Supplementary Materials

File S1

Details of the algorithm magicMap

HMM

The HMM in multiparental populations has been described in detail for haplotype reconstruction (Zheng *et al.* 2015) and genotype imputation (Zheng *et al.* 2018a). Here we give a short summary. Conditional on phased parental genotypes, offspring are assumed to be independent. The hidden process describes how the ancestral origins change along the two homologous chromosomes within an offspring. It is assumed to follow a continuous time Markov process, which is described by an initial distribution $\boldsymbol{\pi}$ and a rate matrix $\boldsymbol{Q}$. The initial distribution $\boldsymbol{\pi}$ is specified to be the stationary distribution of the rate matrix, so that the prior ancestral origin process does not depend on chromosome direction. According to the theory of continuous time Markov chain, the transition probability matrix from markers t to $t+1$ is in the form of the matrix exponential

$$\boldsymbol{P}\left[x_{t+1}|x_t, d_t\right] = e^{\boldsymbol{Q}d_t},$$

where $x_t = (x_t^m, x_t^p)$ with x_t^m (x_t^p) being the ancestral origin on the maternally (paternally) derived chromosome at marker t, and d_t is the genetic distance in Morgan between markers t and $t+1$.

The calculation of the rate matrix $\boldsymbol{Q}$ depends the relationship between the two homologous chromosomes within an offspring. Following Zheng *et al.* (2015), "depModel" denotes the ancestral origin process along the maternally derived chromosome is completely determined by the ancestral origin process along the paternally derived chromosome, "indepModel" denotes that the two ancestral origin processes are independent, and "jointModel" denotes that two processes are modeled jointly. Here we assume that there is no genetic interference. The rate matrix $\boldsymbol{Q}$ takes a general form for "jointModel". It holds that $\boldsymbol{Q} = \boldsymbol{Q^m} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{Q^p}$ for "indepModel", where $\boldsymbol{Q^m}$ ($\boldsymbol{Q^p}$) denotes the rate matrix for the continuous time Markov processes along the maternally (paternally) derived chromosome, $\otimes$ denotes the Kronecker product, and $\boldsymbol{I}$ is an identity matrix with appropriate dimension. For "depModel", we only need to consider one representative homologous chromosome, and the rate matrix $\boldsymbol{Q} = (\boldsymbol{Q^m} + \boldsymbol{Q^p})/2$ for autosomes or female X chromosomes, and $\boldsymbol{Q} = \boldsymbol{Q^m}$ for male X

chromosomes. The calculations of $\boldsymbol{\pi}$, $\boldsymbol{Q}^m$, $\boldsymbol{Q}^p$, and $\boldsymbol{Q}$ using the available breeding design information have been described (Zheng *et al.* 2014; Zheng 2015; Zheng *et al.* 2018b).

For example, consider four-way RILs by many generations of selfing so that the population becomes fully inbred. We use "depModel" for the ancestral origin process. The rate matrix is given by

$$\boldsymbol{Q} = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} R,$$

and the transition probability matrix is given by

$$P(x_{t+1}|x_t) = e^{\boldsymbol{Q}d_t} = \begin{pmatrix} 1-3r/4 & r/4 & r/4 & r/4 \\ r/4 & 1-3r/4 & r/4 & r/4 \\ r/4 & r/4 & 1-3r/4 & r/4 \\ r/4 & r/4 & r/4 & 1-3r/4 \end{pmatrix}$$

where the state space consists of four inbred founder origins, $1/3$ denotes the probability that current state changes into one of other three equally probable parental origins, the map expansion R denotes the average number of recombination breakpoints per Morgan and $R = 3$ for four-way RILs by selfing, and the scaled recombination fraction r is given by

$$r = 1 - e^{-\frac{4}{3}Rd_t},$$

which is similar to the Haldane's map function. Here we have scaled the recombination fraction so that the maximum value is always 1, independent of the number of founders.

Two-locus linkage analysis

We denote the two loci by subscripts 1 and 2. For the sake of computational efficiency in two-locus linkage analysis but not in the final map refinement stage, we construct the transition probability from the stationary distribution vector $\boldsymbol{\pi}(x)$ instead of the rate matrix $\boldsymbol{Q}$, and allow only "depModel" and "indepModel". For a maternally or paternally derived homologous chromosome, the transition probability matrix is constructed as follow,

$$\boldsymbol{P}\left[x_2^a|x_1^a,r\right]=(1-r)\boldsymbol{I}_n+r\left[\mathbf{1}_n\otimes\boldsymbol{\pi}(x_2^a)\right],a=m,p$$

where $\mathbf{1}_n$ is a vector with all elements being 1, the first term on the right-hand side denotes the case of no recombination, and the second term denotes that x_2^a follows the stationary distribution in the case of recombination. For "depModel", the transition probability matrix $\boldsymbol{P}(x_2|x_1,r)$ is calculated as $\frac{1}{2}\boldsymbol{P}\left[x_2^m|x_1^m,r\right]+\frac{1}{2}\boldsymbol{P}\left[x_2^p|x_1^p,r\right]$ for autosomes or female X chromosome, and $\boldsymbol{P}\left[x_2^m|x_1^m,r\right]$ for male X chromosomes. For "indepModel", it is given by $\boldsymbol{P}\left[x_2^m|x_1^m,r\right]\otimes\boldsymbol{P}\left[x_2^p|x_1^p,r\right]$.

