Improving Estimates of Genetic Maps: A Meta-Analysis-Based Approach

William C. L. Stewart[*]

Department of Biostatistics, Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, Michigan

Inaccurate genetic (*or linkage*) maps can reduce the power to detect linkage, increase type I error, and distort haplotype and relationship inference. To improve the accuracy of existing maps, I propose a meta-analysis-based method that combines independent map estimates into a single estimate of the linkage map. The method uses the variance of each independent map estimate to combine them efficiently, whether the map estimates use the same set of markers or not. As compared with a joint analysis of the pooled genotype data, the proposed method is attractive for three reasons: (1) it has comparable efficiency to the maximum likelihood map estimate when the pooled data are homogeneous; (2) relative to existing map estimation methods, it can have increased efficiency when the pooled data are heterogeneous; and (3) it avoids the practical difficulties of pooling human subjects data. On the basis of simulated data modeled after two real data sets, the proposed method can reduce the sampling variation of linkage maps commonly used in whole-genome linkage scans. Furthermore, when the independent map estimates are also maximum likelihood estimates, the proposed method performs as well as or better than when they are estimated by the program CRIMAP. Since variance estimates of maps may not always be available, I demonstrate the feasibility of three different variance estimators. Overall, the method should prove useful to investigators who need map positions for markers not contained in publicly available maps, and to those who wish to minimize the negative effects of inaccurate maps. *Genet. Epidemiol.* 31:408–416, 2007. © 2007 Wiley-Liss, Inc.

Key words: map estimation; meta-analysis; linkage map.

Contract grant sponsor: National Institutes of Health; Contract grant number: T32 HG00040; R01 HG002651.

[*]Correspondence to: William C. L. Stewart, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: wstew@umich.edu

Received 19 December 2006; Revised 6 February 2007; Accepted 13 February 2007
Published online 18 April 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20221

INTRODUCTION

Linkage maps are an essential component of many family-based methods of genetic analysis. Multipoint linkage analysis, genotype error detection, and relationship inference all use linkage maps together with the observed genotypes to infer the unobserved pattern of inheritance at the markers. Recently, several methods have demonstrated the growing importance of linkage maps in population genetics as well. For example, Tapper et al. [2005] and Maniatis et al. [2002] used existing linkage maps in conjunction with measures of linkage disequilibrium to estimate the time in generations after one or more population bottlenecks. However, these methods generally assume that the linkage map is known; when in fact, all linkage maps are subject to sampling error.

Inaccurate maps can reduce the power or inflate the type 1 error of multipoint linkage analysis [Barber et al., 2006; Daw et al., 2000; Fingerlin et al., 2006; Halpern and Whittemore, 1999]. Consequently, several authors have attempted to reduce map inaccuracy by estimating linkage maps from large amounts of pedigree data [Broman et al., 1998; George, 2005; Kong et al., 2002, 2004; Matise et al., 2003]. This approach requires either the costly ascertainment of large samples or the pooling of existing samples, which may be complicated by the lack of availability of the original genotype data or the difficulty in sharing such data. In addition, maximum likelihood (ML) methods for estimating the map tend to be computationally demanding, especially in the presence of missing data. The program CRIMAP [Lander and Green, 1987] avoids most of these computational bottlenecks by ignoring allele frequencies and partially informative inheritance information. CRIMAP implements a method that resembles the expectation maximization algorithm [Dempster et al., 1977]. However, CRIMAP does not yield the ML map estimate when genotype data are missing. By contrast, the program LM_MAP [Stewart and Thompson, 2006] uses Markov chain Monte Carlo (MCMC) to estimate the ML estimate (MLE) of the map, whether genotype data are missing or not.

A meta-analysis-based *composite map* is the linkage map that corresponds to the set of distinct markers used among a series of independent map estimates. When the map estimates use the same markers, Stewart and Thompson [2006] defined the composite map estimator (CME) as the weighted average of independent map estimates. Although the CME does avoid the difficulties associated with the sharing of human subjects data, its widespread use is impractical since different studies will tend to use different sets of markers. Therefore, I present: (1) the extended CME (ECME), which combines a set of map estimates whether they share a common set of genetic markers or not; and (2) a variance estimator that accurately estimates the variance of each component of the ECME. The proposed method is implemented in the computer program METAMAP, which is freely available at http://www-personal.umich.edu/~wstew/.

In what follows, I describe the proposed method and assess its performance through the analysis of simulated data. I show that (1) either LM_MAP or the nonparametric bootstrap procedure [Efron and Tibshirani, 1993] can be used to obtain reliable estimates of the variance of each independent map estimate; (2) when genotype data are missing, CRIMAP gives biased estimates of linkage maps and has increased variability relative to LM_MAP; (3) the ECME is almost as efficient as the MLE of the map obtained from the joint analysis of the pooled genotype data; (4) the variance of the ECME is accurately estimated from a parametric bootstrap procedure [Efron and Tibshirani, 1993]; and (5) the proposed method can be used to combine map estimates from individual linkage studies and publicly available maps to improve the estimates of linkage maps commonly used in whole-genome linkage scans. Consequently, the proposed method should protect investigators from the negative effects of inaccurate maps, and is especially helpful when the marker set of a particular study is not a subset of the markers used in any single publicly available map.

