Genome Scanning by Composite Likelihood

Newton Morton, Nikolas Maniatis, Weihua Zhang, Sarah Ennis, and Andrew Collins

Ambitious programs have recently been advocated or launched to create genomewide databases for meta-analysis of association between DNA markers and phenotypes of medical and/or social concern. A necessary but not sufficient condition for success in association mapping is that the data give accurate estimates of both genomic location and its standard error, which are provided for multifactorial phenotypes by composite likelihood. That class includes the Malecot model, which we here apply with an illustrative example. This preliminary analysis leads to five inferences: permutation of cases and controls provides a test of association free of autocorrelation; two hypotheses give similar estimates, but one is consistently more accurate; estimation of the false-discovery rate is extended to causal genes in a small proportion of regions; the minimal data for successful meta-analysis are inferred; and power is robust for all genomic factors except minor-allele frequency. An extension to meta-analysis is proposed. Other approaches to genome scanning and meta-analysis should, if possible, be similarly extended so that their operating characteristics can be compared.

Like other sciences, genetic epidemiology is both limited and driven by the techniques at its command. For nearly a century, gene localization was dominated by linkage and cytogenetics, with little opportunity to map a gene through associated markers. Finally, short physical maps of regions identified by linkage provided a basis for localization of rare major genes in haplotypes.[1] The Malecot model was useful for this purpose,[2,3] with subsequent extension to oligogenes and diplotypes.[4,5] Maturity, if not completion, of the Human Genome Project accelerated this development by providing a physical map, which led, in turn, to genetic maps in linkage disequilibrium units (LDUs).[4] Various efforts, including the HapMap Project, undertook to provide evidence on marker diversity, with the goal of using that information to localize disease genes and to investigate other aspects of the diversity revealed by genetic polymorphism.[6–8]

After 3 years of LDU development, map construction is now rapid and accurate. Theory and practice for allelic association in small regions appear stable, but further progress will come from experience with scans of much longer tracts. From their upper limit, they have been called "genome scans," a misnomer, since a large tract, chromosome, or set of chromosomes delimited without regard to association is analyzed in the same way: by division into contiguous, nonoverlapping regions. Whatever the terminology or method, such a scan is only stage 1 in a multistage design, since the later stages are concerned with regions selected for evidence of association. The number of markers in a dense scan is extremely large compared with the number of causal sites likely to be detected even in a large sample, whether or not supported by functional tests. This is a challenge to the false-discovery rate (FDR), originally introduced for localization of major loci by linkage.[9] Unfortunately, it cannot be conventionally calculated when the probability of the null hypothesis approaches 1.[10,11] In that situation, alternative corrections have been developed to determine the real significance corresponding to nominal significance in a large number of tests.

Like its predecessors in small regions, genome scans use composite likelihood that adds together individual component log likelihoods, each of which corresponds to a marginal or conditional event.[12] Each component is a function of location on a map in LDUs or, less efficiently, on a correlated linkage or physical map.[13] This use of composite likelihood combines three different statistical problems. First, most statistical hypotheses in genetics are composite, in the sense that the hypothesis is "composed" of a group of simple hypotheses that may specify gene frequency, effect, or location, which results in a treble infinity of possible values. Composite likelihood can estimate location conditional on other variables or concurrently. Second, the number of markers varies even among regions of the same length, with corresponding variation in the estimate of composite likelihood. Third, markers in proximity are not independent but autocorrelated to an extent that is partially predictable from their LDU location but less accurately from their physical location. The effect of this autocorrelation is small (although perhaps not negligible) at low resolution but increases to an unknown extent with marker density. Until this limitation is removed, the genetic epidemiologist must choose between uncertain reliability at high density and loss of power by heavy selection of markers (commonly called "tag SNPs" because most of them are SNPs). We shall now show how these difficulties can be resolved, both in single studies and in meta-analyses of unbiased reports, recently advocated or launched without identification of the metrics that provide reliable combination of evidence.[8,14]

From the Human Genetics Division, University of Southampton, Southampton General Hospital, Southampton, United Kingdom

Received August 18, 2006; accepted for publication October 24, 2006; electronically published December 5, 2006.