Offspring are assumed to be independent conditional on parental haplotypes, and thus the likelihood l for all offspring is given by

$$\begin{aligned}l(r,h_1,h_2)&=\prod_o P(y_1^o,y_2^o|r,h_1,h_2),\\P(y_1^o,y_2^o|r,h_1,h_2)&=\sum_{x_1,x_2}P(y_1^o|x_1,h_1,\epsilon,\epsilon_F)P(y_2^o|x_2,h_2,\epsilon,\epsilon_F)P(x_2|x_1,r)\pi(x_1),\end{aligned}$$

where r is scaled recombination fraction with maximum being 1, parental haplotype h_i ($i=1,2$) accounts for missing genotypes in founders and unknown parental genotype phases, ϵ (ϵ_F) denotes the probability of allelic errors in offspring (founders), and $P(y_i^o|x_i,h_i,\epsilon,\epsilon_F)$ ($i=1,2$) is the probability of genotype y_i^o of offspring o and it is described in Zheng *et al.* (2015, 2018a). Since the transition probability matrix element $P(x_2|x_1,r)$ is a polynomial function of r, the individual likelihood $P(y_1^o,y_2^o|r,h_1,h_2)$ can also be expressed as a polynomial function of r. The function is linear for "depModel" and quadratic for "indepModel". Because both genotypes y_i^o and parental haplotypes h_i ($i=1,2$) take only discrete values, we calculate individual likelihood $P(y_1^o,y_2^o|r,h_1,h_2)$ as a polynomial function of r for all possible combinations of genotypes and

haplotypes, which can be saved in a table to increase computational efficiency. For each pair of markers, we estimate scaled recombination fraction r and parental haplotypes h_1 and h_2 by maximizing the likelihood $l(r, h_1, h_2)$ using Brent's numerical method (Brent 1973), and calculate the linkage LOD score under the null model of $r = 1$.

Spectral clustering

Given the weight matrix $\boldsymbol{W}$, we group markers based on the spectral clustering algorithm (Shi and Malik 2000) that uses a tool called graph Laplacian. The three variants of graph Laplacians are given by

$$\begin{aligned} \boldsymbol{L} &= \boldsymbol{D} - \boldsymbol{W}, \\ \boldsymbol{L}_{sym} &= \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}, \\ \boldsymbol{L}_{rw} &= \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W} \end{aligned}$$

where $\boldsymbol{D}$ is the degree matrix, a diagonal matrix with ith diagonal element being $\sum_j w_{ij}$. All three graph Laplacians are positive semi-definite and have non-negative and real-valued eigenvalues. The number of connected components of the similarity graph is given by the number of zero eigenvalues, and the components can be specified according to the corresponding eigenvectors. The random walk related $\boldsymbol{L}_{rw}$ is more favored than the symmetric normalized $\boldsymbol{L}_{sym}$ and the unnormalized $\boldsymbol{L}$ (von Luxburg 2007), and thus it is used by default in our method.

In relation to the variants of graph Laplacians, there are also several well-known versions of spectral clustering algorithms. We describe as follows a modified version of spectral clustering for marker grouping with the input being the weight matrix $\boldsymbol{W}$ and the number n_{group} of groups to construct.

1. Compute the graph Laplacian $\boldsymbol{L}_{rw}$ from the weight matrix $\boldsymbol{W}$.
2. Compute the first $2n_{group}$ eigenvalues and eigenvectors of $\boldsymbol{L}_{rw}$, which are ordered with increasing eigenvalues.
3. Select the first $n_{group} \leq n' \leq 2n_{group}$ eigenvectors so that the ratio of the $(n'+1)^{th}$ to the $(n')^{th}$ eigenvalues is a local maximum.
4. Let $\boldsymbol{U}$ be the matrix using the n' eigenvectors as columns.
5. Cluster the rows of $\boldsymbol{U}$ with the agglomerative hierarchical clustering algorithm with linkage being average

and distance metric being cosine.

where in the last step the simple k-means algorithm with the Euclidean distance that is used in the spectral clustering (Shi and Malik 2000) is replaced by the hierarchical clustering with the cosine distance, because the latter has been observed with better performance in our preliminary simulation studies. In addition, we select eigenvectors in steps 2 and 3, in contrast to the usual spectral clustering algorithm that uses only the first n_{group} eigenvectors; the maximum number $2n_{group}$ of eigenvectors for selection is somewhat arbitrary.

Spectral ordering

In comparison with spectral clustering, spectral ordering uses a single eigenvector of graph Laplacian for ordering vertices of a connected similarity graph. Let $\boldsymbol{q} = (q_1, q_2, ..., q_n)$ be the relative ordering of n markers in a given linkage group, subject to rescaling and shifting. The goal of spectral ordering is to place similar markers in adjacent orders and to place dissimilar markers far apart. Formally, it minimizes the following weighted sum of squares

$$f(\boldsymbol{q}) = \frac{1}{2}\sum_{i,j}(q_i - q_j)^2 w_{ij} = \boldsymbol{q}'\boldsymbol{L}\boldsymbol{q},$$

subject to the constraints $\boldsymbol{q}'\mathbf{1} = 0$ and $\boldsymbol{q}'\boldsymbol{q} = 1$, where the prime $'$ denotes transpose, and $\mathbf{1} = (1, ..., 1)$. The solution $\boldsymbol{q}$ satisfies the eigenvalue equation $\boldsymbol{L}\boldsymbol{q} = \lambda\boldsymbol{q}$ under the second constraint. The trivial eigenvector corresponding to eigenvalue $\lambda = 0$ is $\boldsymbol{q} = \mathbf{1}$. The markers are ordered by the element values of the Fiedler vector, the eigenvector associated with the non-zero smallest eigenvalue (Fiedler 1973, 1989).

Ding and He (2004) described a modified version using the weighted constraints $\boldsymbol{q}'\boldsymbol{D}\mathbf{1} = 0$ and $\boldsymbol{q}'\boldsymbol{D}\boldsymbol{q} = 1$, and the Fiedler vector satisfies $\boldsymbol{L}_{rw}\boldsymbol{q} = \lambda\boldsymbol{q}$. The authors have shown that the weighting leads to better ordering because it tends to keep vertices with more edges in the middle. By default, we use the modified spectral ordering with graph Laplacian $\boldsymbol{L}_{rw}$, since we also use $\boldsymbol{L}_{rw}$ in the spectral clustering. There are no noticeable differences between the normalized $\boldsymbol{L}_{rw}$ and $\boldsymbol{L}_{sym}$ in our preliminary studies.