METHODS

I assume an invertible map function $\theta(x)$ that relates the recombination fraction to map distance x (in Morgans); and, I express linkage maps as ordered vectors of inter-marker recombination fractions. The data consists of N independent linkage map estimates (which may use different sets of markers) and variance estimates for each estimated recombination fraction within each map. Using only the N map estimates and not the original genotypes, the goal is to estimate the composite map $\mathbf{C}\equiv(\theta_1, \theta_2, \ldots, \theta_{L-1})$, where θ_i is the recombination fraction between markers i and $i+1$ of the L distinct markers used among the N map estimates .

I estimate the composite map by recursion. The starting value may be arbitrary, such as a map derived from published linkage and/or physical maps. Assume that t iterations of the recursion are complete and let $\mathbf{C}^t$ denote the current estimate of the composite map. The next iteration consists of two steps: imputation and weighted averaging. In the first step, I impute the recombination fraction $\tilde{\theta}_i^{(k)}$ and its associated weight $w_i^{(k)}$ for each independent map estimate, $k \in \{1, 2, \ldots, N\}$ and for each of the $L-1$ composite map intervals. In the second step, I update each component of $\mathbf{C}^t$ with the weighted average of the N imputed recombination fractions

$$\theta_i^{t+1} = \frac{\sum_k w_i^{(k)} \tilde{\theta}_i^{(k)}}{\sum_k w_i^{(k)}}.$$

The ECME is the limit of this recursion. In the special case that the independent map estimates use the same L markers, the ECME is attained in one iteration and is equivalent to the CME.

IMPUTING RECOMBINATION FRACTIONS

Let r and s $(r \leq i < i+1 \leq s)$ index the markers in map k that most closely bracket markers i and $i+1$ of the composite map. Let $\hat{\theta}_{rs}^{(k)}$ denote the estimated recombination fraction between markers r and s of map k, and let $[\hat{\sigma}_{rs}^{(k)}]^2$ denote the corresponding estimated variance. For notational convenience, I suppress the dependence of r and s on i and k. Then, using linear interpolation, the imputed recombination fraction between markers i and $i+1$ based on map estimate k is

$$\hat{\theta}_i^{(k)} = \begin{cases} \theta\left[\frac{x(\hat{\theta}_{rs}^{(k)})x(\theta_i^t)}{x(\theta_{rs}^t)}\right] & \text{if interval } [r, s] \quad \text{exists} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, θ_{rs}^{t} is the recombination fraction between markers r and s, according to the current estimate of the composite map; $\theta[\cdot]$ is the map function, and $x(\cdot)$ is its inverse. For each i and k, the interval $[r, s]$ exists if there are markers in map k that span composite map interval $[i, i{+}1]$. Note that whenever the interval $[r, s]$ is also an interval of the composite map, equation (1) reduces to $\tilde{\theta}_{i}^{(k)} = \hat{\theta}_{rs}^{(k)}$.

CONSTRUCTING WEIGHTS

Let $w_i^{(k)}$ denote the weight assigned to imputed recombination fraction $\tilde{\theta}_i^{(k)}$. I define the weight as

$$w_i^{(k)} = \begin{cases} \frac{\hat{\theta}_{rs}^{(k)}(1-\hat{\theta}_{rs}^{(k)})}{[\hat{\sigma}_{rs}^{(k)}]^2 n_{rs}^{(k)}} & \text{if interval } [r,s] \quad \text{exists} \\ 0 & \text{otherwise.} \end{cases}, \qquad (2)$$

where $n_{rs}^{(k)}$ is the number of intervals in the composite map spanned by interval $[r, s]$ of map estimate k. Under the assumption that the information contained in $\hat{\theta}_{rs}^{(k)}$ is distributed evenly along the interval $[r, s]$, the weight $w_i^{(k)}$ measures the information about θ_i contained in map estimate k. This follows from the fact that

$$\eta_{rs}^{(k)} \equiv \frac{\hat{\theta}_{rs}^{(k)}(1-\hat{\theta}_{rs}^{(k)})}{[\hat{\sigma}_{rs}^{(k)}]^2}, \qquad (3)$$

estimates the smallest number of independent, fully informative meioses that, if sampled, could provide the same amount of information about $\theta_{rs}^{(k)}$ as does $\hat{\theta}_{rs}^{(k)}$ [Stewart and Thompson, 2006]. Hence, an equivalent expression for $w_i^{(k)}$ in terms of the average information is

$$w_i^{(k)} = \begin{cases} \frac{\eta_{rs}^{(k)}}{n_{rs}^{(k)}} & \text{if interval } [r,s] \quad \text{exists} \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

Whenever interval $[r, s]$ is also an interval of the composite map, equation (4) reduces to $w_i^{(k)} = \eta_{rs}^{(k)}$. In addition to the computational convenience of (2), its general form is supported by the results of simulated data involving unequal sets of markers (data not shown).

