we know that

$$\mathbb{E}(s_{xz}) \simeq \frac{\mathbb{E}_u(n_{x.z})}{n}.$$

We take as our estimator the robust M-estimator (RM)

$$\hat{r}_{13} = \underset{0 \le r_{13} \le \frac{1}{2}}{\operatorname{argmax}} f(n\mathbf{s}, r_{13}), \tag{10}$$

where we substitute $n\mathbf{s}$ for $\mathbf{n}$.

The M-estimator (maximum likelihood-type estimator) was originally proposed by Huber (1964). If $x_1, \ldots, x_n$ are an independent and identically distributed sample from the density $f(x; \theta)$, then any estimator $\widehat{\theta}$ of the form

$$\widehat{\theta} = \arg\min_{\theta} \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right)$$

is said to be an M-estimator. If $\rho(x, \theta) = -\log f(x; \theta)$ then the usual maximum likelihood estimator (MLE) is obtained. The sample mean can be recovered by setting

$$\rho(x, \theta) = \frac{1}{2}(x - \theta)^2.$$

Other common estimators that are also M-estimators include the sample median and least squares regression estimator. Under suitable regularity conditions such estimates converges to $\arg\min_\theta \mathbb{E}[\rho(X, \theta)]$. Further theory relating to M-estimators can be found in Hampel et al. (2005) and Huber (2009).

We note that the M-estimator described here has some similarity to the Horvitz–Thompson estimator (1952) commonly used in survey analyses. Essentially, we stratify by the marker allele at the SDL and weight our data by the prevalence of each allele to reduce the influence of segregation distortion on our estimate of r_{13}.

For simplicity we have presented our estimator with reference to codominant markers, but the extension to the case where the markers are all dominant is straightforward. If a, b and c are 0 or 1 then let $\hat{p}_{abc}$ denote the empirical proportion of progeny lines which had multi-locus marker genotype a, b, c. Let p_{abc} be the corresponding theoretical proportion. Assume that founder b is dominant at marker M_2. Then the value of s_{xz} corresponding to $M_1 = 0$ and $M_3 = 1$ is

$$s_{01} = \frac{\hat{p}_{001}}{4\hat{p}_{.0.}} + \frac{3\hat{p}_{011}}{4\hat{p}_{.1.}}. \tag{11}$$

If founder a was dominant at marker M_2, then we would instead define

$$s_{01} = \frac{3\hat{p}_{001}}{4\hat{p}_{.0.}} + \frac{\hat{p}_{011}}{4\hat{p}_{.1.}}. \tag{12}$$

We can similarly define the values s_{00}, s_{10} and s_{11}. If these are written as a vector $\mathbf{s}_{\text{dom}}$ then we can construct the estimator as in Eq. 10 by replacing $\mathbf{s}$ by $\mathbf{s}_{\text{dom}}$, and the probability mass function f by the corresponding version for dominant markers.

Note that for two dominant markers M_1 and M_3, even the MLE of r_{13} has no analytic form and must be computed numerically. Although the bias of the MLE is asymptotically zero, for a finite population size the bias will be non-zero and cannot be computed explicitly. Similarly, for our estimator we can only characterize the distribution of the bias through simulation rather than analytically.

MAGIC population

We can use the same type of estimator to estimate recombination fractions in MAGIC populations. We now consider an eight-parent MAGIC population of n inbred individuals with three biallelic markers M_1, M_2 and M_3. Since the progeny are inbred lines, all genotypes are homozygous and we can represent them in terms of single alleles. Marker genotypes are coded as 0 or 1, and the underlying founder genotypes as $\{A, B, C, D, E, F, G, H\}$. Note that if we wish to apply the same idea as previously we need to be able to estimate the pattern and strength of segregation distortion. The simplest situation where this is possible occurs when

1. In the distorted region all founders are equally represented in the progeny, except for a single founder (without loss of generality assumed to be A) which is over- or under-represented.
2. There is a biallelic marker in the distorted region for which founder A carries the 1 allele, and the other founders carry the 0 allele.

Fortunately this situation is believed to occur often in practice. In particular, translocations are common in major crop plants, are commonly associated with segregation distortion and the resulting populations generally have markers that uniquely identify lines carrying the translocation (Tsilo et al. 2008; Farr et al. 2011; Gill et al. 2011; Xie et al. 2012).

In the MAGIC case we have the additional complication that different orderings of the founder lines in the first generation cross (known as funnels) result in populations which are genetically distinct. Typically a large number of different funnels are used in a single population (Bandillo et al. 2013), and probabilities of inheriting haplotypes from different founders vary among funnels (Broman 2005). Let F be the collection of all funnels, $\hat{p}_{xyzf}$ be the proportion of lines from funnel f having genotype x, y and z at the markers of interest, and p_f be the proportion of lines from funnel f. Let $\mathbb{P}_{f,u}$ and $\mathbb{P}_{f,d}$ refer to the probability models for lines drawn from funnel f, under the undistorted and distorted models respectively.