Change of marginal likelihood

We calculate the marginal likelihood for an individual, since offspring are assumed to be independent given the current phased parental genotypes. Let x_t be the latent ancestral origins at locus t, y_t be the observed

genotype data, and $y_{(t+1):t+w}$ be the observed genotypic data in the window of loci from $t+1$ to $t+w$. The standard forward algorithm recursively calculates the joint probability $P(x_t, y_{1:t})$ forwardly for $t = 1, ..., M$, where M is the number of markers. The backward algorithm recursively calculates the conditional probability $P\left[y_{(t+1):M}|x_t\right]$ backwardly for $t = M, ..., 1$ with initial $P\left[y_{(M+1):M}|x_M\right] = 1$. The marginal likelihood is given by

$$P(y_{1:M}) = \sum_{x_t} P(x_t, y_{1:t}) P\left[y_{(t+1):M}|x_t\right]$$

since $P\left[y_{(t+1):M}|x_t, y_{1:t}\right] = P\left[y_{(t+1):M}|x_t\right]$ according to the Markov approximation.

The above calculation of $P(y_{1:M})$ is valid for any locus t in a given genetic map, which enables a fast way to calculate the change $\Delta logl$ of log marginal likelihood. We update inter-marker distances one by one from left to right along chromosomes, and update local ordering by sliding a small window along chromosomes. We first perform the backward algorithm for all offspring. Synchronizing with an update window sliding along chromosomes, we erase the backwardly calculated $P\left[y_{(t+1):M}|x_t\right]$ and meanwhile update the forwardly calculated $P(x_t, y_{1:t})$ up to the rightmost locus t of the window. For a given update window, we can calculate the log marginal likelihood for a proposal map and thus $\Delta logl$ by only re-calculating the forward probabilities for the loci within the window.

Running setups of map construction packages

magicMap

The Mathematica command line used for magicMap is given by

```
magicMap[magicsnpfile,model,popdesign,ngroup,options]
```

where the `magicsnpfile` specifies the input genotypic data. For simulated data, we set `model` to be "indepModel" for the F2, "jointModel" for the CP, and "depModel" for the 8-way RIL. For the real data, we set `model` to be "jointModel" for the Apple CP, and "depModel" for the Arabidopsis 2-way RIL, the Arabidopsis MAGIC, and the barley multiparental population.

The `popdesign` specifies the breeding design information, which is used to compute the process parameter values of the HMM. `popdesign` is set to be the corresponding simulated pedigree file for all simulated

data, the real apple CP data, and the real barley data. For the real Arabidopsis 2-way RIL, it is set to be {"Pairing", "Pairing", "Selfing", ..., "Selfing"} where "Selfing" is repeated for 8 times. For the AI-RIL, it is set to be {"FullDiallel","RM1-E-1000","RM1-E", ..., "RM1-1E", "Selfing", ...,"Selfing"} where "RM1-E" is repeated for 3 times, and "Selfing" is repeated for 6 times. The ngroup specifies the number of linkage groups.

There are many options for magicMap to specify the details of map construction. We use default option values, except that we set the option isFounderInbred -> False for the CP where two parents are outbred, and we set minLodSegregateBin -> Automatic for the maize MAGIC and the maize EU-NAM.

MSTmap

The command line used for MSTmap is given by

```
MSTmap.exe input.txt output.txt
```

Here input file input.txt specifies genotypic data and some parameter values, and output file input.txt outputs the constructed genetic map. We set distance_function to Haldane, missing_threshold to 1.00, estimation_before_clustering to no, detect_bad_data to yes, and objective_function to ML.

We set population_type to RIL2 for the simulated F2 and to RIL8 for the real *Arabidopsis* 2-way RIL. For grouping from simulated data, we set cut_off_p_value to 10^{-6} for the population size 50 and to 10^{-10} for the other population sizes, and we set no_map_dist to 15.0 and no_map_size to 2. For ordering from simulated data, we use the grouping results of magicMap, and set cut_off_p_value to 2, no_map_dist to 10^6, and no_map_size to 0. For the real *Arabidopsis* 2-way RIL, we set cut_off_p_value to 10^{-12}, no_map_dist to 15.0 and no_map_size to 2.

Lep-MAP3

The command line used for Lep-MAP3 parental genotype calling is given by

```
java -cp ParentCall2  data=geno.post >geno.call
```

The command line used for Lep-MAP3 filtering is given by

```
java -cp Filtering2  data=geno.call dataTolerance=0.001 >geno_f.call
```

For Lep-MAP3 grouping, first run the following command line

```
java -cp SeparateChromosomes2 data=geno_f.call
         lodLimit=threshold sizeLimit=5 >geno_f_grouping.txt
```

and iteratively run the following command until the grouping does not change

```
java -cp JoinSingles2All map=geno_f_grouping.txt
         lodLimit=4 >geno_f_grouping_iter.txt
```

The command line used for Lep-MAP3 ordering is given by

```
java -cp OrderMarkers2 map=geno_f_grouping_iter.txt
          data=geno_f.call numMergeIterations=6
          useKosambi=0 sexAveraged=1 >geno_f.order
```

Here input file `geno.post` specifies genotypic data, output file `geno_f_grouping.txt` saves the grouping, and output file `geno_f.order` saves the constructed genetic map. Here we set the threshold of `lodLimit` to 6 for simulated data, and 8.5 for the real apple CP data. For marker ordering in the simulated data, we use the grouping obtained from magicMap.

mpMap

The R command lines used for mpMap grouping in the simulated 8-way RIL are given by

```
dat.rf <- mpestrf(sim.dat)
grouped <- mpgroup(dat.rf, groups=5,clusterBy="combined", method="average")
```

Here the object `sim.dat` specifies all genotypic data and pedigree information.