VARIANCE ESTIMATORS FOR MAP ESTIMATES

For the estimated recombination fractions of each independent map estimate, the proposed method requires estimates of their variances. In the case of ML map estimation, the required variance estimates can be obtained from LM_MAP, which uses MCMC to estimate the inverse observed information [Louis, 1982]. Alternatively, for a given map estimation method, a nonparametric bootstrap procedure [Efron and Tibshirani, 1993] could be applied to the families of the linkage mapping data set to obtain the required variance estimates.

An attractive feature of the proposed method is its ability to estimate accurately the variance of each component of the ECME. I assume that each independent map estimate follows a multivariate normal distribution with mean μ_k and variance Σ_k, where μ_k is the k^{th} independent map estimate and Σ_k is its variance with off diagonal elements set to zero. Then, I use the parametric bootstrap procedure [Efron and Tibshirani, 1993] to estimate the variance of each estimated recombination fraction of the composite map.

SIMULATION DESCRIPTION

I simulated data in three settings. In the *complete data* setting, each replicate contains genotype data simulated at 20 markers with four equi-frequent alleles spaced at 5 cm intervals on 90 copies of a 12-member pedigree (Fig. 1). For *study 1* (families 1–30), I used the genotype data at all 20 markers. For *study 2* (families 31–60), I masked the data at markers 2, 4, 6, …, 20. For *study 3* (families 61–90), the genotypes at markers 1, 3, 5, …, 19 were masked. The *pooled data* contain the marker data and families of all three studies. For each study, I simulated a total of 100 replicates.

In the *incomplete data* setting, I introduced missing data in two ways: either the genotypes

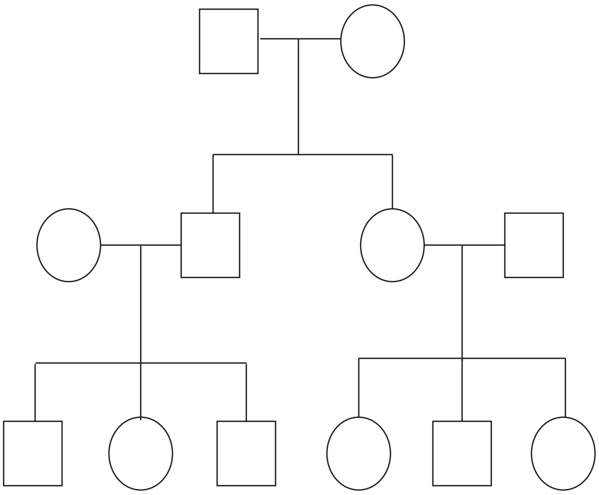

Fig. 1. In total, 27,000 copies of this pedigree structure were used in the *complete data* and *incomplete data* settings.

of grandparents were masked or the genotypes of both parents and grandparents were masked.

In the *screening sets* setting, I used information from a real linkage study [Abkevich et al., 2003] and a real linkage mapping data set [Matise et al., 2003] to simulate three realistic, independent linkage mapping designs. For the first two designs, denoted MSS9 and MSS16, I simulated marker data on 110 families at 10 markers in Marshfield screening set 9 and at 10 (not necessarily equal) markers in screening set 16 [Broman et al., 1998]. These families contain 1,900 individuals with 38% missing data and range in size from 4 to 7 generations and from 4–53 individuals per family. For the third design, denoted RUT, I simulated marker data on a mixture of Centre d'Etude du Polymorphisme Humain and deCODE [Kong et al., 2002] families at 13 markers taken from the Rutgers map [Matise et al., 2003]. These families contained 1,000 individuals with almost no missing data.

Each design in the *screening sets* setting uses a different map, and the three maps share 5 of the 15 distinct markers (Table I). For each design, the allele frequencies and missing data patterns were chosen to reflect those found in the corresponding real data, and a total of 100 replicates were simulated and analyzed. This yielded 100 sets of map estimates containing three independent map estimates per set. For each set, I used the proposed method to compute the ECME and hence, the meta-analysis-based map estimate for each design.

RESULTS

I analyzed simulated data to evaluate the proposed method in three different settings. In the *complete data* setting, CRIMAP calculates the exact MLEs of the recombination fractions in the maps of *studies 1*, *2*, and *3*. However, LM_MAP estimates the MLEs using MCMC. The difference between the two methods yields an estimate of the MCMC error. On average, less than one-fifth of one percent (.0012) of the total variation of any MLE estimated by LM_MAP was attributable to MCMC error.

