Genetic Epidemiology 20:17–33 (2001)

A Bayesian Hierarchical Model for Allele Frequencies

John R. Lockwood,[1]* Kathryn Roeder,[1] and B. Devlin[2]

[1]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania
[2]Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania

Genetic epidemiological methodologies, such as linkage analysis, often require accurate estimates of allele frequencies. When studies involve multiple sub-populations with different evolutionary histories, accurate estimates can be difficult to obtain because the number of subjects per sub-population tends to be limited. Given allele counts for a collection of loci and sub-populations, we propose a Bayesian hierarchical model that extends existing empirical Bayesian approaches by allowing for explicit inclusion of prior information about both allele frequencies and inter-population divergence. We describe how such information can be derived from published data and then incorporated into the model via prior distributions for model parameters. By analysis of simulated data, we highlight how the hierarchical model, as implemented in the publicly available program AllDist, combines prior information with the observed data to refine allele frequency estimates. Genet. Epidemiol. 20:17–33, 2001. © 2001 Wiley-Liss, Inc.

Key words: empirical Bayes; linkage analysis; population substructure; shrinkage estimators

INTRODUCTION

Many genetic epidemiological analyses are quite sensitive to estimates of allele frequencies. One of the best examples is linkage analysis involving ambiguous identity-by-descent (IBD) status. In this instance, IBD must be inferred as a function of population allele frequencies, with the probability of IBD increasing with the rarity of shared alleles [Weeks and Lange, 1988; Risch, 1990; Kruglyak et al., 1996]. Clearly, underestimation of allele frequencies can lead to false linkage, whereas overestima-

Contract grant sponsor: National Institutes of Health; contract grants MH57881 and MH56193; Contract grant sponsor: National Science Foundation, Contract grant DMS-9803433.

*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: roeder@stat.cmu.edu

Received for publication 30 November 1999; revision accepted 1 April 2000

tion can lead to reduced power. Although recent work underscores the severity of this problem for affected relative-pair methods without genotyped parents [Ott, 1992; Friemer et al., 1993; Tores et al., 1996], it remains pertinent to linkage analyses of extended pedigrees [Risch and Giuffra, 1992].

An important trend in linkage analysis of complex diseases is the conglomeration of families of different ethnic origins, thereby presumably increasing the power to detect liability genes of subtle effect. Often such studies use pooled allele frequencies for inference, ignoring sub-population heterogeneity and implicitly assuming sub-population heterogeneity is negligible. This assumption, although harmless if true, may seriously compromise inference regarding linkage when it is false.

When pedigrees are drawn from heterogeneous sub-populations, the ideal linkage analyses employ sub-population-specific allele frequency estimates. However, such estimates can be error-prone when only a small number of independent subjects are drawn from each sub-population. Fortunately, this problem has a solution that involves "shrinking" the error-prone sub-population estimates toward some common value. Such solutions have already been implemented in the field of genetic epidemiology in the context of forensic inference from genetic markers [Devlin et al., 1991; Lange, 1995].

Most relevant to our discussion of linkage analysis is Lange's [1995] empirical Bayes estimator for allele distributions. Although his approach is developed rigorously in the sequel, in which it forms the nucleus for our model, it can be paraphrased as follows. First, estimate the degree of allele frequency heterogeneity for a locus of interest. Then shrink sub-population-specific allele frequencies toward their pooled estimates as a function of the estimated sub-population heterogeneity: when heterogeneity is large, shrinkage is relatively small and vice versa.

Our goal is to refine Lange's empirical Bayes model for allele frequency data by using Bayesian hierarchical methods. By re-casting the problem, we take advantage of several sources of information not incorporated into the empirical Bayes analysis. For instance, multiple loci are assessed for most modern epidemiological analyses, especially linkage analysis. These loci all share a common population history causing the degree of sub-population heterogeneity across loci to be similar. Therefore, to estimate sub-population heterogeneity, our model "borrows strength" across loci, allowing the estimated heterogeneity at each locus to be informed by all other loci. Moreover, the model allows for explicit inclusion of prior information about both allele frequencies and inter-population divergence. These extensions produce more accurate estimates of sub-population-specific allele frequencies. In this article, we first describe the model, establish notation, and provide an explanation of the model structures. We then discuss prior distributions, propose an approximation to the model that reduces computational burden, and address issues concerning underdispersed data. Finally, we give an example of the proposed methods and a discussion. The methods we develop for Bayesian estimation of allele frequencies are implemented in the publicity available software AllDist.

MODEL SPECIFICATION

In this section, we describe and motivate a Bayesian hierarchical model that incorporates prior information for each marker, estimates the level of heterogeneity

among the allele distributions of the sub-populations, and shrinks the allele frequency estimates accordingly toward both the pooled and prior frequencies. The model has four stages: the first stage is the multinomial likelihood model for allele counts, stratified by sub-population (*allele count* stage); the second stage provides a model for the variability of allele frequencies across sub-populations (*sub-structure* stage); the third stage allows for the input of information about allele frequencies obtained from prior studies (*allele count prior* stage); and the final stage quantifies prior evidence about the level of heterogeneity among the sub-populations under study (*dispersion prior* stage).

The data consist of allele counts for L loci and S sub-populations. Let $l = 1,...,L$ index loci, $s = 1,...,S$ index sub-populations, $a = 1,...,M_l$ index alleles at the *lth* locus, and X_{asl} be the allele count and $\mathbf{X_{sl}}$ be the vector of allele counts for a sub-population at a given locus.

Allele Count Stage

For $s = 1,...,S$ and $l = 1,...,L$, we assume that the allele counts $\mathbf{X_{sl}}$ are distributed Mult(n_{sl};$\boldsymbol{p}_{sl}$), independently with probabilities $\mathbf{p_{sl}} = (p_{1sl},...,p_{M_lsl})$. Assuming the allele distribution is multinomial for each sub-population implies Hardy-Weinberg equilibrium within sub-populations. This is a standard assumption for population genetic models.