The R command lines used for mpMap ordering and spacing in the simulated 8-way RIL are given by

```
dat.rf <- mpestrf(sim.dat))
grouped <- mpgroup(dat.rf, groups=1, clusterBy="combined", method="average")
ordered2<-mporder(grouped,1,type="2")
orderedm1<-mporder(ordered2,1,type="m",mapfx="haldane",window=5,repeats=1)
orderedm1d<-computemap(orderedm1,mapfx = "haldane",maxOffset=40))
```

Here the object `sim.dat` specifies genotypic data for a given linkage group and pedigree information. The grouping is given by magicMap.

Preparation of the real datasets

The marker data for the *Arabidopsis* RIL, the CP, and the barley MAGIC have been filtered by West *et al.* (2006), Gardner *et al.* (2014), and Liller *et al.* (2017), respectively, and they were used directly as the test data. For the *Arabidopsis* MAGIC, the markers with the number of missing founder genotypes being greater than 4 and the markers without known physical locations were deleted, there remain 1228 out of 1259 markers. For the tomato MAGIC, 5 out of 1487 markers were deleted because offspring genotypes at those markers are completely missing. For the maize MAGIC (Dell'Acqua *et al.* 2015), we select 303 out of 529 offspring with eight founders. And we deleted 6700 marker with missing fraction greater than 0.5, 6025 monomorphic markers, and 36 markers with the number of missing founder genotypes being greater than 4, reducing the number of markers from 54234 to 41473. For the marker data of the maize EU-NAM (Bauer *et al.* 2013), only the markers the are present in the genetic map derived by Giraud *et al.* (2014) were selected. For the maize US-NAM and the maize EU-NAM, the parental genotypes were imputed and corrected independently for each biparental family, based on single marker analyses; the inconsistent genotypes for the shared parents were then re-set to be missing.

To study the effects of the number of markers, we extract four sub-datasets by randomly choosing 2000, 5000, 10000, 20000 out of the 41473 markers in the maize MAGIC, and similarly for the 34223 markers in the maize EU-NAM.

Literature Cited

Bauer, E., M. Falque, H. Walter, C. Bauland, C. Camisan, L. Campo, N. Meyer, N. Ranc, R. Rincent, W. Schipprack, T. Altmann, P. Flament, A. E. Melchinger, M. Menz, J. Moreno-Gonzalez, M. Ouzunova, P. Revilla, A. Charcosset, O. C. Martin, and C. C. Schon, 2013 Intraspecific variation of recombination rate in maize. Genome Biology **14**: R103.

Brent, R. P., 1973 *Algorithms for Minimization Without Derivatives*. Courier Corporation.

Dell'Acqua, M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens, G. Magris, A. L. Hlaing, H. H. Aung, H. Nelissen, J. Baute, E. Frascaroli, G. A. Churchill, D. Inze, M. Morgante, and M. E. Pe, 2015 Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea *mays*. Genome Biology **16**: 167–189.

Ding, C. and X. He, 2004 Linearized cluster assignment via spectral ordering. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 30–37, New York, NY, USA, ACM.

Fiedler, M., 1973 Algebraic connectivity of graphs. Czechoslovak Mathematical Journal **23**: 298–305.

Fiedler, M., 1989 Laplacian of graphs and algebraic connectivity. Banach Center Publications **25**: 57–70.

Gardner, K. M., P. Brown, T. F. Cooke, S. Cann, F. Costa, C. Bustamante, R. Velasco, M. Troggio, and S. Myles, 2014 Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3 (Bethesda) **4**: 1681–1687.

Giraud, H., C. Lehermeier, E. Bauer, M. Falque, V. Segura, C. Bauland, C. Camisan, L. Campo, N. Meyer, N. Ranc, W. Schipprack, P. Flament, A. E. Melchinger, M. Menz, J. Moreno-Gonzalez, M. Ouzunova, A. Charcosset, C. C. Schon, and L. Moreau, 2014 Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the Flint and Dent heterotic groups of maize. Genetics **198**: 1717–1734.

Liller, C. B., A. Walla, M. P. Boer, P. Hedley, M. Macaulay, S. Effgen, M. von Korff, G. W. van Esse, and M. Koornneef, 2017 Fine mapping of a major QTL for awn length in barley using a multiparent mapping population. Theoretical and Applied Genetics **130**: 269–281.

Shi, J. B. and J. Malik, 2000 Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**: 888–905.

von Luxburg, U., 2007 A tutorial on spectral clustering. Statistics and Computing **17**: 395–416.

West, M. A. L., H. van Leeuwen, A. Kozik, D. J. Kliebenstein, R. W. Doerge, D. A. St Clair, and R. W. Michelmore, 2006 High-density haplotyping with microarray-based expression and single feature polymorphism markers in arabidopsis. Genome Research **16**: 787–795.

Zheng, C., 2015 Modeling X-linked-linked ancestral origins in multiparental populations. G3 (Bethesda) **5**: 777–801.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2014 A general modeling framework for genome ancestral origins in multiparental populations. Genetics **198**: 87–101.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics **200**: 1073–1087.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2018a Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics **210**: 71–82.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2018b Recursive algorithms for modeling genome blocks in a fixed pedigree. G3 (Bethesda) **8**: 3231–3245.