In the same setting, accurate estimation of the variance of each recombination fraction was more difficult. For example, both the LM_MAP and the nonparametric bootstrap variance estimators appear to have difficulty near the ends of the chromosome. Specifically, for intervals 1 and 8 of the *study 2* map (odd markers) and for intervals 1 and 7 of the *study 3* map (even markers), the distribution of the difference between LM_MAP and the empirical variance estimates is visibly shifted away from zero. The same comment applies to the distribution of the difference between the nonparametric bootstrap and the empirical variance estimates (Fig. 2A and B). Despite these shared difficulties, the LM_MAP variance estimates are in greater agreement with the corresponding empirical variance estimates. A test of the null hypothesis that the mean difference between the LM_MAP and the empirical variance estimates is normally distributed,

TABLE I. Marker Map Information

Name	Position	Composite map	RUT map	MSS9 map	MSS16 map	Heterozygosity	Number of alleles
D20S103	0.0	√	√	√	—	.733	7
D20S482	12.5	√	√	√	√	.714	6
D20S602	22.4	√	√	—	√	.667	3
D20S851	29.0	√	√	√	—	.837	10
D20S604	37.6	√	√	√	—	.653	6
D20S1143	41.0	√	—	—	√	.707	5
D20S470	44.4	√	√	√	—	.867	12
D20S477	53.9	√	√	√	√	.746	7
D20S478	61.3	√	√	√	√	.687	7
D20S481	70.7	√	√	√	√	.746	7
AAT269	77.6	√	—	—	√	.750	4
D20S480	84.5	√	√	√	√	.769	7
D20S451	95.8	√	√	—	√	.793	6
D20S171	95.9	√	√	√	—	.471	2
D20S164	106.4	√	√	—	√	.504	5

Note: Each row contains information pertaining to 15 markers of chromosome 20. Columns 3–6 denote different maps (described in the text). √ or — indicates the presence or absence of the corresponding marker on the corresponding map. Map positions are given in centiMorgans based on the Kosambi map function.

yielded a P-value of 0.136. The same test based on the nonparametric bootstrap variance estimator yielded a P-value of 0.040. As a final point to the results of the *complete data* setting, the parametric bootstrap procedure accurately estimates the marginal variance of the ECME. Specifically, the estimates based on the parametric bootstrap procedure are clustered tightly around the corresponding empirical estimate of the marginal variance of the ECME (Fig. 2C).

From the analysis of data simulated in the *incomplete data* setting, the average variance of the ECME is smaller than the average variance of the map estimate based on *study 1* data, provided that the independent map estimates are ML estimates (Fig. 3D and E). For the same map estimates and for the same caveats, the average absolute bias was relatively unchanged (Fig. 3A and B). For each statistic (average absolute bias and average variance) and for each estimation procedure, the average is taken over the intervals of the corresponding map. Note that while the ECME and the *study 1* map use the same 20 markers, they do not use the same data.

When the independent map estimates are estimated by CRIMAP and both parental and grandparental data are missing, the average variance of the ECME is not smaller than the average variance of the *study 1* map (Fig. 3D and E). Relative to ML map estimation, the average absolute bias and the average variance of the CRIMAP estimator are markedly increased when the marker data of both parents and grandparents are missing (Fig. 3A and D).