Substructure Stage

For $s = 1,...,S$ and $i = 1,...,L$, we assume that $\mathbf{p_{sl}}$ are independent random vectors with a locus-specific Dirichlet distribution. The Dirichlet parameters $\boldsymbol{\alpha}_l = (\alpha_{1l},...,\alpha_{M_ll})$ are the key to inferences concerning allele frequencies $\mathbf{p_{sl}}$. These parameters are the expected values of p_{asl} times a simple function of the variability among populations. Let $\alpha_{\cdot l} = \Sigma_{a=1}^{M_l} \alpha_{al}$. Then, $\alpha_{al}/\alpha_{\cdot l}$ is the expected value of p_{asl} and $\alpha_{\cdot l}$ quantifies the dispersion of $\mathbf{p_{sl}}$ about its mean, with larger values of $\alpha_{\cdot l}$ implying smaller dispersion. The dispersion parameter can be expressed as a function of the standardized measure of inter-populational variability of allele frequencies θ_l (or F_{st} [Wright, 1951], namely $\alpha_{\cdot l} = (1 - \theta_l)/\theta_l$.

The *allele count* and *sub-structure* stages together form the likelihood for the data $\mathbf{X_{sl}}$. Combining these two modeling stages yields the marginal distribution of $\mathbf{X_{sl}}$, which is the compound multinomial or Dirichlet multinomial distribution with parameter $\boldsymbol{\alpha}_l$ [Johnson et al., 1997]. Introduced by Mosimann [1962], it is the multivariate generalization of the beta-binomial distribution [Skellam, 1948]. Although the Dirichlet model does not provide a perfect fit to evolutionary theory, it induces correlation among alleles within a sub-population matching that predicted by the theory [Jiang and Cockerham, 1987].

The final two stages define the prior distribution for the parameters of the compound multinomial likelihood model.

Allele Count Prior

With a transformation of variables, the prior information for $\boldsymbol{\alpha}_l$ can be modeled more easily. For $a = 1,...,M_l - 1$, let $\psi_{al} = \log(\alpha_{al}/\alpha_{M_ll})$ and let $\beta_l = \log(\alpha_{\cdot l})$. By invoking the "logit" transformation, we can separate the cell frequencies $\boldsymbol{\alpha}_l/\alpha_{\cdot l}$ of the Dirichlet distribution from the dispersion parameter $\alpha_{\cdot l}$. This separation has two ad-

vantages: the components have different dependence structures across sub-populations; and the re-parameterization improves the validity of the normal approximation we will use to implement the theory in AllDist. It is natural to assume that $\boldsymbol{\psi_l}$ and β_l are independent a priori for each l. With this assumption, the prior for $(\boldsymbol{\psi_l},\beta_l)$ splits into two independent parts:

1. Independently for each l we assume that $\boldsymbol{\psi_l} \equiv (\psi_{1l},...,\psi_{M_l-1,l})$ has a multivariate normal distribution. The mean $\boldsymbol{\mu_l}$ is a $M_l - 1$ dimensional vector and the covariance $\boldsymbol{\tau_l^{-1}}$ is a $(M_l - 1) \times (M_l - 1)$ matrix. Both are derived from the prior observations.
2. We assume that β_l is distributed normally with mean μ_β and variance τ_β^{-1}, independently for each l.

Dispersion Prior

To complete the model, we suppose that μ_β is distributed normally with mean δ and variance ϵ^{-1}, and that τ_β is distributed Gamma with shape parameter a and scale parameter b. The parameters δ, ϵ, a, b are known as hyperparameters because they come from external information independent from the study's data and are not estimated by the model. This final stage has two purposes: it allows input of prior information about dispersion and encourages shrinkage of the individual β_l toward a common mean value for the entire set of loci. The amount of shrinkage depends on the prior and the data. Ultimately, however, the procedure estimates distinct values of β_l for each locus.

Table I gives a guide to the various transformations of the Dirichlet parameters $\boldsymbol{\alpha_l}$ that we will utilize. To summarize, a larger value of the dispersion parameter, $\boldsymbol{\alpha_{\cdot l}}$, means less dispersion among sub-populations. The θ_l parameterization is useful because it is well known in the population genetics literature. The transformations β_l and ψ_{al} are necessary because estimates of these quantities are more nearly normally distributed than the corresponding un-transformed quantities. Ultimately results about these transformations of $\boldsymbol{\alpha_l}$ are back-transformed to the original scale for inference.

Empirical Bayes Estimation

Lange's [1995] empirical Bayes method treats the *allele count* and *sub-structure* stages as a complete model for allele frequencies. Independently for each locus, the compound-multinomial likelihood function is used to estimate $\boldsymbol{\alpha_l}$. The analysis then proceeds (locus by locus) to obtain the posterior distribution of the cell frequencies for each sub-population. The allele frequencies are estimated using the posterior mean,

TABLE I. Useful Transformations of the Dirichlet Parameters $\boldsymbol{\alpha_l}$

Parameter	Function	Meaning
$\alpha_{\cdot l}$	$\Sigma_a\, \alpha_{al}$	Dispersion parameter
θ_l	$1/(1 + \alpha_{\cdot l})$	Wright's F_{st}
β_l	$\log_e(\alpha_{\cdot l})$	Transformed dispersion parameter
ψ_{al}	$\log_e(\alpha_{al}/\alpha_{Ml})$	Logit of expected allele frequencies

$$\tilde{p}_{asl} = \frac{X_{asl} + \alpha_{al}}{n_{sl} + \alpha_{.l}}. \quad (1)$$

Because the relative magnitudes of the components of $\boldsymbol{\alpha}_l$ are reflective of the corresponding magnitudes of the marginal allele frequencies, larger values of $\alpha_{.l}$ shrink the sub-population estimates more closely toward the marginal frequencies. This behavior is sensible as a large $\alpha_{.l}$ suggests very little divergence among sub-populations.

Bayesian Estimation

Our proposed model essentially inherits the empirical Bayes structure, but adds two important features: locus-to-locus dependence in the degree to which the sub-populations differ in allele frequencies is modeled directly, and prior information about allele frequencies for each locus is incorporated. This additional structure is built into the *allele count* and *dispersion priors*. The log-odds vectors of expected allele frequencies, $\boldsymbol{\psi}_l$, are treated separately for each locus, with the goal of shrinking the estimated cell frequencies towards the marginal distribution of the population and toward the prior distribution for that locus. If the observed frequency for an allele in a particular sub-population is zero, which will be common when the number of sampled individuals is small, the estimated frequency of that allele is not forced to be zero. This helps to estimate frequencies of rare alleles, which may not be present in all sub-populations when sample sizes are limited.