Similar to the F2 population, we decompose $\hat{p}_{x.z}$ into eight components and reweight each component. Unlike the F2 population we must also decompose by funnel. This results in the 64 values of the form

$$s_{xz} = \sum_{f\in F} p_f \left(\frac{\hat{p}_{xAzf}}{8\hat{p}_{.A.f}} + \sum_{y\neq A} \frac{7\hat{p}_{xyzf}}{8\left(1-\hat{p}_{.A.f}\right)} \right).$$

Replacing the denominators with their expectations and accumulating the error in ϵ_{xz} gives

$$\begin{aligned} s_{xz} &= \sum_{f\in F} p_f \left(\frac{\hat{p}_{xAzf}}{8p_{.A.f}} + \sum_{y\neq A} \frac{7\hat{p}_{xyzf}}{8\left(1-p_{.A.f}\right)} \right) + \epsilon_{xz} \\ &\stackrel{\text{def}}{=} t_{xz} + \epsilon_{xz}. \end{aligned}$$

Excluding the error term, the expectation is

$$\begin{aligned} \mathbb{E}(t_{xz}) &= \sum_{f\in F} p_f \left(\frac{1}{8}\mathbb{P}_{f,d}(M_1 = x, M_3 = z | M_2 = A) \right. \\ &\qquad \left. + \sum_{y\neq A} \frac{7p_{xyzf}}{8\left(1-p_{.A.f}\right)} \right) \\ &= \sum_{f\in F} p_f \left(\mathbb{P}_{f,u}(M_1 = x, M_2 = A, M_3 = z) \right. \\ &\qquad \left. + \frac{7}{8}\mathbb{P}_{f,d}(M_1 = x, M_3 = z | M_2 \neq A) \right). \end{aligned} \tag{13}$$

It can be shown (see "Proof of Eq. 14" in Appendix) that if $p_f = |F|^{-1}$, so that an equal proportion of lines come from each funnel, then it is approximately true that

$$\mathbb{E}(t_{x,z}) = \sum_{f\in F} |F|^{-1}\mathbb{P}_{f,u}(M_1 = x, M_3 = z). \tag{14}$$

This approximation is equally valid when M_2 no longer lies between M_1 and M_3, and is essentially the same result as in Eq. 9. Hence as in the F2 case, we expect the value of $\mathbb{E}(s_{xz})$ to be primarily affected by variation in r_{13} rather than any of the other parameters, and again we can construct the M-estimator in a similar manner to that defined in Eq. 10.

Simulation studies

We performed three types of simulations to characterize the performance of the RM for F2 and MAGIC populations. First, we considered the distribution of the error term for an F2 population, to demonstrate that its magnitude was sufficiently small for the approximation to be reasonable. Second, we compared the bias and variability of the RM estimator in an F2 population with those for the EM estimator (Cheng et al. 1996) and the uncorrected estimator. Third, we compared the bias of the RM estimator to that of the uncorrected estimator for a MAGIC population. The simulations of error are described further in "Simulation of error terms" in Appendix.

For the F2 population three dominant markers M_1, M_2 and M_3 were simulated, with M_2 being the SDL. The markers were assumed to be in the correct order. The aim was to estimate the recombination fraction r_{13} between M_1 and M_3. Parameters varied were the distortion model, distortion strength t, recombination fraction between M_1 and M_2 (r_{12}) and between M_2 and M_3 (r_{23}). Estimation with correction for distortion was performed using the EM algorithm presented in Cheng et al. (1996), with 50 EM iterations. Uncorrected estimation was performed using the MLE for r_{12} and r_{13}. This was obtained by a grid search, with 501 equally spaced values for r_{13}.

We generated data from the three models considered in Cheng et al. (1996): gametic selection in model 1, which results in a viability ratio of $1 : 1 + t : t$ for genotypes $a : h : b$; gametic selection in model 2, which results in a viability ratio of $1 : 2t : t^2$; and zygotic selection in model 3, with a ratio of $1 : 2 : t$. We varied the level of distortion t with values 0.2, 0.6, and 1.0 under model 1, and values 0.3, 1.15 and 2 under models 2 and 3. The recombination fractions r_{12} and r_{23} had values 0.05, 0.275 and 0.5. For each set of models, we generated 1,000 replicates of F2 populations containing 300 individuals. The dominant founder allele was selected randomly for all three markers.

We performed similar simulations to assess the performance of the RM estimator for a MAGIC population. Here we compared the RM estimator against numerical maximum likelihood using a two-dimensional grid search for the parameters r_{12} and r_{23}. We generated 1,000 replicates of

eight-parent MAGIC populations containing 2,000 individuals, genotyped at three biallelic markers M_1, M_2 and M_3, with M_2 being an SDL. We varied three parameters: the recombination fraction between M_1 and M_2 (r_{12}); the recombination fraction between M_2 and M_3 (r_{23}); and the proportion of the population expected to carry the segregation distortion allele ($p_{.A}$). We considered all possible combinations of the values 0.05, 0.25, and 0.50 for the three parameters. For the distortion parameter this corresponds to 0.4, 2 and 4 times as much inheritance of the distorted founder allele (A) as would be expected under Mendelian segregation. A set of 31 segregation patterns for founders at M_1 and M_3 were generated for each combination of parameters, with results averaged over segregation patterns. For M_2, founder A carried allele 1 and the other founders carried allele 0. Individuals were drawn from random funnels, but the same set of funnels was used for every simulated population. Uncorrected estimation was also performed using a numerical grid search to maximize the likelihood; computation of both the uncorrected and the RM estimators has been implemented in the R/mpMap package (Huang and George 2011).

Wheat MAGIC

We calculated the RM estimator described above for markers from the 9K SNP chip (Cavanagh et al. 2013) genotyped on 1,743 inbred progeny from an eight-parent bread wheat MAGIC population. The eight parents (AC-Barrie, Alsen, Baxter, Pastor, Volcani, Westonia, Xiaoyan54, and Yitpi) were crossed in a total of 306 funnels out of the possible 315 funnels. The number of lines per funnel ranged from 1 to 17.

We focused on map construction for Chromosome 2B, as the parent Baxter is known to carry the rye introgression Sr36 (Tsilo et al. 2008; Huang et al. 2012). While this introgression has been previously mapped (Huang et al. 2012) no correction was made at that time for the observed distortion in distances between markers. All analyses were performed using R/mpMap (Huang and George 2011).