Address for correspondence and reprints: Dr. Newton Morton, Human Genetics Division, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, United Kingdom. E-mail: nem@soton.ac.uk

Am. J. Hum. Genet. 2007;80:19–28. 0002-9297/2007/8001-0004$15.00

Table 1. Frequencies Observed and Expected in Random Samples of n Haplotypes ($ad - bc \geqslant 0$, $b \leqslant c$)[a]

Status and Sample	Allele G	Allele g	Total
Affected or normal:			
Count	a	b	a + b
Probability	$f[z + (1 - z)R]$	$f(1 - z)(1 - R)$	f
Normal or affected:			
Count	c	d	c + d
Probability	$(R - f)z + R(1 - f)(1 - z)$	$(1 - R)[z + (1 - f)(1 - z)]$	$1 - f$
Total:			
Count	a + c	b + d	n
Probability	R	$1 - R$	1

[a] See the work of Maniatis et al.[5]

Methods

Maps in LDU under the Malecot Model

All n/2 pairs of codominant diallelic diplotypes under random mating can be reduced from a 3 × 3 table to a 2 × 2 count of haplotype frequencies

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

that, by interchange of rows and/or columns, satisfy $ad - bc \geqslant 0$ and $b \leqslant c$.[4] The optimal measure of allelic association is

$$\hat{\rho} = \frac{ad - bc}{(a + b)(b + d)} \ ,$$

with information

$$K_\rho = \frac{n(a + b)(b + d)}{(a + c)(c + d)}$$

under the null hypothesis that $\rho = 0$, which is tested by $\chi_1^2 = \hat{\rho}^2 K_\rho$.[3] The Malecot model predicts ρ as $(1 - L)Me^{-\Sigma\varepsilon_h d_h} + L$, where d_h is the distance (in kilobases [kb]) between adjacent markers h and h + 1. This is an enhancement of the less reliable prediction from the physical map that approximates distance by $\varepsilon\Sigma d_h$.[2] The composite likelihood over all markers in a region is $lk = e^{-\Lambda/2}$, where $\Lambda = \Sigma K_\rho(\hat{\rho} - \rho)$.[2] Given a physical map, the parameters M, L, and ε are estimated from pairs of diallelic markers. The estimate of ε in the physical map is small and highly variable. The LDU map estimates M, L, and the ε_h that represent the block-and-step pattern of LD. Optionally, the exponent can be taken as $\varepsilon\Sigma\varepsilon_h d_h$, where ε is estimated for a candidate region, but ε is usually too close to 1 to warrant refinement. These calculations are performed by the LDMAP program,[5] the current version of which exploits grid technology to accommodate the high SNP density in HapMap.

Association Mapping with the Malecot Model

Quantitative traits can be studied by regression, giving scope under some sampling schemes to valid covariance analysis and inference of gene-environment interaction. Quantitative traits are especially frequent for anthropometrics, behavior, and response to drugs or other pharmaceutical agents but can occur in any branch of human genetics. Commonly, however, phenotypes are dichotomous; the two classes are "affected" and "normal" if sampled at random or "cases" and "controls" if sampled selectively. Controls may be random, matched with cases for age and other variables, or hypernormal. The last is more powerful but difficult to analyze by regression, because the sample-enrichment factor is poorly specified and environmental covariates may be distorted. For example, hypernormal controls selected to be older than cases do not imply that affection decreases with age.

In the simplest case, affection is determined by a rare dominant or recessive gene and is studied in families so that the disease-associated allele can be recognized.[2] Alternatively, inheritance may be more complex, and then alleles are assumed to be additive. This is not restrictive, because Maclauren's calculus theorem predicts that causal markers of small effect tend to approach additivity, which is enhanced for predictive but noncausal SNPs by recombination. This simplifying assumption reduces diplotype data to a 2 × 2 table of allele counts by affection status,[4] deferring more elaborate analysis until causal SNPs are identified. In this way, n haplotypes are scored from n/2 diplotypes. The association parameter is $z = \gamma\rho$, where $\gamma = Qw/f$ is the attributable risk under additivity, Q is the frequency of an allele predisposing to affection, f is the frequency of affected diplotypes, w is the penetrance in GG homozygotes with corresponding penetrance w/2 in Gg heterozygotes, and ρ is the association probability when $\gamma = 1$ (table 1).[2,5] The score for z is

$$U_z = \frac{\partial \ln lk}{\partial z} = \frac{(ad - bc)n}{(a + c)(c + d)} \ ,$$

with information

$$K_z = \frac{n(a + b)(b + d)}{(a + c)(c + d)} \ ,$$

which is formally the same as that for K_ρ in the previous section, but the counts are different. Then,

$$\hat{z} = \frac{U_z}{K_z} = \frac{ad - bc}{(a + b)(b + d)}$$

and $\chi_1^2 = \hat{z}^2 K_z$. Since SNPs with $(a + c)(c + d) = 0$ or $(a + b)(b +$

d) = 0 are omitted because their information K_z is 0 or indeterminate, K_z must satisfy

$$0<K_z=\frac{\chi_1^2}{\hat{z}^2}\leqslant n\ .$$

At present, we do not consider an enrichment factor, which combines with affection status to create complications not met in construction of LD maps or mapping of rare genes with high penetrance.[2] Omission of an enrichment factor does not violate constraints on observed counts or K, but it may not be optimal.