The logarithms of the dispersion parameters β_l are assumed to be a sample from a common distribution across loci. This allows distinct parameters at different loci while promoting shrinkage across loci toward a common value. The coherence of values is expected because the loci share a common population history. Despite their shared history, distinct processes and rates of mutation promote inter-locus variability [Chakraborty and Jin, 1992]. Our structure directly accounts for both features of the data. The prior structure incorporated in the *dispersion prior* stage allows this shrinkage to be a function of the observed data.

PRIOR DISTRIBUTIONS

To implement the model, fixed values for δ, ϵ, a, and b from the *dispersion prior* stage and for $\boldsymbol{\mu}_l$ and $\boldsymbol{\tau}_l$ from the *allele count prior* stage are required. Judicious choice for these values incorporates the information from prior studies while ensuring that the priors do not dominate the results for the current study. In this section, we outline a procedure for determining values for these hyperparameters based on prior information coming from the literature and reference databases of allele frequencies (e.g., CEPH).

Deriving Values for δ, ϵ, *a, b*

A plausible prior for β_l follows from basic reasoning. This prior is initially formulated for the transformation θ_l because it is for this parameterization that historical records are available. Presumably, interest lies in sub-populations with θ_l ranging between 0.003 and 0.03. For smaller values, there is little concern about heterogene-

ity in allele frequencies, and for larger values it may not be appropriate to pool populations in a common genetic analysis. This information can be utilized to choose δ, ϵ, a, and b. A reasonable specification of the hyperparameters given θ_l between 0.003 and 0.03 is $\delta = 4.5$, $\epsilon = 2.0$, $a = 18$, $b = 6$. Using these parameters, a prior distribution for β_l is induced that transforms to a prior distribution for θ_l with median approximately 0.01, first quartile equal to 0.007, third quartile equal to 0.016, and $Pr(\theta > 0.03)$ approximately 0.05. Although the investigator can input other parameter choices in AllDist, our specification is consistent with reasoning developed here, and it provides enough prior variance to allow the data to update the distribution.

Deriving Values for the Vectors μ_l and the Matrices τ_l

A prior for allele frequencies requires a prior for ψ_l in terms of parameters μ_l and τ_l for the allele distribution. The information for specifying the prior distribution comes from one or more allele frequency distributions observed from prior studies. To translate these "prior data" into a prior distribution, we use the properties of the compound multinomial distribution.

Procedure 1: Single-Reference Population

For each locus, assume a vector of allele counts is available from only one reference population. Let $\mathbf{Y_l} = (Y_{1l},...,Y_{Ml})$ be the allele counts and $n_l = \Sigma_{a=1}^{M_l} Y_{al}$. Our goal is to use $\mathbf{Y_l}$ to determine values for the hyperparameters μ_l and τ_l, assuming that $\mathbf{Y_l}$ follows a compound-multinomial distribution with parameter α_l. To do so, we assume a fixed value for $\alpha_{.l} = e^{\beta_l}$ because inference about this dispersion parameter with data from only a single population is impossible. From the prior described previously, a sensible choice is the transformed mean of the prior distribution of β_l, $e^{4.5}$. We use the sampling distribution of the prior data $\mathbf{Y_l}$ to form the prior for our current data analysis. See Appendix A for details on how this translates into numerical values for μ_l and τ_l. From the form of the covariance matrix described therein, it can be deduced that the prior information cannot exert excessive influence on the Bayesian estimation procedure regardless of the size of the sample n_l. Prior samples with more than $n_l = 500$ alleles have essentially the same influence as a sample of size 500.

Procedure 2: Multiple-Reference Populations

In principle, when allele counts from multiple populations are available, the structure of the hierarchical model could be used to find the maximum likelihood estimate (MLE) of the underlying compound multinomial distribution. However, because of occasional difficulties associated with calculating the MLE for the compound multinomial model, we have not introduced this additional structure in AllDist. Instead, if multiple populations are available a priori, AllDist requires that only the marginal allele counts be used. This is equivalent to treating the multiple samples as a single population, where the calculation of μ_l and τ_l follows as in the single-population procedure.

When data are unavailable for a particular locus, we suggest setting μ_l to a vector of zeroes (corresponding to equal cell frequencies) and τ_l to a small number times the identity matrix. This allows the frequencies for that locus to be entirely determined by the data. If one or more alleles has a frequency of zero in the prior data, we suggest augmenting the counts of *all* alleles for that locus by a small value

(commonly 0.5) called a *flattening constant* [Schafer, 1997]. The effect is to shrink the parameter estimates slightly toward a uniform distribution on the allele frequencies, while eliminating computational difficulties associated with the zero count.

INFERENCE

To improve computational efficiency, we introduce a simplification and an approximation to our model. The simplification uses the fact that the *allele count* and *sub-structure* stages imply that for each s and l, $\mathbf{X_{sl}}$ given $\boldsymbol{\alpha}_l$ follows a compound multinomial distribution with parameter $\boldsymbol{\alpha}_l$. The approximation replaces the assumed model with an asymptotically normal version [Daniels and Kass, 1997] that allows calculation of some model features in closed form and greatly facilitates simulation from the joint posterior of the model parameters.

The approximation treats the MLE $\hat{\boldsymbol{\alpha}}_l$ as the data. Based on the asymptotic properties of MLEs, $\hat{\boldsymbol{\alpha}}_l$ is distributed multivariate normal with mean $\boldsymbol{\alpha}_l$ and variance equal to the observed Fisher information matrix. The latter quantity $\mathbf{I}_{\alpha l}(\hat{\boldsymbol{\alpha}}_l)$ is the matrix with (a,s) entry $-\partial^2/\partial\alpha_a\partial\alpha_s \log \mathcal{L}_l\,(\boldsymbol{\alpha}_l)$ where $\mathcal{L}_l\,(\boldsymbol{\alpha}_l)$ is the compound multinomial likelihood function. Because the revised model has a hierarchical normal structure, Gibbs or successive substitution sampling from the joint posterior distribution of parameters can be used to calculate characteristics of the model [Geman and Geman, 1984; Schervish, 1995; Gelman et al., 1995; Gilks et al., 1996]. Such calculations yield the posterior means of the vectors $(\boldsymbol{\psi}_l, \boldsymbol{\beta}_l)$, which can be back-transformed to obtain estimates of $\boldsymbol{\alpha}_l$. Finally, the allele frequencies are estimated using equation (1). It should be clear at this point that the Bayesian estimation procedure differs from the empirical Bayesian procedure only in how it estimates $\boldsymbol{\alpha}_l$. Given an estimate of these quantities, both procedures estimate the sub-population-specific allele frequencies by the same process.