For Chromosome 2B, we first estimated recombination fractions without accounting for segregation distortion, using the function 'mpestrf' with default parameters. Markers were grouped using hierarchical clustering, and linkage groups aggregated interactively. Groups of markers corresponding to Chromosome 2B were identified based on previous mapping studies (Cavanagh et al. 2013). We then ordered markers with the function 'mporder', which aims to minimize the number of Anti-Robinson events among the matrix of recombination fraction estimates using the software package R/seriation (Hahsler et al. 2008).

Once a marker ordering was achieved, we used the function 'computemap' to estimate map positions. This converts the matrix $\hat{\mathbf{R}}$ of pairwise recombination fraction estimates to a matrix of genetic distances using the Haldane mapping function. The true matrix of genetic distances is unknown, but it must satisfy a large number of additivity constraints. For example, if we have three markers M_1, M_2 and M_3 in the correct order and x_{ij} is the genetic distance from marker i to j, then $x_{12} + x_{23} = x_{13}$. These constraints can be written as a matrix equation of the form $\mathbf{A}\hat{\mathbf{x}} = f(\hat{\mathbf{R}})$ where $\mathbf{A}$ is a known matrix, f is a known function and $\hat{\mathbf{x}}$ is the vector of genetic distances between adjacent markers. We find an approximate solution for $\hat{\mathbf{x}}$ using non-linear least squares, and this determines the map positions of the markers.

Next, we chose markers that were believed to identify the translocation. These markers were all highly linked, with pairwise recombination fraction estimates <0.005, and were all highly distorted. The Chi squared statistic for distortion was over 140 for all these markers, giving a p value that was numerically equal to 0. All the chosen markers had a segregation pattern that uniquely identified the Baxter founder. For 38 lines it was unclear whether the translocation was present or not; these lines were discarded. The recombination fractions for markers on 2B were then re-estimated using the RM estimator to correct for distortion. We then followed the same steps as above using 'mporder' and 'computemap' to construct a corrected map.

Results

Simulation studies

Our simulations of F2 populations aimed to both assess the magnitude of the approximation made in construction of the RM estimator, and the performance of the estimator in comparison to other approaches. In all cases, the expectations $\mathbb{E}[\epsilon_{xz}]$ were small. More detail can be found in "Simulation of error terms" in Appendix.

We next compared the RM estimator against other approaches, including that which has no correction for distortion. For the F2 population our simulations show that some correction for distortion is clearly necessary. Figure 1c shows the case where data was generated according to model 2, with distortion parameter $t = 0.3$. This was the case where the uncorrected estimation performed worst, with an uncorrected bias of 30 %. The EM algorithm approach outlined in Cheng et al. (1996) outperformed the RM and uncorrected estimates in all cases. It is unbiased and not shown in any figures. Note however that this approach carries a significant computational cost. We compare the RM and uncorrected estimates on the basis of the mean square error (MSE), defined for an estimator $\widehat{\theta}$ of θ as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\theta - \widehat{\theta})^2\right].$$