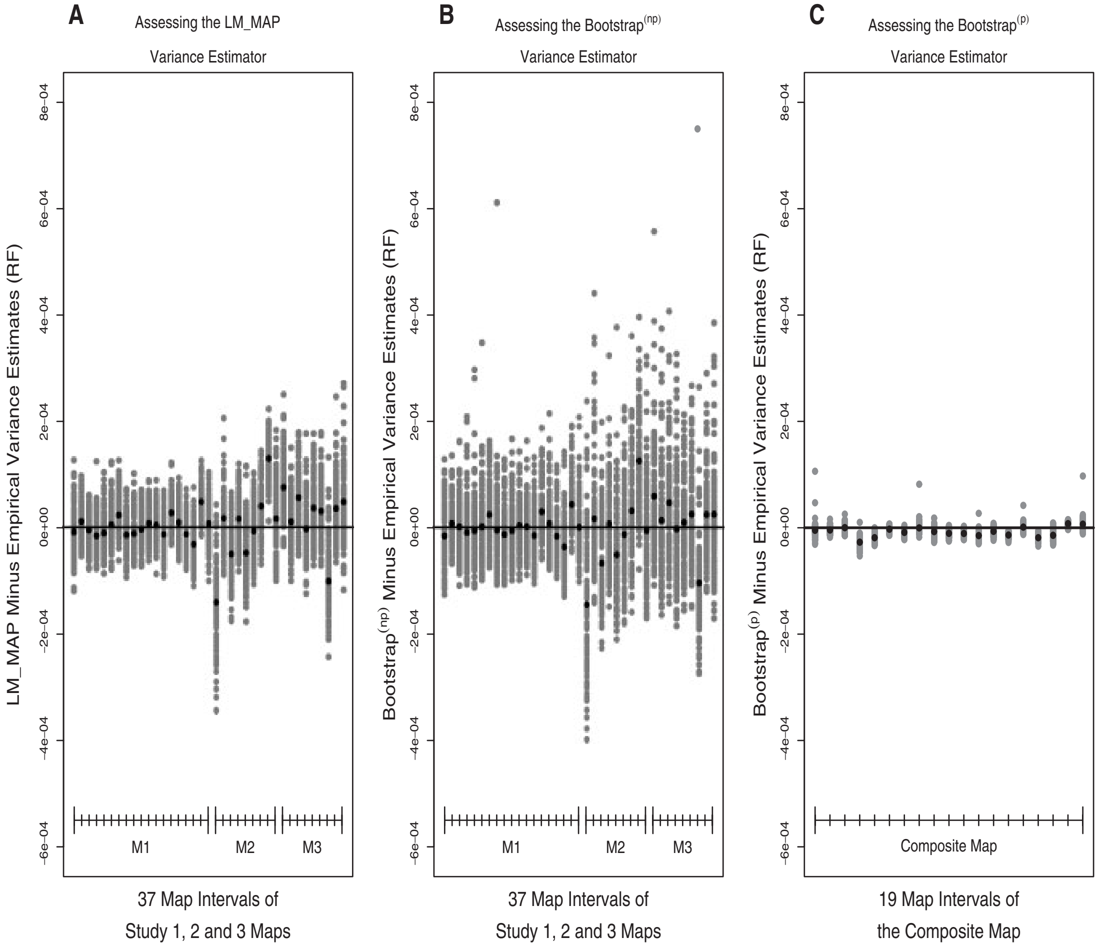

Fig. 2. For each recombination fraction (RF) estimate in each map, the difference between a selected variance estimator and the corresponding empirical estimate of the variance is shown. (A) LM_MAP uses a variance estimator that is based on the inverse observed information. (B and C) The nonparametric and parametric bootstrap estimators are denoted by bootstrap$^{(np)}$ and bootstrap$^{(p)}$, respectively. Gray dots are obtained from 100 replicates and black dots represent the mean. The thick black line demarcates zero and thin black horizontal lines represent marker maps. The maps of *studies 1, 2,* and *3* are labelled M1, M2, and M3, respectively. Interior and exterior intervals of each map are denoted by small and large ticks, respectively.