UNDERDISPERSED DATA

For some loci, $\alpha_{.l}$ is likely to be very large, suggesting homogeneity across subpopulations or very little divergence. For these loci, divergence may not be statistically detectable either because it does not exist or because small effects may be obscured by sampling error. When this occurs, the marginal distribution of $\mathbf{X_{sl}}$ will correspond to a degenerate compound multinomial distribution because observations appear to come from a common multinomial distribution. In this case, a proper MLE for the compound multinomial distribution does not exist: the MLE for $\alpha_{.l}$ is infinite, and estimation is impossible. Fortunately, there is a simple patch. We assume the parameter θ_l takes on some baseline value, say $\theta_B = 0.001$, which corresponds to large $\alpha_{.l}$. This value yields a proper compound-multinomial distribution close to a simple multinomial distribution, eliminating computational difficulties caused by underdispersion. Under these conditions, a closed form solution exists for estimation of $\boldsymbol{\alpha}_l$ (see Appendix C).

The only shortcoming of this approach is that the information about heterogeneity contained in the underdispersed loci is not utilized when making inferences about the other loci. That is, the estimates of all parameters relating to heterogeneity are driven only by the loci for which the heterogeneity is observable. Because the

bias errs on the side of less shrinkage, thereby retaining more of the sub-population data structure, we believe it is the most sensible approach to the problem.

Finally, we offer a suggestion for the choice of the baseline value θ_B. The primary issue is that it be chosen small enough to be beyond the range of values where the practitioner desires to discriminate between divergence and homogeneity while, at the same time, not so extreme as to allow the estimates of the underdispersed data to introduce numerical problems. Our experience indicates that values of θ_B in the range 0.001–0.003 are sufficient to satisfy both of these concerns. In our implementation, AllDist determines the data are underdispersed based on the behavior of the mode-finding algorithm we used to fit the model (see Appendix C).

SIMULATIONS

To highlight the distinctions between the Bayesian hierarchical model and other estimation methods, we performed a set of simple simulations. We simulated the assessment of 200 loci, each with the same fundamental allele frequency distribution (0.05, 0.1, 0.2, 0.3, 0.2, 0.1, 0.05). This probability distribution over alleles is comparable in heterozygosity to STR loci commonly used for linkage analysis. To introduce sub-population heterogeneity at each locus, we sampled from this distribution to produce six new probability distributions, representing sub-populations, with dispersion parameter $\alpha_{\cdot} = (1 - \theta)/\theta$. Finally, from each of the (200×6) sub-population allele distributions, we sampled N subjects (corresponding to $2N$ sampled alleles). We performed this simulation independently for all six combinations of the values of $\theta = 0.003$, 0.017 and 0.030 and $N = 15$ and 30.

The hierarchical model was fit to the six simulated data sets using the program AllDist. For illustrative purposes, we describe in detail the analysis of a single data set, beginning with the necessary inputs to the program. We first specify the number of loci ($L = 200$), the number of sub-populations ($S = 6$), and the number of alleles at each locus ($M_l = 7$ for every locus in this example), and then supply the simulated sample allele count for each cell in this (L stacked tables of dimension $S \times M_l$). For each of the 200 loci, we also need to specify prior counts for each of the alleles at that locus. In practice, this quantity is derived from an external source such as CEPH. In this simulation, the prior distribution for each locus is the same, namely allele counts of (10, 20, 40, 60, 40, 20, 10) from an assessment of 100 individuals. Following the previous discussion of prior distributions, we then specify values for the hyperparameters of $\delta = 4.5$, $\epsilon = 2.0$, $a = 18$, and $b = 6$. Values of the hyperparameters μ_l and τ_l are automatically determined by AllDist using the procedure outlined in Appendix A. Finally, we specify a baseline value $\theta_B = 0.001$, a burn-in period for the Markov chain of 500 iterations, and a number of iterations on which to base inferences of 2,000.

The fundamental output of the program is the estimated frequencies for each allele in each sub-population at each locus. Table II traces the transformation of prior counts into posterior frequency estimates for a single locus. AllDist begins with the (6×7) table of allele counts for the locus. The information provided by these counts is augmented by the prior allele counts for that locus via the *allele count* prior. Additional input about sub-population heterogeneity is provided by the *dispersion prior* stage and the dispersion information at the other observed loci. The net result is an

TABLE II. Details of Estimation for a Single Locus

S	A1	A2	A3	A4	A5	A6	A7	
1	3	2	12	10	2	1	0	Observed counts
6	2	2	7	8	8	2	1	
	10	20	40	60	40	20	10	Prior counts
	4.792	13.645	23.010	29.947	18.225	8.720	6.232	$\hat{\alpha}_l$
1	0.058	0.116	0.260	0.297	0.150	0.072	0.046	Posterior frequencies
6	0.050	0.116	0.223	0.282	0.195	0.080	0.054	

The column "S" labels sub-populations (1 and 6 are depicted), and the columns A1–A7 refer to different alleles.

estimate of the Dirichlet parameter α_l for this locus, which immediately provides posterior frequency estimates for each allele in each sub-population by Equation (1).

The simulated data were also analyzed with three other estimation procedures: pooled allele frequencies (i.e., marginal frequencies, ignoring sub-population distinctions), observed frequencies (gene counting for each sub-population), and the empirical Bayes model (shrinking observed sub-population frequencies toward the marginal frequencies for that locus by an amount that depends only on the data for that locus). The results from these four procedures were compared by a discrepancy statistic $[(200\times 6\times 7)^{-1}\Sigma_{l=1}^{200}\Sigma_{s=1}^{6}\Sigma_{a=1}^{7}(\hat{p}_{asl}-p_{asl})^2]^{1/2}$. For this root mean squared error statistic, smaller values indicates more accurate estimates.