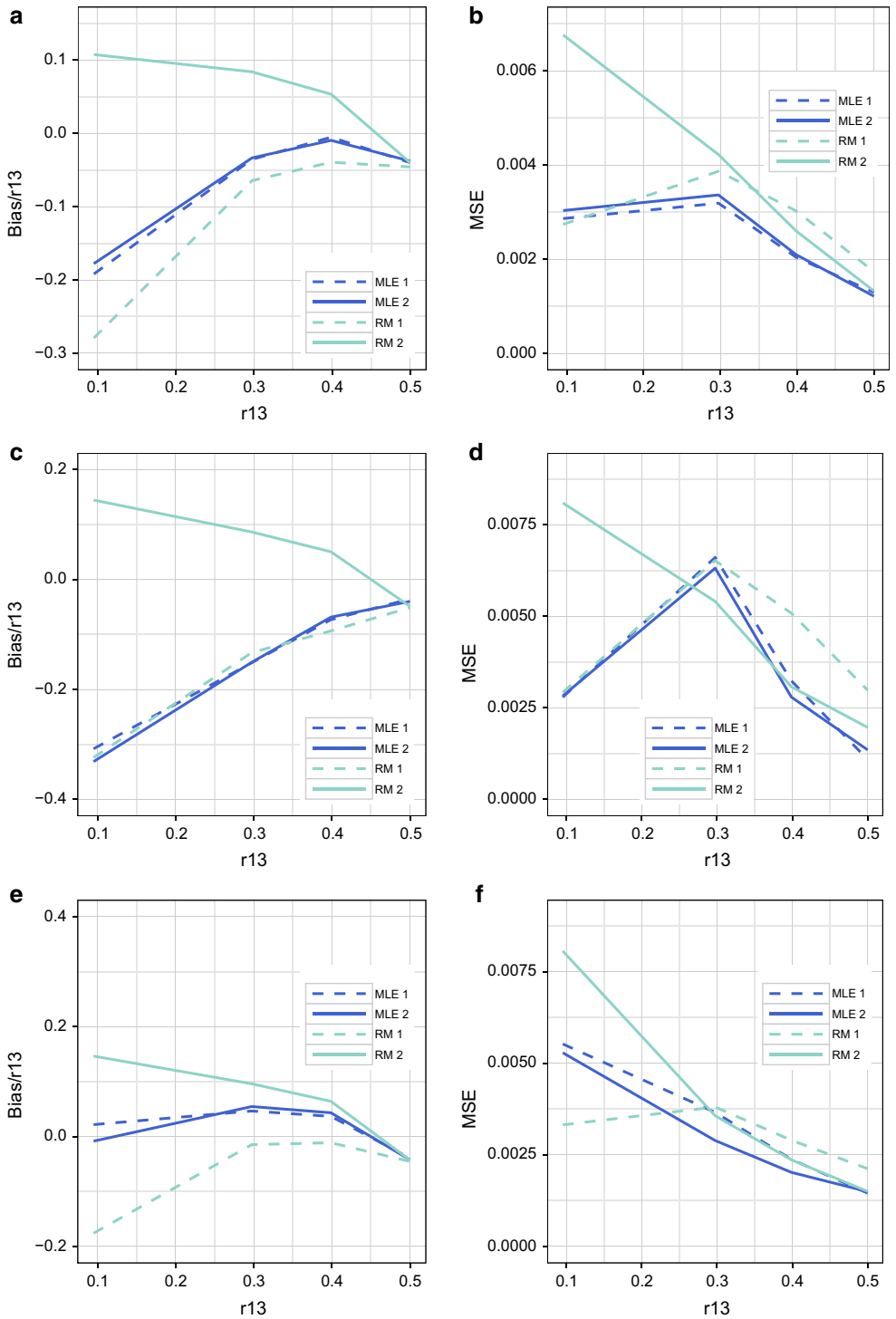

Fig. 1 Bias and mean squared error (MSE) of recombination fraction estimates using the proposed robust M-estimator (RM) and the uncorrected MLE in the presence of segregation distortion for 1,000 replicates of an F2 population of 300 individuals. Selected distortion scenarios are presented, with the suffix *1* denoting founder *a* being dominant at marker M_2 and the suffix *2* denoting founder *b* being dominant. **a** Bias as a proportion of r_{13}, under model 1 with $t = 0.2$. **b** Mean squared error under model 1 with $t = 0.2$. **c** Bias as a proportion of r_{13}, under model 2 with $t = 0.3$. **d** Mean squared error under model 2 with $t = 0.3$. **e** Bias as a proportion of r_{13}, under model 3 with $t = 0.3$. **f** Mean squared error under model 3 with $t = 0.3$

For the F2 population we found the RM estimator to have mixed performance, which depended strongly on the dominant founder at the SDL and the strength of distortion. Where there was an improvement in proportional bias this was typically offset by an increase in variance, and vice versa. This resulted in only small to moderate changes in the MSE overall.

For model 1 (Fig. 1a, b) the performance of the uncorrected and RM estimators was essentially identical for $t = 0.6$ and $t = 1.0$. For $t = 0.2$ we observed a generally higher bias and variance for the RM estimator. The exception was the case where $r_{12} = r_{23} = 0.05$. In this case if founder a was dominant at M_2 then the absolute bias was larger and the variance was smaller than for the uncorrected estimator, resulting in a slightly smaller MSE. If founder b was dominant at M_2 then we observed a smaller absolute bias and higher variance for the RM estimator, also resulting in a smaller MSE.

For model 2 (Fig. 1b, c) the RM and uncorrected estimators performed identically for $t = 1.15$ and $t = 2.0$. For $t = 0.3$ performance was also identical for small recombination fractions if founder a was dominant, although the uncorrected estimator performed slightly better for recombination fractions close to 0.5. If founder b was dominant, the combination of a smaller absolute bias but correspondingly higher variance resulted in MSE similar in size to that of the uncorrected estimator.

For model 3 the RM and uncorrected estimators performed identically for $t = 1.15$ and $t = 2.0$. For $t = 0.3$ the bias of the RM estimator was larger than the bias of the uncorrected estimator, regardless of recombination fraction. However, if founder a was dominant at M_2, there was an offsetting decrease in the variance of the RM estimator, resulting in a moderate decrease in the MSE compared to the uncorrected estimator.

For the MAGIC population the performance gain in correcting for segregation distortion was more clear, as uncorrected estimation showed considerable bias for tightly linked markers (Fig. 2).

Indeed, Fig. 2 illustrates the situation where distortion is least problematic. If the distorted founder allele is present at the SDL in 50 % of progeny, the bias of uncorrected estimation increases to 80 % for tightly linked markers. The RM estimator performs slightly worse when the markers are unlinked. However in practice bias is a far greater problem for markers which are closely linked, as they will have the greatest impact on local ordering for map construction.

Wheat MAGIC

We first compared the uncorrected and corrected maps to each other on the basis of chromosome length. The length of the uncorrected map of Chromosome 2B is 470 cM,

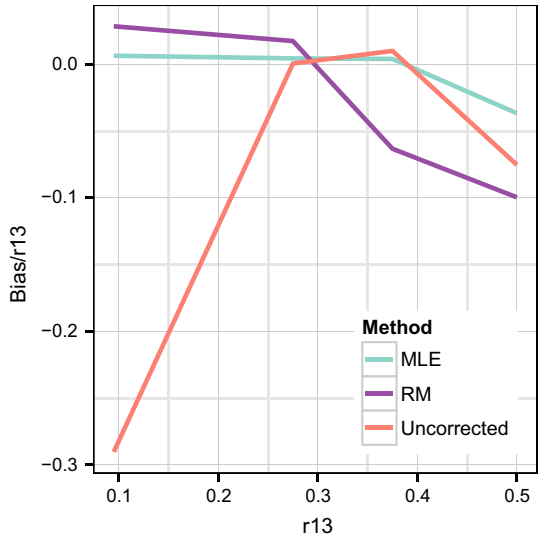

Fig. 2 Bias in recombination fraction as a proportion of the true value r_{13}, for estimates using three approaches—a multidimensional maximum likelihood estimator (MLE); the proposed robust M-estimator (RM); and the uncorrected estimator. All estimates are made for an eight-parent MAGIC population of 2,000 individuals, with the distorted founder allele present in 25 % of progeny at the SDL