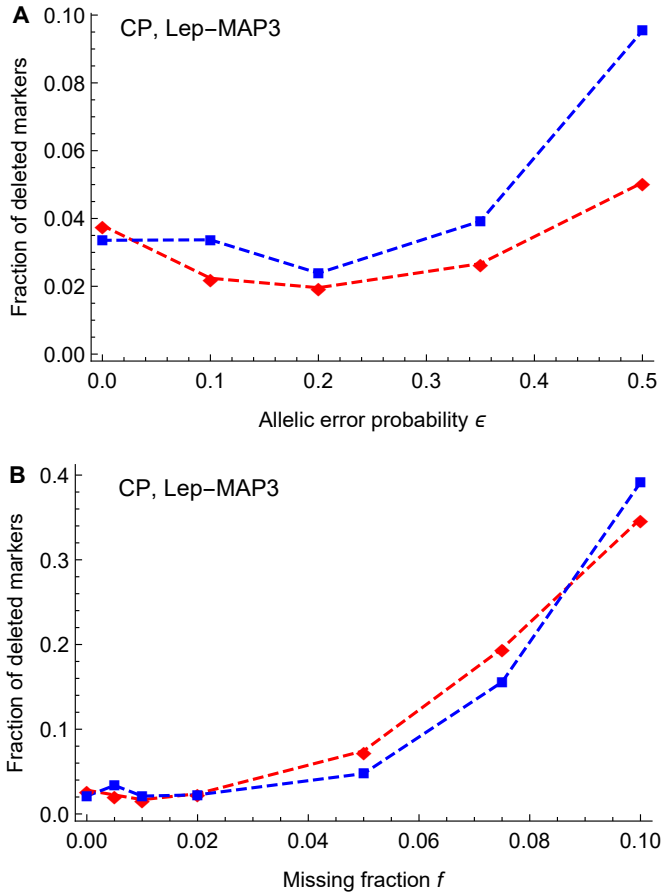

Figure S1 Number of markers deleted during the Lep-MAP3 ordering. The red diamonds (◆) and blue rectangles (■) refer to the CP population sizes of 100 and 200, respectively. (A) Effects of allelic error probability. (B) Effects of missing fraction.

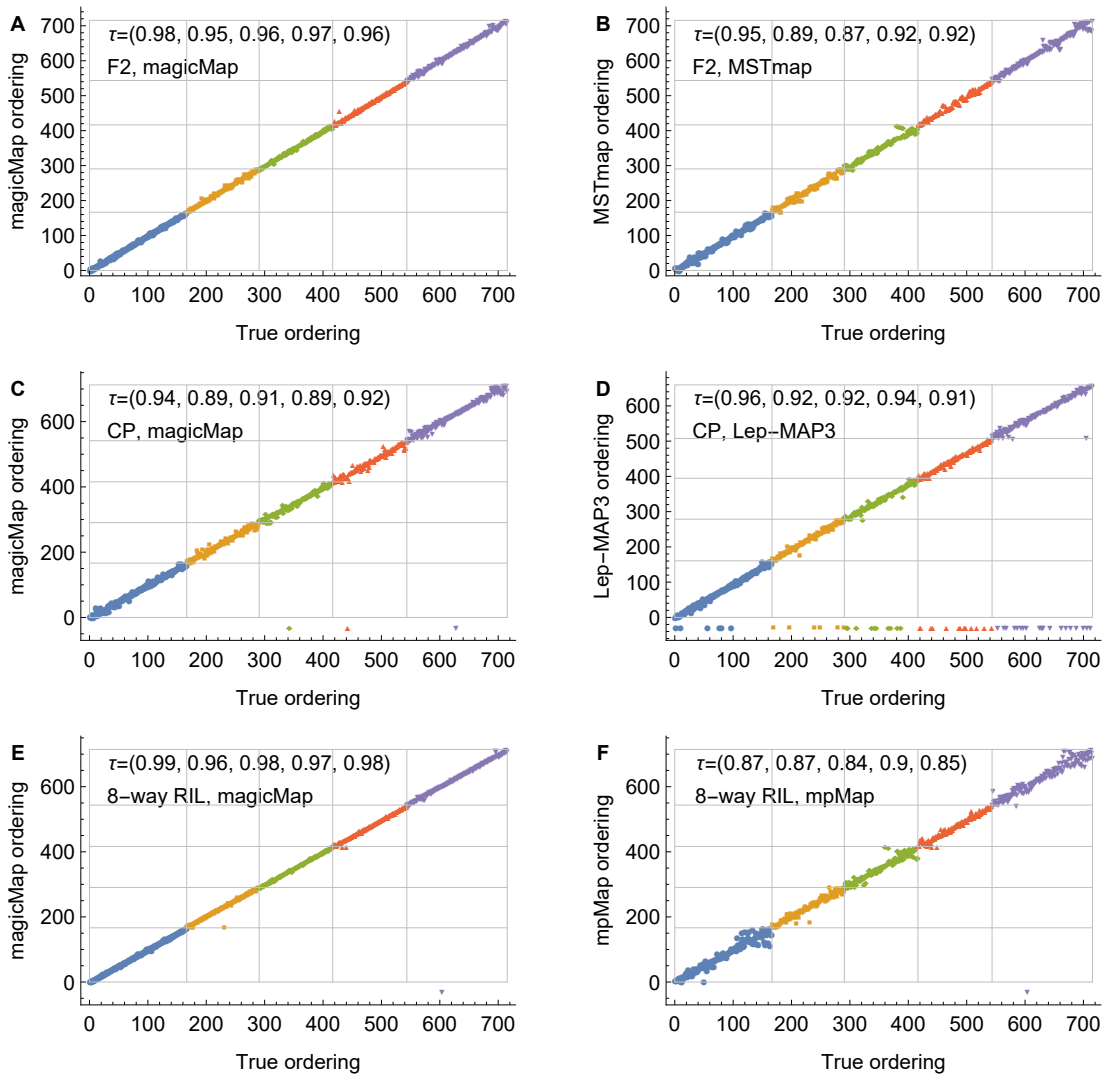

Figure S2 Illustrative comparisons between estimated marker ordering and the true ordering. The left panels refer to magicMap (A, C, and E), and the right panels refer to MSTmap (B), Lep-MAP3(D), and mpMAP (F). The gray grid lines denote the chromosome boundaries, and the dots with negative y-values denote the ungrouped markers. The values of Kendall's tau (τ) for five chromosomes are given for the F2 (A&B), the CP (C&D), and the 8-way RIL (E&F), respectively. The results are obtained from the simulated data with allelic error probability $\epsilon = 0.05$, missing fraction $f = 0.1$, and medium population size 100 for both the F2 and the CP and 400 for the 8-way RIL.