Table III summarizes the simulation results. Before comparing the performance of the estimation methods to one another (the rows of the table), it is instructive to note how the performance of each method varies with θ and N (the columns of the table). As expected, for fixed N, the pooled, empirical Bayes and Bayes frequency estimates perform more effectively as θ decreases. This is owing to the fact that smaller values of θ imply less divergence among sub-populations, a feature exploited by pooling across sub-populations. On the other hand, the performance of the observed frequency estimates is entirely independent of θ because these estimates make no use of the relationships between sub-populations at the same locus. Not surprisingly, however, the estimates do depend heavily on N, with better accuracy as N increases.

TABLE III. Simulation Results Comparing Four Estimation Methods By Root Mean Squared Error From True Allele Frequencies

Simulation	Pooled	Observed	Empirical Bayes	Bayes
$\theta = 0.003$				
$N = 15$	0.031	0.063	0.031	0.028
$N = 30$	0.025	0.044	0.024	0.023
$\theta = 0.017$				
$N = 15$	0.048	0.063	0.044	0.040
$N = 30$	0.045	0.043	0.035	0.033
$\theta = 0.030$				
$N = 15$	0.059	0.062	0.048	0.045
$N = 30$	0.056	0.043	0.038	0.036

Comparisons within the rows of the table indicate that the Bayesian hierarchical model yields more accurate allele frequency estimates than the other methods over the range of θ and N examined. More specifically, the empirical Bayesian and hierarchical Bayesian methods achieve similar advantages over the other estimation methods, especially in the presence of substantial divergence. This is owing to the fact that both of these methods provide moderate shrinkage toward marginal frequencies, as opposed to the extreme behavior of observed frequencies (no shrinkage) and pooled frequencies (total shrinkage). For each combination of N and θ, we also see that the Bayesian estimates offer a slight but systematic improvement over the empirical Bayesian estimates because of the more efficient use of both the data and prior information. As an overall assessment, we see that the greatest power of the Bayesian approach versus the usual observed frequencies is realized when limited data are available (N small) and population divergence is minimal (θ small).

One of the reasons that the Bayesian approach is consistently superior to the empirical Bayes analysis is that estimates of divergence across sub-populations are highly variable when they are based on individual loci, as is the case for the empirical Bayes analysis. The fully Bayesian divergence estimates are more accurate and less variable because they are based on the data from all loci, as well as the prior information. These improved divergence estimates facilitate a more unified degree of shrinkage, which is less affected by sampling variability at each locus.

Figure 1 graphically displays the performance of the Bayesian estimates compared to the observed frequencies. Figure 1 top and bottom each shows the following: for a single allele, we calculate the standardized residual $(\hat{p} - p)/[p(1 - p)]^{1/2}$ for each of the six sub-populations where $\hat{p}$ is an estimated frequency and p is the true frequency for each sub-population. We plot these residuals for $\hat{p}$ calculated from the Bayesian procedure and the observed frequencies. Figure 1 (top) shows an example, taken from the data with $\theta = 0.003$ and $N = 15$, for which the Bayesian procedure is clearly superior to observed frequencies. Figure 1 (bottom), taken from the data with $\theta = 0.030$ and $N = 30$, shows that occasionally the Bayesian estimates offer no improvement over the observed frequencies. In accordance with the numerical results of Table III, the plots for most alleles are more similar to Fig. 1 (top).

Remark: For reference, random samples from European countries yield estimates of θ for STR loci slightly smaller than 0.003 [Chakraborty et al., 1999, and references therein]. Using evolutionary simulations [see Lam et al., 2000] and the above allele distribution, we obtain divergence of slightly less than 0.03 for seven sub-populations given the following conditions: each sub-population was founded by 1,000 individuals drawn at random from the same founder population; sub-populations remain isolated for 100 generations; individuals mate at random within sub-population; and the population grows exponentially to a final size of 50,000. Analyses based on these evolutionary simulations yield results and conclusions similar to those we report here.

DISCUSSION

The Bayesian hierarchical model presented herein provides an answer to the vexing problem of how to avoid biasing the analyses depending on samples from multiple sub-populations when sample sizes are too small to provide reliable sub-

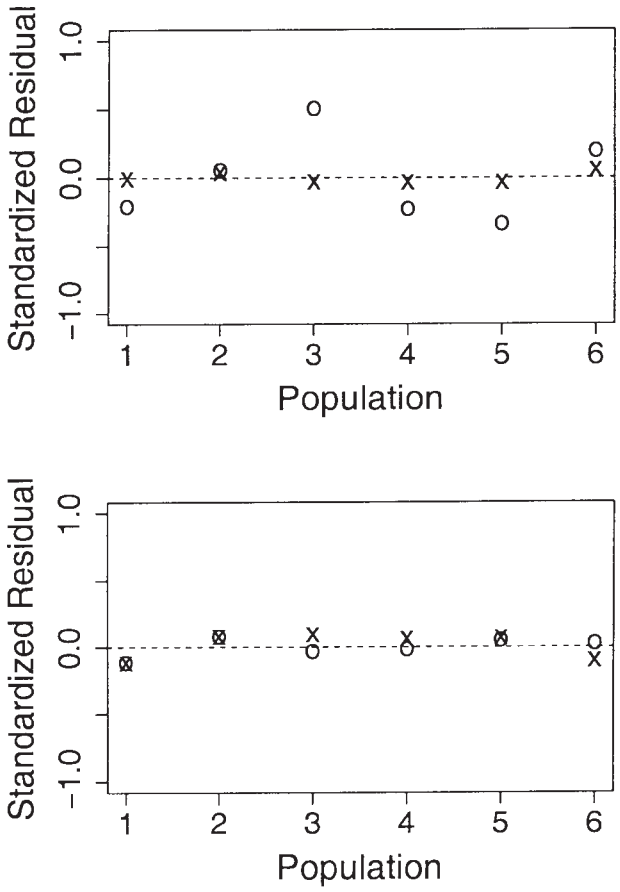

Fig. 1. Plot of standardized residuals comparing the Bayesian estimates to the observed frequencies. **Top:** An example of an allele for which the Bayesian frequency estimates (x) are superior to observed frequencies (o) for all sub-populations. **Bottom:** shows that, on occasion, the Bayesian estimates offer no improvement to observed frequencies. The results for most alleles are closer to those at top, consistent with the fact that the Bayesian procedure has lower mean squared error than using observed frequencies.

population-specific frequency estimates. The answer is appropriate "shrinkage" of sub-population-specific allele frequency distributions toward the distribution obtained by pooling sub-populations and toward published prior distributions. Appropriate shrinkage for a given locus is determined by the joint behavior of all examined loci, prior information on the amount of sub-population divergence, and the variability inherent in the prior distributions of allele frequencies.