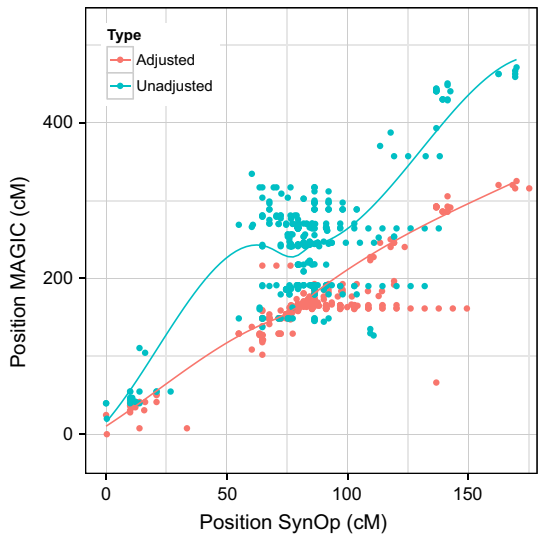

Fig. 3 Map positions of markers from the 9K SNP chip mapped in the SynOp population compared to MAGIC maps corrected and uncorrected for segregation distortion. *Lines* denote smoothed loess curves (span = 0.75) to indicate general trend of positions

which is significantly longer than the estimated length in other maps (Cavanagh et al. 2013). The procedure based on the RM estimator appears to produce more reasonable lengths, resulting in a corrected chromosome size of 335 cM (Cavanagh et al. 2013).

Both maps constructed from the MAGIC population were also compared to a map constructed from the

Synthetic × Opata (SynOp) population which was also genotyped on the 9K SNP chip (Cavanagh et al. 2013). As this population does not contain the Sr36 introgression, we expect markers to be ordered with less bias; hence conflicts between marker positions may indicate errors in the MAGIC map.

Figure 3 compares the positions of the 428 markers mapped in both the SynOp and MAGIC populations to Chromosome 2B. Complete agreement between the maps with respect to order and position would result in a straight line along the diagonal. Differences in positions due to map expansion will result in angled lines; differences in order will result in points displaced from the line. The adjusted map shows good agreement with the SynOp map, with the possible exception of one marker. However, the unadjusted map shows significant disagreement, with markers at the same position in the SynOp map commonly located 150 cM apart in the MAGIC map.

Discussion

The accurate inclusion of distorted markers into the mapping process is important for several reasons. It has been shown that the presence of SDLs can have a positive effect on detection power in QTL mapping (Xu 2008; Zhang et al. 2010). Additionally, these loci may have important biological functions which are not detected if excluded from the genetic map. A number of SDL mapping approaches have been proposed, some of which jointly estimate segregation distortion effects and marker recombination fractions (Zhu et al. 2007), or map QTL and SDL jointly (Xu and Hu 2009). These approaches require a known marker ordering and cannot be used independently of approaches such as those we have described. Additionally, where a large number of markers are available they require a very large number of parameters be estimated jointly using the EM algorithm.

We have presented here an estimator for recombination fractions in the presence of segregation distortion, and applied it to F2 and 8-parent MAGIC populations. These designs represent two different extremes. For an F2 population, estimation of the recombination fraction r_{13} between codominant markers M_1 and M_3 is completely robust against distortion at M_1 or M_3, and no correction for distortion needs to be made. If the distortion occurs at some other marker M_2, then uncorrected estimation will be biased. If all three markers are dominant then the uncorrected MLE can be very biased, but the RM estimator does not exhibit a large increase in performance. Thus, there are relatively few situations in which the RM estimator is necessary in F2 populations; in general the uncorrected estimation will perform as well, with similar computational cost.

For the 8-parent MAGIC population, however, we notice considerable improvement using the RM estimator in certain situations. One particular situation in which we find that the uncorrected MLE has high proportional bias corresponds to specific combinations of founder genotypes for markers M_1 and M_3, which result in a low-information likelihood. In these cases, there is low certainty about estimates of the true distance between markers, especially in the presence of segregation distortion. Including M_2 in the estimation procedure increases the information content of the likelihood and hence improves precision in estimates. While this would in theory be true for other estimators, the notable aspect of the RM estimator is that it does so at no extra computational cost. Importantly, these situations do not occur in F2 populations, explaining the differences in performance.

We found that the EM algorithm approach proposed in Cheng et al. (1996) is highly effective in F2 populations, and significantly outperformed both the RM and ML estimators. However this approach has significant disadvantages in modern studies. First, it requires a model of the type of selection (gametic, zygotic, etc.) occurring at the SDL, which may not be known a priori. Second, it assumes that the SDL lies between two flanking markers, which is potentially different from a model where the SDL lies to one side of both markers. The RM estimator does not require this assumption.

The third disadvantage is the computational burden: while individual applications of the EM algorithm are not overly time-consuming, it must be applied on the order of 10^7 times for estimation of recombination fractions between all markers in a linkage group containing 5,000 markers. To deal with genotyping throughput of this magnitude or larger, such as that produced by genotyping-by-sequencing technology (Elshire et al. 2011), a more computationally efficient approach is required. The RM estimator can be computed using a one-dimensional parametric sweep, and is also suitable for use on highly parallel processors such as Graphics Processing Units (GPUs). It can be computed in an identical amount of time to the MLE from the undistorted probability model. The EM algorithm approach requires two numerical optimisations per iteration of the algorithm, and a number of iterations must be performed. The iterative nature of the algorithm means that it is more difficult to employ GPU acceleration.

For the F2 simulation study the EM algorithm was significantly slower, with over 2 days of computation required. Applying the RM estimator to the same datasets required only 3 h. Both approaches were implemented in Mathematica, and neither implementation was highly optimised, so we consider this a valid comparison. The uncorrected estimation using a grid search was implemented in the R statistical package, and required <2 min of computation time.

A similarly optimised version of the RM estimator would be equally fast, as the computations that must be performed are nearly identical.

The computational differences were more pronounced for the MAGIC population. For the MAGIC simulation study using multi-dimensional maximum likelihood, over a week of computing time was required. By contrast, the RM estimator required only 6 min of computation. Both approaches for the MAGIC population were implemented in highly optimised C code. One drawback of the RM estimator is that as only two markers are used in the model, some pairs of markers will be completely uninformative for the parameter of interest. Taking all the analysis in MAGIC simulation study together, this happened around 7 % of the time.