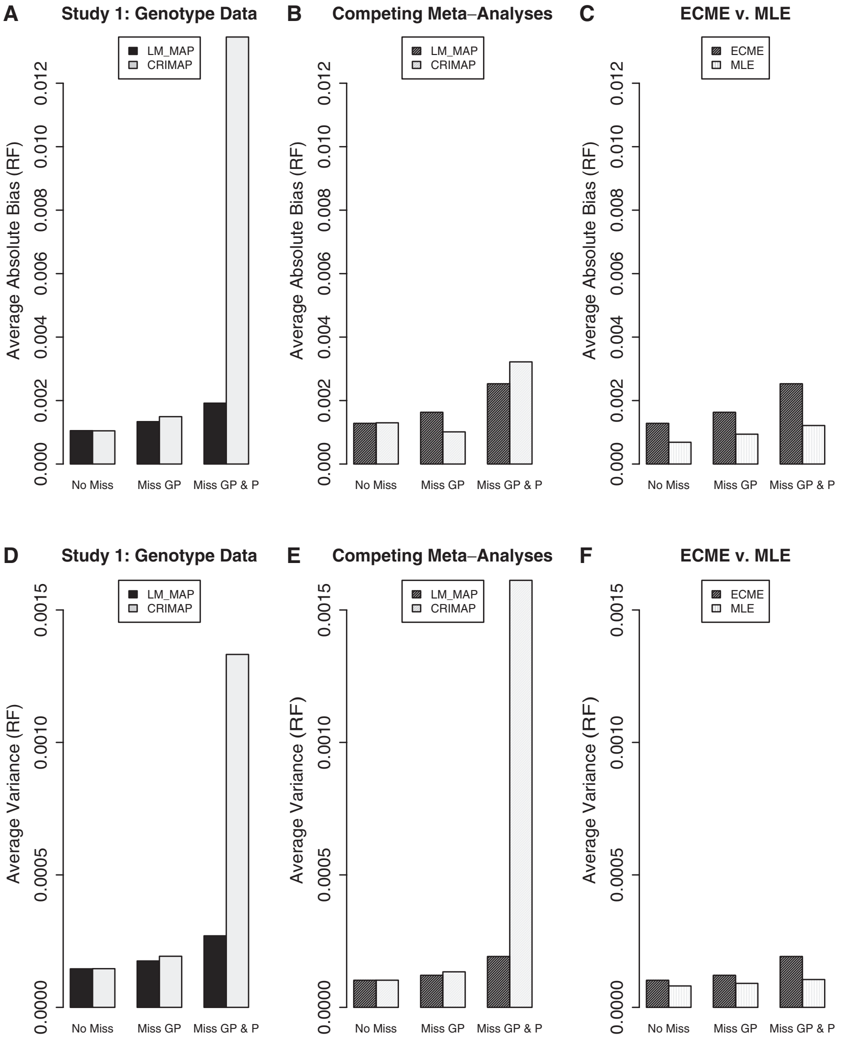

Fig. 3. For each map estimator in (A), (B), and (C), the interval-specific estimates of the absolute bias in recombination fraction (RF) are averaged over the intervals in the corresponding map. Similarly, (D), (E), and (F) give the corresponding estimates of the average variance. (A and D) Compare LM_MAP and CRIMAP estimation programs using the genotype data of *study 1*. (B and E) Compare the meta-analysis of independent map estimates generated by LM_MAP to the meta-analysis of independent map estimates generated by CRIMAP. (C and F) Compare the ECME to the MLE obtained from the joint analysis of the *pooled data*. ECME, extended composite map estimates; MLE, maximum likelihood estimate.

In addition to achieving smaller average variance than the *study 1* map estimate based on genotype data, the ECME has efficiency comparable to that of the MLE based on the joint analysis of the *pooled data* (Fig. 3E and F). Moreover, the ECME agrees well with the simulation truth (Fig. 4).

From the analysis of simulated data in the *screening sets* setting, the average variance of each meta-analysis-based map estimate is smaller than the average variance of each independent map estimate. By contrast, the average absolute bias is about the same for both estimation procedures. This is shown in Table II, which gives the average percent reduction in variance and the change in the average absolute bias for all three designs. For each statistic and for each design, the average is taken over the intervals of the corresponding map.

DISCUSSION

Meta-analysis can be a useful approach to map estimation when the independent map estimates of multiple studies are available but the original data are not easily pooled. In particular, if the