Who should care about refined estimates of sub-population-specific allele frequencies? Although such estimates are important for many real-world problems, such as forensic inference from DNA [Roeder, 1994; Roeder et al., 1998], we target linkage analyses that draw families from different sub-populations. As described previously, the ideal linkage analysis uses sub-population-specific allele frequencies, yet many investigators use estimates derived from the pooled population. There are ac-

tually arguments for both approaches, as revealed by our simulations (Table III). With larger samples and non-negligible sub-population sub-structure, the observed sub-population frequencies are closer to their true values than are pooled frequencies, whereas the opposite occurs for smaller samples from more homogeneous sub-populations. These results are simple to understand from the perspective of shrinkage estimators. There are two sources of variability to consider: sampling variance and population sub-structure. Although population sub-structure does not depend on sample size (N), sampling variance does. For small N, sampling variance often will dominate population divergence and therefore a single pooled estimate will be closer, on average, to each sub-population's allele frequency distribution because it is an (unadaptive) shrinkage estimator. As N increases, however, the converse must occur.

A Bayesian hierarchical model is a natural way to navigate this cross-current. It adaptively shrinks sub-population-specific allele distributions, with the amount of shrinkage depending on the data and the strength of prior studies. If samples from each sub-population are large relative to the amount of sub-population sub-structure, shrinkage will be minimal. On the other hand, when only limited data are available, the shrinkage is more formidable. The net result is that the Bayesian procedure automatically yields the most efficient estimation, moderating the degree of shrinkage in a manner consistent with characteristics of the available data.

Surprisingly, despite the oft-documented sensitivity of linkage analysis to estimated allele frequencies, especially affected relative pair analysis, we are unaware of any linkage studies using adaptive shrinkage estimators. We suspect two impediments have contributed to this result: limited software to obtain such estimates and the fact that linkage software often cannot automatically accommodate such a procedure. Regarding the former impediment, we implemented both the Bayesian and empirical Bayes estimators in a C program, AllDist, which has structure and format requirements similar to programs such as GeneHunter [Kruglyak et al., 1996]. AllDist also provides estimates of dispersion for each locus.

Regarding the latter impediment, combining sub-population-specific results for standard linkage methods is trivial because one simply adds the LOD scores. Combining the results from non-parametric analyses such as the NPL scores for GeneHunter is not much more challenging. In fact, we provide a program to combine GeneHunter results over sub-populations at our URL listed below.

Our method makes explicit use of estimated allele distributions, derived from prior population studies, to refine estimates for the current set of sub-populations. Ideally, the prior and data distributions would be derived from the same population or, at the least, from closely related populations. Because of the limited availability of publicly available databases (e.g., http://watson.hgen.pitt.edu/AlleleFreq.html), this feature may be less valuable than it could be. Our hope is that data from many laboratories, as well as independent analyses from the same laboratory, will soon be placed on the Web. In addition, other prior information could be quite useful for other aspects of linkage analysis, such as estimated recombination fractions. Publication of these data would help refine linkage analyses for the entire research community at little cost to individual researchers.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants MH57881 and MH56193 and National Science Foundation grant DMS-9803433. The program AllDist and related materials can be obtained at http://wpicr.wpic.pitt.edu/wpiccompgen/ or http://www.stat.cmu.edu/~roeder/.

REFERENCES

Chakraborty R, Jin L. 1992. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum Genet 88:267–72.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B. 1999. The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis 20:1682–96.

Daniels M, Kass R. 1997. A note on first-stage approximation in two-stage hierarchical models. Technical Report 663. Pittsburgh: Carnegie Mellon University.

Devlin B, Risch N, Roeder K. 1991. Estimation of allele frequencies for VNTR loci. Am J Hum Genet 48:662–76.

Friemer N, Sandkuijl L, Blower S. 1993. Incorrect specification of marker allele frequencies: effects on linkage analysis. Am J Hum Genet 52:1102–10.

Gelman A, Carlin J, et al. 1995. Bayesian data analysis. London: Chapman and Hall.

Geman A, Geman D. 1984. Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. IEEE Trans Pattern Anal Machine Intell 6:721–41.

Gilks WR, Richardson S, et al. 1996. Markov chain Monte Carlo in practice. London: Chapman and Hall.

Jiang CJ, Cockerham CC. 1987. Use of the multinomial Dirichlet model for analysis of subdivided genetic populations. Genetics 115:363–6.

Johnson N, Kotz S, et al. 1997. Discrete multivariate distributions. New York: John Wiley and Sons, Inc.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–63.

Lam JC, Roeder K, Devlin B. 2000. Haplotype fine-mapping by evolutionary trees. Am J Hum Genet 66:659–73.

Lange K. 1995. Applications of the Dirichlet distribution to forensic match probabilities. Genetica 96:107–17.

Levin B, Reeds J. 1977. Compound multinomial likelihood functions are unimodal: proof of a conjecture of I.J. good. Ann Stat 5:79–87.

Mosimann J. 1962. On the compound multinomial distribution, the multivariate beta distribution, and correlation among proportions. Biometrika 49:65–82.

Powell M. 1977. Restart procedure for the conjugate gradient method. Math Program 12:241.

Ott J. 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283–90.

Roeder K. 1994. DNA fingerprinting: a review of the controversy (with discussion). Stat Sci 9:222–78.

Roeder K, Escobar M, Kadane JB, Balazs I. 1998. Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 85:269–87.

Risch N. 1990. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–53.

Risch N, Giuffra L. 1992. Model misspecification and multipoint linkage analysis. Hum Hered 42:77–92.