While our approach does not require either of the model assumptions made by the EM-algorithm approaches, we do assume some knowledge about the SDL. However, this is not difficult to satisfy in practice. Current high-density SNP chips allow for the placement of a marker at sub-centiMorgan distances, and we can reasonably expect the existence of a marker tightly linked to the SDL. In some cases such markers have been previously identified, such as *stm773* for the Sr36 introgression in wheat (Tsilo et al. 2008).

If such a marker is unknown, we suggest the following procedure. First, group the markers by chromosome. In our experience this grouping is unambiguous, regardless of the presence of segregation distortion. Next compute recombination fractions assuming no distortion and order the markers on each chromosome. A visual inspection of the ordered recombination fractions will suggest chromosomes where the ordered recombination fractions are biologically implausible. For these chromosomes perform a chi-squared goodness-of-fit test to assess whether the markers follow the Mendelian segregation ratios. If a large number of apparently tightly linked markers are significantly distorted then this is evidence for segregation distortion. If these markers all have the same segregation pattern then this is further evidence for distortion and may have a biological interpretation. This procedure can identify distorted regions unambiguously; for our wheat MAGIC population the Chi squared test statistic for distortion is regularly above 200. For comparison, a value over 60 represents a p-value on the order of 10^{-15}. Ability to correctly identify distorted regions depends primarily on the strength of distortion and the marker density.

Performance of the RM estimator on real data indicates it can make a substantial contribution to the accuracy of map construction. In practical applications we expect to see the greatest improvement for closely linked markers, which are common with high-density genotyping. Greater improvement was seen for the MAGIC population than the F2; again, this supports use of this estimator in high-density mapping, since the MAGIC populations have greater genetic resolution and diversity than biparental populations. By allowing the inclusion of segregation distortion loci in maps even with large number of markers, we anticipate better positioning of loci in the genome.

In conclusion, it is highly desirable to include distorted markers in the map construction process. Doing so can improve map quality and avoid discarding markers of important biological significance. We propose a robust and computationally efficient estimator of recombination fraction, which is more suitable for map construction using high-density SNP chips than traditional EM algorithm techniques. Further work is necessary to determine the best ways to integrate map construction and QTL analysis in the presence of segregation distortion, particularly for the situation of high-throughput genotyping.

Author contributions R.S. conceived of the idea, performed simulations and analysis of data, and drafted and revised the manuscript. C.R.C. acquired the data and contributed to revisions of the manuscript. B.E.H. contributed to the design of the study, the analysis and interpretation of data, and drafting and revising the manuscript.

Acknowledgments Dr Huang is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE120101127).

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Proof of Eq. 9

$$
\begin{aligned}
\mathbb{E}(s_{xz}) &= \frac{p_{xaz}}{4p_{.a.}} + \frac{p_{xhz}}{2p_{.h.}} + \frac{p_{xbz}}{4p_{.b.}} + \mathbb{E}(\epsilon_{xz}) \\
&= \frac{1}{4}\mathbb{P}_d(M_1 = x, M_3 = z | M_2 = a) \\
&\quad + \frac{1}{2}\mathbb{P}_d(M_1 = x, M_3 = z | M_2 = h) \\
&\quad + \frac{1}{4}\mathbb{P}_d(M_1 = x, M_3 = z | M_2 = b) + \mathbb{E}(\epsilon_{xz}) \\
&= \frac{1}{4}\mathbb{P}_u(M_1 = x, M_3 = z | M_2 = a) \\
&\quad + \frac{1}{2}\mathbb{P}_u(M_1 = x, M_3 = z | M_2 = h) \\
&\quad + \frac{1}{4}\mathbb{P}_u(M_1 = x, M_3 = z | M_2 = b) + \mathbb{E}(\epsilon_{xz}) \\
&= \mathbb{P}_u(M_1 = x, M_2 = a, M_3 = z) \\
&\quad + \mathbb{P}_u(M_1 = x, M_2 = h, M_3 = z) \\
&\quad + \mathbb{P}_u(M_1 = x, M_2 = b, M_3 = z) + \mathbb{E}(\epsilon_{xz}) \\
&= \mathbb{P}_u(M_1 = x, M_3 = z) + \mathbb{E}(\epsilon_{xz})
\end{aligned}
$$

Proof of Eq. 14

It is *not* true that

$$\mathbb{P}_{f,d}(M_1 = x, M_3 = z|M_2 \neq A) = \mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 \neq A)$$

However if we take $p_f = |F|^{-1}$ we can rewrite

$$\sum_{f \in F} |F|^{-1} \mathbb{P}_{f,d}(M_1 = x, M_3 = z|M_2 \neq A)$$

as

$$|F|^{-1} \sum_{f \in F} \sum_{B \leq y \leq H} \mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 = y) \mathbb{P}_{f,d}(M_2 = y|M_2 \neq A)$$

We now make the approximation that $\mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 = y)$ has the same value for all funnels when x, y and z are distinct. Similarly, assume that there is a unique value across all funnels for $x = y, y \neq z$ and another value for $x \neq y, y = z$. Note that this holds exactly for $r_{12} = r_{23} = 0$ and $r_{12} = r_{23} = 0.5$, and it is always true that $\mathbb{P}_{f,u}(M_1 = x, M_3 = x|M_2 = x)$ is funnel-independent. If $x \neq A$ and $z \neq A$, we now have

$$\begin{aligned} |F|^{-1} \Bigg(& \sum_{f \in F} \mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 = x) \mathbb{P}_{f,d}(M_2 = x|M_2 \neq A) \\ & + \sum_{f \in F} \sum_{\substack{B \leq y \leq H \\ y \neq x, y \neq z}} \mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 = y) \mathbb{P}_{f,d}(M_2 = y|M_2 \neq A) \\ & + \sum_{f \in F} \mathbb{P}_{f,u}(M_1 = x, M_3 = z|M_2 = z) \mathbb{P}_{f,d}(M_2 = z|M_2 \neq A) \Bigg) \end{aligned}$$