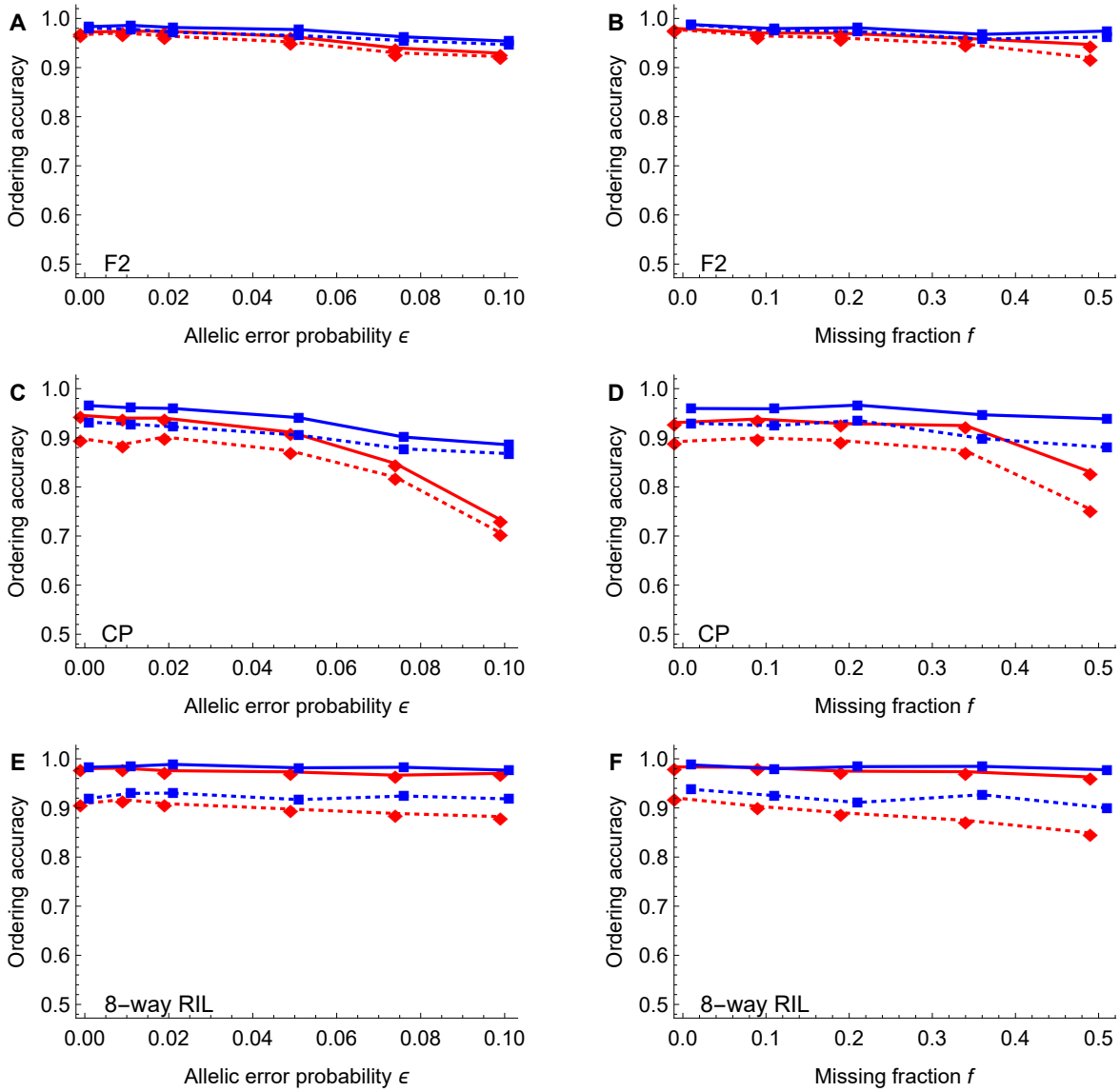

Figure S3 Effect of magicMap ordering refinement. The dotted lines denote the initial spectral orderings, and the solid lines denote the refined ordering. The red diamonds (◆) and blue rectangles (■) refer to medium and large population sizes, respectively. The population sizes are 100 and 200 for the F2 (A&B) and the CP (C&D), and they are 400 and 800 for the 8-way RIL (E&F).