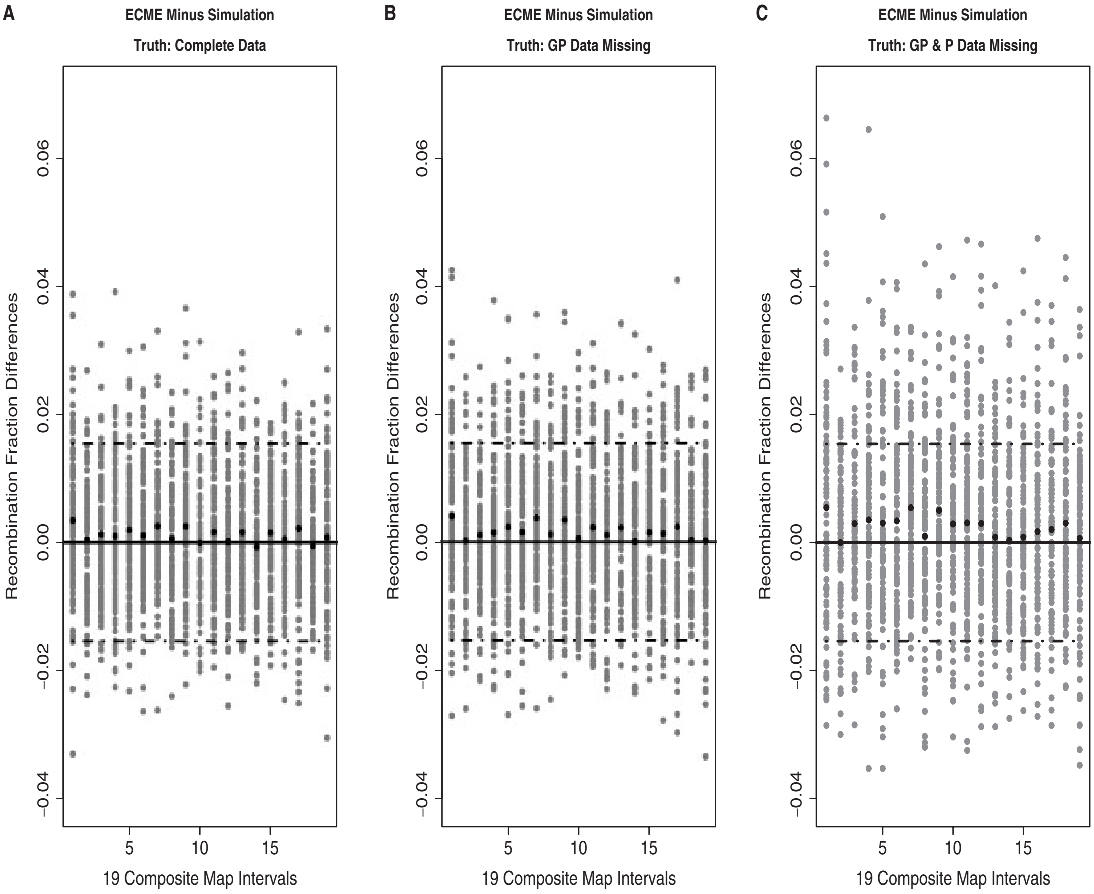

Fig. 4. For each interval of the composite map and for each pattern of missing data: no missing data (A), missing grandparental data (B), and missing both parental and grandparental data (C), the difference between the estimated recombination fraction and the simulation truth based on 100 Monte Carlo realizations of the ECME is shown (gray dots). Black dots denote the mean difference. The labels GP and P denote grandparental and parental, respectively. The solid black line demarcates zero and dotdash black lines demarcate the 95% confidence limits for a sample of 800 fully informative meioses.

TABLE II. Variance Reduction in Marshfield Maps

Linkage mapping data set	Percent reduction in variance	Change in average absolute bias
MSS9	57	−0.0007
MSS16	42	+0.0000
RUT	57	−0.0001

Note: For each linkage map (described in the text) estimates are obtained in two ways: (1) maximum likelihood is applied to the genotype data of the corresponding linkage mapping design, and (2) meta-analysis is applied to the three independent map estimates obtained in (1). The percent reduction in variance and the change in average absolute bias are shown.

independent map estimates are ML map estimates then, for the settings considered, the ECME has (1) reduced variability relative to the variability of any map estimate used in the meta-analysis; and (2) efficiency comparable to that of the MLE based on the joint analysis of the pooled genotype data.

To investigate the sensitivity of these results to heterogeneity, I estimated ML map estimates from highly differentiated sets of marker data and computed the ECME. The composition of each data set differed markedly in terms of family size, stucture, data availability, marker heterozygosity, and map density. In each case, the ECME continued to exhibit reduced variation relative to the individual map estimates used in the meta-analysis (data not shown).

In contrast to ML map estimation, when the independent map estimates are obtained via CRIMAP and there is missing data, the individual map estimates used in the meta-analysis and the ECME may both be affected sharply. In general, when genotype data are missing and CRIMAP is used, the magnitude and direction of bias for each independent map estimate will not be known. Therefore, a prudent course of action seeks to minimize the degree to which underlying assumptions of the model are violated. For example, the presence of linkage disequilibrium, undetected genotyping errors, and inaccurate allele frequencies are all potential sources of bias for linkage analysis [Abecasis and Wigginton, 2005; Ott, 1999; Clerget-Darpoux et al., 1986]; and in principle, for linkage maps too.

Two immediate extentions of the proposed method include: (1) sex-specific map estimation and (2) map estimation from heterogeneous data. The first is important since sex-specific maps are more variable than sex-averaged maps estimated from the same data. The second is important since heterogeneous data can violate underlying assumptions of the models used by existing map estimation methods. As an example of this type of application, consider a heterogeneous data set composed of families sampled from two populations that differ markedly in terms of their allele frequencies and the number of alleles at each marker. Existing methods could be used to analyze the families of each population separately, as each group of families is homogeneous. Then, the proposed method could be used to combine efficiently the resulting map estimates.

The proposed method requires accurate map estimates and estimates of their variances, both of which LM_MAP provides. In principle, LM_MAP can analyze arbitrary amounts of pedigree data; however, CPU time becomes increasingly important as the size and complexity of a data set increases. Alternatively, CRIMAP could be used in conjunction with the nonparametric bootstrap procedure to estimate the map and its variance, but caution is advised when there are substantial amounts of missing data. In either case, the variance estimates for recombination fractions in intervals near the ends of the chromosome may not be reliable. Fortunately, the distribution of the ECME was insensitive to differences between weights constructed from either variance estimator (data not shown). Moreover, since direct estimates of the variance may not always be available, weights could be constructed from existing estimates of marker informativeness. In particular, the proposed method could be used to combine the information of individual multipoint linkage studies with the information of the Rutgers linkage mapping data set [Matise et al., 2003].

For many studies in which map estimates are potentially available, it may be difficult to gain authorized access to the original genotype data. Even if these data are accessible, a joint analysis may be computationally infeasible due to their size and complexity, or inappropriate due to the presence of heterogeneity. In such situations, a meta-analysis of map estimates may be the only way to combine efficiently the information contained in the original genotype data. In particular, a vast collection of map estimates is potentially available from the analysis of multipoint marker data contained in large linkage and linkage mapping studies. The proposed method could be used to combine these map estimates into a single, comprehensive linkage map of unprecedented precision.