From our assumption about funnel independence this becomes

$$\begin{aligned} & \mathbb{P}_u(M_1 = x, M_3 = z|M_2 = x) c_x + \sum_{\substack{y \neq A \\ y \neq x, y \neq z}} \mathbb{P}_u(M_1 = x, M_3 = z|M_2 = y) c_y \\ & + \mathbb{P}_u(M_1 = x, M_3 = z|M_2 = z) c_z \end{aligned}$$

Here $\mathbb{P}_u$ refers to the approximate funnel-independent values. However c_x, c_y and c_z are all exactly $\frac{1}{7}$, so this is equal to

$$\frac{8}{7} \mathbb{P}_u(M_1 = x, M_2 \neq y, M_3 = z).$$

Similar arguments can be made for the cases $x = A, z = A$, etc. Substituting this approximation back into Eq. 13 gives the desired result. Note that as we did not use the fact that M_2 was between M_1 and M_3, these approximations are equally as valid if the marker order is M_2, M_1, M_3.

Simulation of error terms

As the expectation and distribution of ϵ_{xz} from Sect. 2.2 are analytically intractable, we characterized them through simulation. We considered the case of three dominant markers M_1, M_2 and M_3. The recombination fractions r_{12} and r_{23} took on values $0.05, 0.1, 0.2, 0.3, 0.4$ and 0.5. The true genotypic probabilities $p_{.a.}$ and $p_{.h.}$ took on values $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{8}{10}$, with the restriction that $p_{.a.} + p_{.h.} \leq 0.9$. The last genotypic fraction $p_{.b.}$ had value $1 - p_{.a.} - p_{.h.}$. All eight possible combinations of dominant founders at the markers were considered. In total, 10,368 different sets of parameters were considered. For each set of parameters, 30,000 F2 populations of 300 individuals were generated. For each population, the values $\epsilon_{00}, \epsilon_{01}, \epsilon_{10}$ and ϵ_{11} were calculated.

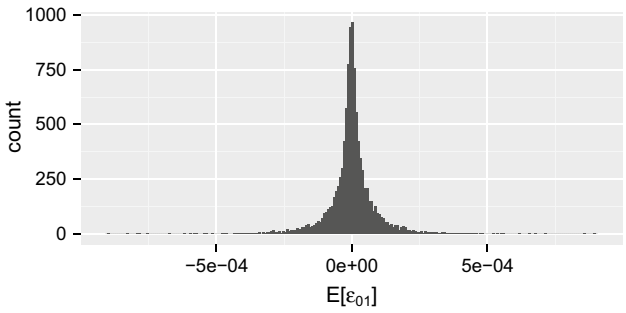

Fig. 4 Histogram of the values of $\mathbb{E}[\epsilon_{00}]$ for three dominant markers simulated in 30,000 F2 populations of size 300. For each population different values of recombination fractions between each pair of markers (M_1 and M_2; M_2 and M_3) were generated from the range 0.05 to 0.5, and true genotypic probabilities ($p_{.a.}$ and $p_{.h.}$) were generated from the range 0.1 to 0.8

Figure 4 shows a histogram of the estimated values of $\mathbb{E}[\epsilon_{00}]$ across all the scenarios considered. Note that Fig. 4 is not a histogram of a distribution, but it shows that the expectation was close to zero in every scenario considered. The behaviour of the other three expectations is similar. In considering whether it is reasonable to assume that $\epsilon_{xz} \simeq 0$, it is therefore sufficient to examine the variance of ϵ_{xz}.

Table 1 lists the five scenarios for which the largest value of $\mathrm{Var}(\epsilon_{01})$ was observed. These scenarios all involve extreme distortion. They also involve specific choices of dominant founders. For example, in the first scenario Eq. 11 implies that

$$\epsilon_{01} = \frac{1}{4} \left(\frac{\hat{p}_{001}}{\hat{p}_{.0.}} - \frac{\hat{p}_{001}}{p_{.0.}} \right) + \frac{3}{4} \left(\frac{\hat{p}_{011}}{\hat{p}_{.1.}} - \frac{\hat{p}_{011}}{p_{.1.}} \right).$$

In this case $p_{.1.} = 1 - p_{.a.} = 0.2$, which is relatively small. So the difference in the second term can potentially be large, whereas the difference in the first term will tend to be small. Now consider the same scenario but with the dominant founder at M_2 being a. From Eq. 12, in this case

$$\epsilon_{01} = \frac{3}{4} \left(\frac{\hat{p}_{001}}{\hat{p}_{.0.}} - \frac{\hat{p}_{001}}{p_{.0.}} \right) + \frac{1}{4} \left(\frac{\hat{p}_{011}}{\hat{p}_{.1.}} - \frac{\hat{p}_{011}}{p_{.1.}} \right).$$

The difference in the first term will tend to be small, and the difference in the second will be potentially large. However

Table 1 Simulation parameters that produced the five largest values of $\mathrm{Var}(\epsilon_{01})$

r_{12}	r_{23}	$p_{.a.}$	$p_{.h.}$	M_1	M_2	M_3	$\mathrm{Var}(\epsilon_{01})$	$\max \lvert\epsilon_{01}\rvert$
0.50	0.05	0.80	0.10	a	b	b	3.96e−03	0.30
0.05	0.50	0.10	0.10	a	a	b	3.95e−03	0.27
0.10	0.50	0.10	0.10	a	a	b	3.70e−03	0.27
0.50	0.10	0.80	0.10	a	b	b	3.67e−03	0.29
0.05	0.40	0.10	0.10	a	a	b	3.45e−03	0.28

Columns M_1, M_2 and M_3 give the dominant founders at those markers. Column $\max |\epsilon_{01}|$ gives the largest observed value of $|\epsilon_{01}|$ across 30,000 simulated populations

it is now multiplied by a factor of $\frac{1}{4}$ rather than $\frac{3}{4}$, and as a result $\mathrm{Var}(\epsilon_{01})$ is expected to be smaller in this case.