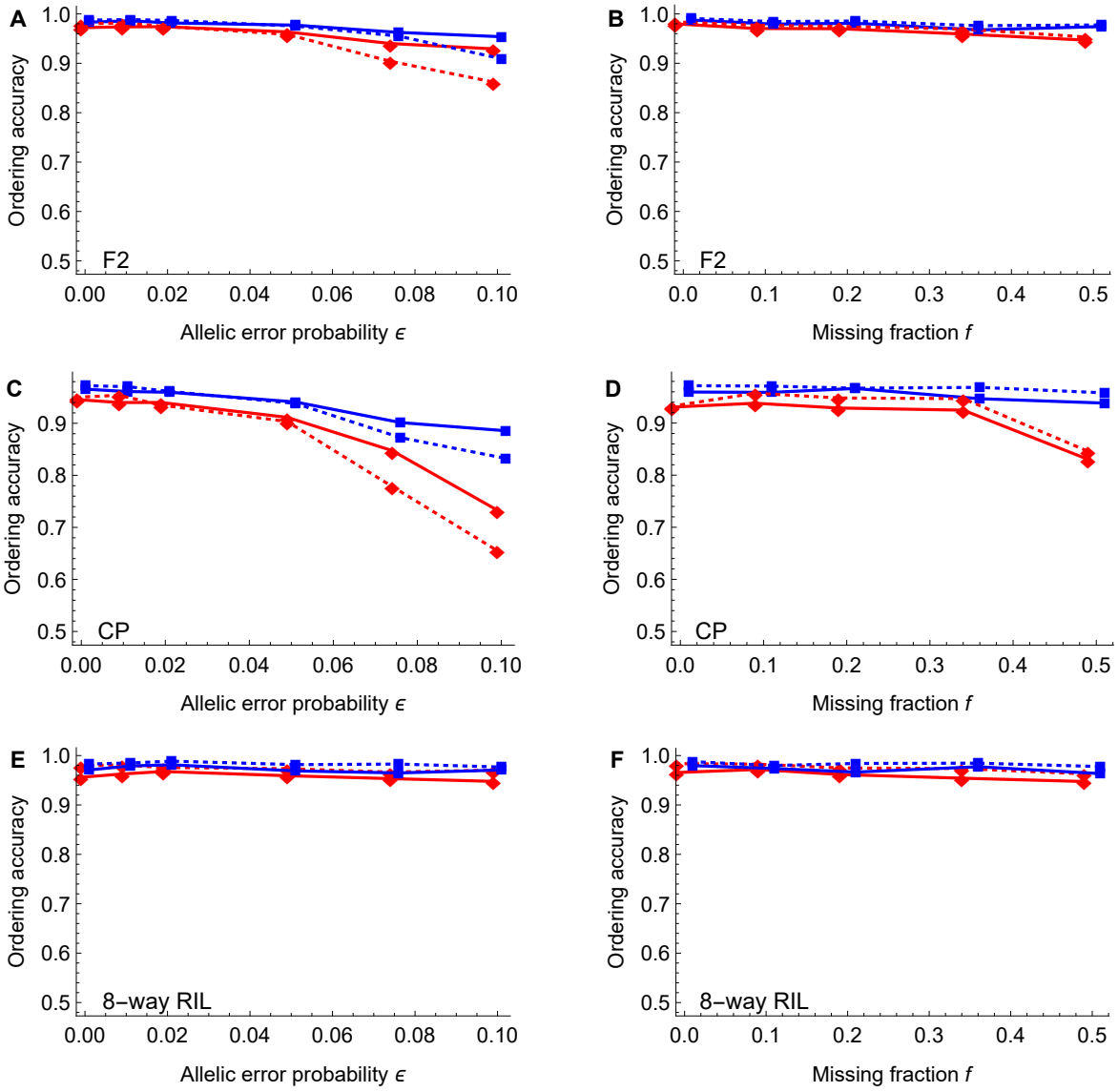

Figure S4 Effect of magicMap error correction on marker ordering. The dotted lines denote the refined orderings without the error correction, and the solid lines denote the orderings with the error correction. The red diamonds (◆) and blue rectangles (■) refer to medium and large population sizes, respectively. The population sizes are 100 and 200 for the F2 (A&B) and the CP (C&D), and they are 400 and 800 for the 8-way RIL (E&F).