Schafer J. 1997. Analysis of incomplete multivariate data. New York: Chapman and Hall.

Schervish MJ. 1995. Theory of statistics, 2nd ed. New York: Springer.

Skellam J. 1948. A probability distribution derived from the binomial distribution by regarding the probability of a success as variable between the sets of trials. JRSS(B) 10:257–61.

Tores F, Barillot E, Philippi A, et al. 1996. Affected sib-pair strategies for late onset diseases: type I error and marker allele frequencies. Genet Epidemiol 13:202.

Weeks DE, Lange K. 1988. The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–26.

Wright S. 1951. The genetical structure of populations. Ann Eugen 15:159–71.

APPENDIX A: PRIOR DISTRIBUTIONS

To derive values for the vectors $\boldsymbol{\mu}_l$ and the matrices $\boldsymbol{\tau}_l$, which describe the prior for $\boldsymbol{\psi}_l$, we first derive a prior for $\boldsymbol{\alpha}_l$, given $\alpha_{.l}$ and then back-transform. We utilize an estimate of $\boldsymbol{\alpha}_l$, $(\alpha_{.l}\mathbf{Y}_l)/n_l$, as the mean of the prior for $\boldsymbol{\alpha}_l$. Similarly, the covariance matrix of the prior for $\boldsymbol{\alpha}_l$ is derived from the covariance of its estimator. Note that $\text{var}(\mathbf{Y}_l) = \mathbf{H}(n_l + \alpha_{.l})/(1 + \alpha_{.l})$, where $\mathbf{H}$ is the covariance matrix of a multinomial random vector of sample size n_l with cell frequencies $\boldsymbol{\alpha}_l/\alpha_{.l}$ [Johnson et al., 1997]. Consequently, a natural choice for the covariance of the prior for $\boldsymbol{\alpha}_l$ is

$$\widehat{\boldsymbol{H}}\left(\frac{n_l+\alpha_{.l}}{1+\alpha_{.l}}\right)\frac{\alpha_{.l}^2}{n_l^2}$$

where $\widehat{\boldsymbol{H}}$ is the multinomial covariance defined above but with the population cell frequencies $\boldsymbol{\alpha}_l/\alpha_{.l}$ replaced with the sample cell frequencies $\mathbf{Y}_l/n_l$. Formally we can think of building the prior in two stages. Start with an implicit diffuse prior regarding the distribution of α_l, which does not favor any particular distribution. Then, the prior data, described under the Prior Distribution section, is used to update the diffuse prior to obtain a "pre-experiment" posterior distribution. This posterior then becomes the prior for the experiment of interest, estimating the allele frequencies for locus l and subpopulation s. To translate these results into values for $\boldsymbol{\mu}_l$ and $\boldsymbol{\tau}_l$, we apply first-order approximations based on the transformation that maps the vector $\boldsymbol{\alpha}_l$ to $\boldsymbol{\psi}_l$:

$$\mu_l = \left[\log\left(\frac{Y_{1l}}{Y_{M_l l}}\right),\ldots,\log\left(\frac{Y_{(M_l-1)l}}{Y_{M_l l}}\right)\right]$$

$$\tau_l^{-1} = \boldsymbol{D}_l\left(\widehat{\boldsymbol{H}}\left(\frac{n_l+\alpha_{.l}}{1+\alpha_{.l}}\right)\frac{\alpha_{.l}^2}{n_l^2}\right)\boldsymbol{D}_l^T$$

where $\mathbf{D}_l$ is a matrix with (a, s) entry $\partial g_{al}/\partial\alpha_s$ evaluated at the estimate for $\boldsymbol{\alpha}_l$ and $g_l = (g_{1l},\ldots,g_{Ml})$ is the transformation that maps $\boldsymbol{\alpha}_l$ to $(\boldsymbol{\psi}_l, \beta_l)$. Note that the value for $\boldsymbol{\mu}_l$ does not depend on the fixed value of $\alpha_{.l}$.

To motivate the procedure just described for obtaining hyperpriors from the sampled allele frequencies, it is illustrative to show a fundamental link between the Bayesian and frequentist perspectives. Let Θ be a p-dimensional parameter and let $\widehat{\Theta}_n$ be a maximum likelihood estimator of Θ based on X^n, a random vector of n observations with some conditional distribution given Θ. Finally, let $I(\widehat{\Theta}_n)$ be the $(p \times p)$ observed Fisher information matrix based on X^n. Then, under fairly general regularity conditions, the following two statements are true:

1. Frequentist asymptotic normality of the MLE: The conditional distribution of $\widehat{\Theta}_n$ given Θ converges to $N[\Theta, I^{-1}(\widehat{\Theta}_n)]$

2. Bayesian asymptotic normality of the posterior distribution: The conditional distribution of Θ given X^n converges to $N[(\widehat{\Theta}_n, I^{-1}(\widehat{\Theta}_n)]$

In other words, the asymptotic distribution of $I^{-1/2}(\hat{\Theta}_n)$ $(\Theta - \hat{\Theta}_n)$ is p-dimensional standard normal regardless of whether we are treating Θ or $\hat{\Theta}_n$ as a random variable. In our implementation (AllDist), this result allows us to use the asymptotic covariance of a function of $\mathbf{Y}_l$ given $\boldsymbol{\alpha}_l$ as a covariance for $\boldsymbol{\alpha}_l$ when $\boldsymbol{\alpha}_l$ is considered random.

APPENDIX B: INFERENCE

The approximate Bayesian model treats $\hat{\boldsymbol{\alpha}}_l$ as the data. For superior finite sample properties, the approximate distribution for $\hat{\boldsymbol{\alpha}}_l$ can be translated directly into an approximate distribution for $(\hat{\boldsymbol{\psi}}_l, \hat{\beta}_l)$ through the delta method:

$$(\hat{\psi}_l, \hat{\beta}_l)^T \mid (\psi_l, \beta_l)^T \sim N_{M_l}\left\{(\psi_l, \beta_l)^T, \boldsymbol{D}_l I_{\alpha_l}^{-1}[g_l^{-1}(\hat{\psi}_l, \hat{\beta}_l)]\boldsymbol{D}_l^T\right\}.$$

Hence, the revised model has the following stages:

Substructure and Allele Count Stages

Independently for $l = 1,...,L$

$$\left(\hat{\psi}_l, \hat{\beta}_l\right)^T \mid ((\psi_l, \beta_l)^T \sim N_{M_l}\left((\psi_l, \beta_l)^T, \boldsymbol{C}_l\right),$$

where $\boldsymbol{C}_l = \boldsymbol{D}_l I_{\alpha_l}^{-1}(g_l^{-1}(\hat{\psi}_l, \hat{\beta}_l))\boldsymbol{D}_l^T$.