When applying the approximation $\epsilon_{xz} \simeq 0$ we actually make four approximations simultaneously. The worst case for each individual approximation is not expected to be representative of the performance when actually applied to recombination fraction estimation. For example, in the first scenario in Table 1, the largest values of $|\epsilon_{00}|, |\epsilon_{10}|$ and $|\epsilon_{11}|$ observed across 30,000 populations were $0.0186, 0.0089$ and 0.11. In the specific population that gave a value of -0.297 for ϵ_{01}, the corresponding values of the other error terms were $0.014, 0.0062$ and -0.022. In general, we observed that scenarios with large values for one of the error terms have very small values for other terms. Hence for nearly all situations we expect the approximation to perform reasonably well.

References

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R, Singh RK, Gregorio G, Redona E, Leung H (2013) Multi-parent advanced generation inter-cross (magic) populations in rice: progress and potential for genetics research and breeding. Rice 6:11

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

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to magic: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11(2):215–221. doi:10.1016/j.pbi.2008.01.002. http://www.sciencedirect.com/science/article/pii/S1369526608000162

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci 110:8057–8062

Cheng R, Saito A, Takano Y, Ukai Y (1996) Estimation of the position and effect of a lethal factor locus on a molecular marker linkage map. Theor Appl Genet 93:494–502. doi:10.1007/BF00417940

Cheng R, Kleinhofs A, Ukai Y (1998) Method for mapping a partial lethal-factor locus on a molecular-marker linkage map of a backcross and doubled-haploid population. Theor Appl Genet 97:293–298. doi:10.1007/s001220050898

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. PLoS ONE 6(e19):379. doi:10.1371/journal.pone/0019379

Farr A, Lacasa Benito I, Cistu L, Jong J, Romagosa I, Jansen J (2011) Linkage map construction involving a reciprocal translocation. Theor Appl Genet 122(5):1029–1037. doi:10.1007/s00122-010-1507-2

Gill BS, Friebe BR, White FF (2011) Alien introgressions represent a rich source of genes for crop improvement. Proc Natl Acad Sci 108(19):7657–7658. doi:10.1073/pnas.1104845108. http://www.pnas.org/content/108/19/7657.short, http://www.pnas.org/content/108/19/7657.full.pdf+html

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90(1):33–38. doi:10.1038/sj.hdy.6800173

Hahsler M, Buchta C, Hornik K (2008) Getting things in order: an introduction to the r package seriation. J Stat Softw 25:3

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2005) The approach based on influence functions. In: Robust statistics. Wiley, New York, pp 100–107

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685

Huang BE, George AW (2011) R/mpmap: a computational platform for the genetic analysis of multi-parent recombinant inbred lines. Bioinformatics 27:727–729

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10(7):826–839. doi:10.1111/j.1467-7652.2012.00702.x

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101

Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken, pp 45–55

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5(7):e1000,551

Liu X, Guo L, You J, Liu X, He Y, Yuan J, Feng Z (2010) Progress of segregation distortion in genetic mapping of plants. Res J Agron 4:78–83

Lorieux M, Goffinet B, Perrier X, León DG, Lanaud C (1995a) Maximum-likelihood models for mapping genetic markers showing segregation distortion. 1. backcross populations. Theor Appl Genet 90:73–80. doi:10.1007/BF00220998

Lorieux M, Perrier X, Goffinet B, Lanaud C, León D (1995b) Maximum-likelihood models for mapping genetic markers showing

segregation distortion. 2. f2 populations. Theor Appl Genet 90:81–89. doi:10.1007/BF00220999

Teuscher F, Broman K (2007) Haplotype probabilities for multiple-strain recombinant inbred lines. Genetics 175:1267–1274

Tsilo TJ, Jin Y, Anderson JA (2008) Diagnostic microsatellite markers for the detection of stem rust resistance gene sr36 in diverse genetic backgrounds of wheat. Crop Sci 48(1):253–261

Wang C, Zhu C, Zhai H, Wan J (2005) Mapping segregation distortion loci and quantitative trait loci for spikelet sterility in rice (*Oryza sativa* l.). Genet Res 86:97–106

Wu R, Ma CX, Casella G (2007) Statistical genetics of quantitative traits: linkage, maps and QTL. Springer, Berlin, pp 52–56

Xie W, Ben-David R, Zeng B, Dinoor A, Xie C, Sun Q, Rder M, Fahoum A, Fahima T (2012) Suppressed recombination rate in 6vs/6al translocation region carrying the pm21 locus introgressed from haynaldia villosa into hexaploid wheat. Mol Breed 29(2):399–412. doi:10.1007/s11032-011-9557-y

Xu S (2008) Quantitative trait locus mapping can benefit from segregation distortion. Genetics 180:2201–2208

Xu S, Hu Z (2009) Mapping quantitative trait loci using distorted markers. Int J Plant Genomics. doi:10.1155/2009/410825

Zhang L, Wang S, Li H, Deng Q, Zheng A, Li S, Li P, Li Z, Wang J (2010) Effects of missing marker and segregation distortion on qtl mapping in f2 populations. Theor Appl Genet 121(6):1071–1082. doi:10.1007/s00122-010-1372-z

Zhu C, Wang C, Zhang YM (2007) Modeling segregation distortion for viability selection i. reconstruction of linkage maps with distorted markers. Theor Appl Genet 114:295–305. doi:10.1007/s00122-006-0432-x