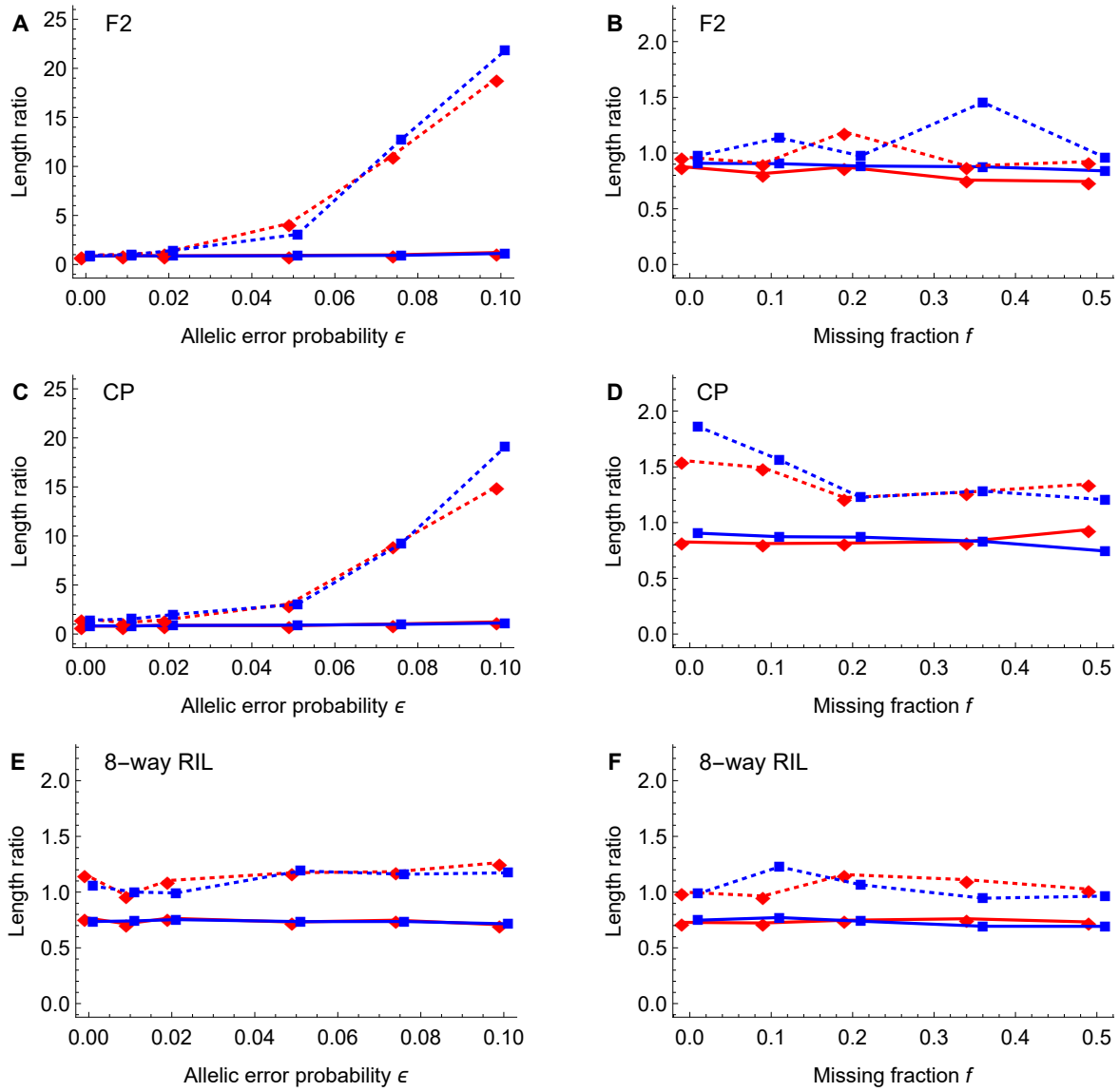

Figure S5 Effect of magicMap error correction on marker spacing. The y-axis denotes the ratio of estimated total chromosome length to true value. The dotted lines denote the refined orderings without the error correction, and the solid lines denote the orderings with the error correction. The red diamonds (◆) and blue rectangles (■) refer to medium and large population sizes, respectively. The population sizes are 100 and 200 for the F2 (A&B) and the CP (C&D), and they are 400 and 800 for the 8-way RIL (E&F).

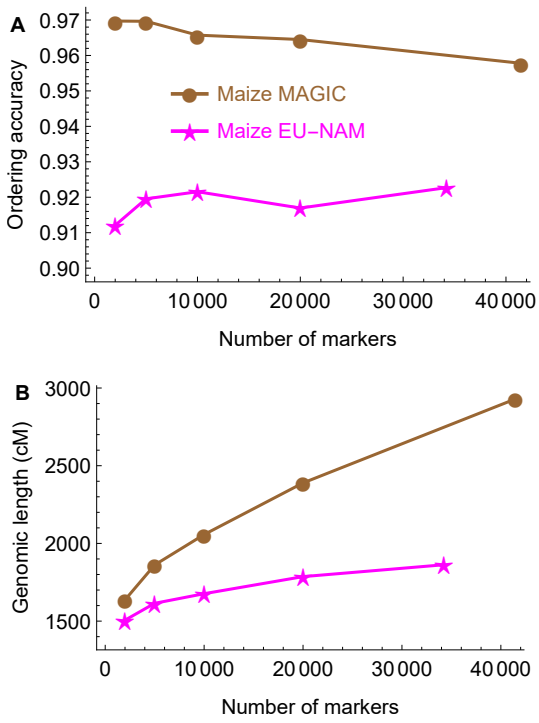

Figure S6 Scaling of magicMap performances with the number of markers in the maize MAGIC and the maize EU-NAM. (A) The ordering accuracy. (B) The genetic map length.

Table S1 List of symbols and their brief descriptions

Symbol	Description
x_t^m	Ancestral origin of the allele on maternally derived chromosome at locus t
x_t^p	Ancestral origin of the allele on paternally derived chromosome at locus t
x_t	Hidden state at locus t. $x_t = (x_t^m, x_t^p)$
d_t	Genetic distance in Morgan between loci t and $t+1$
$P(x)$	Probability that the implicit random variable takes value x
$\boldsymbol{P}[x_2\|x_1, d]$	Transition probability matrix from random variable x_1 to x_2 for a given d value.
"depModel"	Dependent model with $x_t^m = x_t^p$
"indepModel"	Independent model with $P(x_t^m, x_t^p) = P(x_t^m)P(x_t^p)$
"jointModel"	Joint model without assuming a simplified structure of $P(x_t^m, x_t^p)$
$\boldsymbol{I}$	Identity matrix
$\boldsymbol{Q}^m$	Rate matrix of Markov process along the maternal chromosome
$\boldsymbol{Q}^p$	Rate matrix of Markov process along the paternal chromosome
$\boldsymbol{Q}$	Rate matrix for two homologous chromosomes under "jointModel" $\boldsymbol{Q} = \boldsymbol{Q}^m \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{Q}^p$ under "indepModel" $\boldsymbol{Q} = (\boldsymbol{Q}^m + \boldsymbol{Q}^p)/2$ for autosomes or female XX under "depModel" $\boldsymbol{Q} = \boldsymbol{Q}^m$ for male X chromosome under "depModel"
$\boldsymbol{\pi}(x)$	Stationary distribution vector as a function of random variable x
R	Expected number of recombination change-points per Morgan
r	Scaled recombination fraction
f	Fraction of missing offspring genotypes
ϵ	Allelic error probability in offspring
ϵ_F	Allelic error probability in founders
y_i^o	Genotype of offspring o at locus $i = 1, 2$
h_i	Parental haplotype at locus $i = 1, 2$
$l(r, h_1, h_2)$	Two-locus likelihood. Probability of all offspring genotypes at loci 1 and 2
G	Test statistic in G-test
C_{save}	LOD score threshold below which the results of two-locus analysis are not saved
$C_{linkage}$	LOD score threshold in linkage analysis
C_{indep}	LOD score threshold in independence tests
$\boldsymbol{S}$	Similarity matrix $\{s_{ij}\}$
$\boldsymbol{W}$	Sparse similar matrix $\{w_{ij}\}$ obtained from $\boldsymbol{S}$
$\boldsymbol{D}$	Diagonal matrix with ith diagonal element being $\sum_j w_{ij}$
$\boldsymbol{L}$	Un-normalized graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$
$\boldsymbol{L}_{sym}$	Symmetric normalized graph Laplacian $\boldsymbol{L}_{sym} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$
$\boldsymbol{L}_{rw}$	random walk related graph Laplacian $\boldsymbol{L}_{rw} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$
k_{nn}	Number of nearest neighbors
$\Delta logl$	Change of log likelihood by the proposal map
T	Temperature in simulated annealing
T_0	Initial temperature in simulated annealing
T_c	Freezing temperature in simulated annealing
α	Cooling constant. $T \to \alpha T$ if $T > T_c$ and $T \to \alpha^3 T$ if $T \leq T_c$ after each iteration
n_{freeze}	Maximum number of iterations with $T \leq T_c$