ACKNOWLEDGMENTS

I thank Dr. Victor Abkevich for supplying information regarding his recent linkage study and Dr. Tara Matise for providing information regarding the Rutgers data set. I thank Drs. Gonç alo Abecasis, Michael Boehnke, and the reviewers for their many helpful comments.

WEB RESOURCES

The URLs for software presented herein are as follows:

METAMAP, http://www-personal.umich.edu/~wstew/

LM_MAP, http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml

CRIMAP, http://linkage.rockefeller.edu/soft/crimap/

REFERENCES

Abecasis GR, Wigginton JE. 2005. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767.

Abkevich V, Camp NJ, Hensel CH, Neff CD, Russell DL, Hughes DC, Plenk AM, Lowry MR, Richards RL, Carter C, Frech GC, Stone S, Rowe K, Chau CA, Cortado K, Hunt A, Luce K, O'Neil G, Poarch J, Potter J, Poulsen GH, Saxton H, Bernat-Sestak M, Thompson V, Gutin A, Skolnick MH, Shattuck D, Cannon-Albright L. 2003. Predisposition locus for major depression at chromosome 12q22-12q23.2. Am J Hum Genet 73:1271–1281.

Barber MJ, Todd JA, Cordell HJ. 2006. A multimarker regression-based test of linkage for affected sib-pairs at two linked loci. Genet Epidemiol 30:191–208.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic map: individual and sex-specific variation in recombination. Am J Hum Genet 63: 861–869.

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J. 1986. Effects of misspecifying genetic parameters of the trait model in lod score analysis. Biometrics 42:393–399.

Daw EW, Thompson EA, Wijsman EM. 2000. Bias in multipoint linkage analysis arising from map misspecification. Genet Epidemiol 19:366–380.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Series B 39:1–37.

Efron B, Tibshirani RJ. 1993. An Introduction to the Bootstrap. New York: Chapman & Hall.

Fingerlin TE, Abecasis GR, Boehnke M. 2006. Using sex-averaged genetic maps in multipoint linkage analysis when identity-by-descent status is incompletely known. Genet Epidemiol 30:384–396.

George AW. 2005. A novel Markov chain Monte Carlo approach for constructing accurate meiotic maps. Genetics 171:791–801.

Guo SW, Thompson EA. 1994. Monte Carlo estimation of mixed models for large complex pedigrees. Biometrics 50:417–432.

Halpern J, Whittemore AS. 1999. Multipoint linkage analysis. A cautionary note. Hum Hered 49:194–196.

Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML, Lamb NE, Sherman SRGJ, Stefansson K. 2004. Recombination rate and reproductive success in humans. Nat Genet 36:1203–1206.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. 2002. A high resolution recombination map of the human genome. Nat Genet 31:241–247.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Nat Acad Sci USA 84:2363–2367.

Louis TA. 1982. Finding observed information using the EM algorithm. J R Stat Soc Series B 44:98–130.

Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Nat Acad Sci USA 99:2228–2233.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, Toma dC, Ehm M, Glanowski S, He C, Heil J, Markianos K, McMullen I, Pericak-Vance MA, Silbergleit A, Stein L, Wagner M, Wilson AF, Winick JD, Winn-Deen ES, Yamashiro CTMCH, Lai E, L HA. 2003. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet 73:271–284.

Ott J. 1999. Analysis of Human Genetic Linkage, 3rd edition. Baltimore, MD: Johns Hopkins University Press.

Stewart WCL, Thompson EA. 2006. Improving estimates of genetic maps: a maximum likelihood approach. Biometr 62: 1–8.

Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE. 2005. A map of the human genome in linkage disequilibrium units. Proc Nat Acad Sci USA 102:11835–11839.

APPENDIX

COMPUTATIONAL PARAMETERS FOR SIMULATIONS AND ANALYSES

For all LM_MAP analyses, 8 steps of the Monte Carlo Expectation Maximization algorithm [Guo and Thompson, 1994] were used to find the ML estimate of the map. The initial parameter value for the algorithm was sampled from a distribution centered around the simulation truth. The initial inheritance pattern was sampled from the conditional distribution of inheritance patterns given marker data. MCMC estimates of the ML map estimate take 6–15 minutes, whereas MCMC estimates of the variance of the ML map estimate take 3–8 min using a 2.8–3.2 GHz processor. For each replicate of *studies 1*, *2*, and *3*, 100 nonparametric bootstrap samples were used to compute the variance of each component of the corresponding ML map estimate. For each *pooled data* replicate in the *incomplete data* setting (described in Simulation Description), 1,000 parametric bootstrap samples were used to compute the variance of each component of the ECME. The recursive algorithm used to compute the ECME terminates when the absolute change in successive estimates is less than 10e-07.