Allele Count Prior

Independently for $l = 1,...,L$

$$(\psi_l, \beta_l)^T \mid \mu_l, \tau_l, \mu_\beta, \tau_\beta \sim N_{M_l}\left(\boldsymbol{U}_l, \boldsymbol{T}_l^{-1}\right),$$

where

$$\boldsymbol{U}_l = (\mu_l, \mu_\beta)^T, \quad T_l = \begin{bmatrix} \tau_l & \mathbf{0} \\ \mathbf{0}^T & \tau_\beta \end{bmatrix}$$

and $\mathbf{0}$ is a vector of length $M_l - 1$ of zeros.

Dispersion Prior

Independently, $\mu_\beta|\delta, \epsilon \sim N_1(\delta, \epsilon^{-1})$ and $\tau_\beta|a,b \sim \Gamma(a,b)$. Notice that the *allele count* and *sub-structure* stages have been collapsed to form the first stage, whereas the stages describing the prior input are identical to the original model. The model results in the following full conditional distributions, where the term *rest* to the right of a conditioning bar denotes all parameters in the model except those to the left of the bar. With these conditional distributions, the Gibbs or successive substitution sampling can be used to estimate the parameters of the Bayesian model.

1. Independently for $l = 1...,L$

$$(\psi_l, \beta_l)^T \mid rest,\ data \sim N_{M_l}\left\{(\boldsymbol{C}_l^{-1} + \boldsymbol{T}_l)^{-1}[\boldsymbol{C}_l^{-1}(\hat{\psi}_l, \hat{\beta}_l)^T + \boldsymbol{T}_l \boldsymbol{U}_l], (\boldsymbol{C}_l^{-1} + \boldsymbol{T}_l)^{-1}\right\}$$

2. $$\mu_\beta \mid rest,\ data \sim N_1\left[\frac{\epsilon\delta + \tau_\beta \Sigma_{l=1}^{L} \beta_l}{\epsilon + L\tau_\beta}, (\epsilon + L\tau_\beta)^{-1}\right]$$

3. $$\tau_\beta \mid rest,\ data \sim \Gamma\left[a + \frac{L}{2}, b + \frac{1}{2}\sum_{l=1}^{L} (\beta_l - \mu_\beta)^2\right]$$

APPENDIX C: UNDERDISPERSION

Insofar as we are aware, there is no known verifiable condition for determining whether data are underdispersed for fitting a compound-multinomial model. We base our determination of underdispersion on the behavior of a modified Powell hybrid mode-finding algorithm [Powell, 1977]. Levin and Reeds [1977] show that, for every fixed vector of cell frequencies, the compound-multinomial likelihood function is a unimodal function of the dispersion parameter, where the mode may be $+\infty$. A sensible initial parameter value of $\boldsymbol{\alpha_l}$ for the optimization outline is the vector of marginal frequencies times a relatively small value of $\alpha_{.l}$. The algorithm then progresses in a neighborhood of the marginal frequency "ridge," the major steps being taken to adjust the dispersion parameter. The practical import is that if the algorithm drives the dispersion parameter toward $+\infty$, the data are extremely likely to be underdispersed. Alternatively, when a finite MLE exists, it is located quickly. Based on this behavior, our determination of underdispersion proceeds by picking a baseline dispersion parameter defined by $\alpha_B = (1 - \theta_B)/\theta_B$ and starting the algorithm as described. If the algorithm moves beyond α_B from an initial staring value, we repeat the algorithm at 10 different starting values. We declare underdispersion if the algorithm attempts to move beyond α_B from all 10 of the additional trials.

When a proper MLE cannot be found, it is necessary to reduce the dimension of the problem to obtain numerical stability. Fixing $\alpha_{.l} = \alpha_B$, we find the MLE for $\boldsymbol{\psi_l}$ and the corresponding asymptotic covariance matrix $\mathbf{G_l} = I_{\psi l}^{-1}(\hat{\boldsymbol{\psi}}_l)$. With $\alpha_{.l}$ fixed at α_B, the model becomes considerably simpler and the stages of the model reduce as follows:

Sub-structure and Allele Count Stages

Independently for $l = 1,...,L$,

$$\hat{\psi}_l^T \mid \psi_l^T \sim N_{M_l - 1}(\psi_l^T, \boldsymbol{G}_l).$$

Allele Count Prior

Independently for $l = 1,...,L$, $\psi_l^T \mid \mu_l, \tau_l \sim N_{M_l - 1}(\mu_l, \tau_l^{-1})$. Inferences are based on a single full conditional distribution that can be computed analytically.

Independently for $l = 1,...,L$

$$\psi_l^T \mid rest,\ data \sim N_{M_l-1}\left\{(\boldsymbol{G}_l^{-1}+\tau_l)^{-1}[\boldsymbol{G}_l^{-1}\hat{\psi}_l^T+\tau_l\mu_l],(\boldsymbol{G}_l^{-1}+\tau_l)^{-1}\right\}.$$

Using the posterior mean of $\boldsymbol{\psi}_l$ along with the baseline dispersion α_B, we then solve for $\boldsymbol{\alpha}_l$ and proceed using (1) to estimate the allele frequencies for each sub-population.

This treatment of underdispersion has considerable justification provided that α_B is chosen sensibly. Although the estimate of $\boldsymbol{\alpha}_l$ is clearly sensitive to the choice of α_B, this sensitivity is not transferred to the estimate of $\boldsymbol{\psi}_l$ because the latter is not a function of the dispersion parameter. Based on the likelihood surface, $\alpha_{.l}$ is clearly very large, suggesting the sub-populations have not diverged at locus l. Arbitrarily setting $\alpha_{.l}$ to a large value has little impact on the resulting inferences.