However the enrichment issue is resolved, association mapping accepts kb and LD maps and estimates parameters M and L appropriate to the relationship of affection status with diallelic markers and, in addition, the estimated location $\hat{S}$ and SE of a causal marker, regardless of whether that marker was included in the data set. For this analysis, $\varepsilon\Sigma d_h$ in the physical map or $\varepsilon\Sigma\varepsilon_h d_h$ in the LDU map is replaced by $\varepsilon\Delta(S_h - S)$. Since lk is a composite likelihood, the parameters but not the error variance can be estimated by minimizing Λ as for true likelihood. Experience has shown that ε cannot be estimated reliably during association mapping of S, whereas the accuracy with which L can be estimated increases with the length of the LDU region.[5] We therefore examined three analyses that do not attempt to modify ε or a standard LD map derived from pairs of markers (fig. 1). The simplest one compares A and C, where A assumes no causal marker in the region, with L predicted and $M = 0$; the alternative C takes the same value of L but estimates M and S. The second analysis compares A with D, which estimates L, M, and S. The third analysis takes B under the null hypothesis that $M = 0$ with L estimated and compares it with D. This has the least power, because B confounds H_1 with L and will not be considered here.

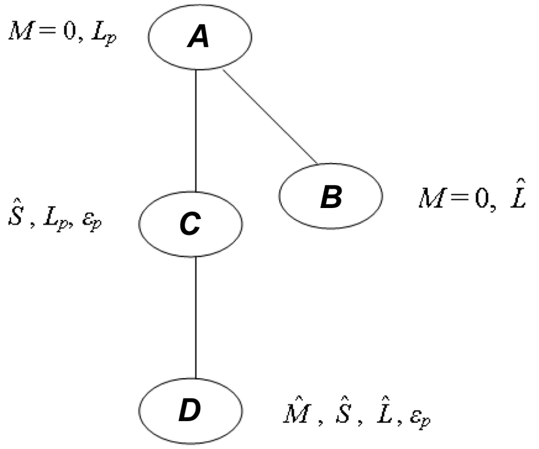
$M=0, L_p$
A
B
$M=0,\ \hat{L}$
$\hat{M}\ ,\ \hat{S}\ , L_p, \varepsilon_p$
C
D
$\hat{M}\ ,\ \hat{S}\ ,\ \hat{L}, \varepsilon_p$

Figure 1. Subhypotheses under the Malecot model. A circumflex (^) indicates a parameter that is estimated in association mapping, and the subscript p indicates a parameter estimated from other data. L is the asymptote at large distance, M is the increment maximized at 0 distance, S is the location of the causal SNP (with the assumption that there is one in the region), and ε is the coefficient of distance near S in the physical or LDU map. The two most frequent tests are AC = A versus C and AD = A versus D.

Control of Type I Error

Under the null hypothesis H_0 that there is no causal marker in a region j with m_j markers, composite likelihood analyses provide an estimate of $x_{ij} = \Lambda_{Aij} - \Lambda_{Cij}$ or $\Lambda_{Aij} - \Lambda_{Dij}$ for the ith replicate in the jth region ($i = 1, \ldots r_j$), designated as AC_{ij} or AD_{ij}, respectively. Each replicate is obtained by shuffling, so that each individual is assigned randomly and without replacement to an observed phenotype, whether case/control, affected/normal, or quantitative. If SNPs were independent, the error variance in a region would be estimated for each replicate by $V_C = \Lambda_C/(m-2)$ and V_D by $\Lambda_D/(m-3)$, and this would provide an estimate of χ_2^2 or χ_3^2 from which the significance level P would be derived. Since autocorrelation violates the independence assumption to an extent that increases with SNP resolution, we must use different estimates of these quantities to recover P_{ij} for replicates beginning with the fractional rank r_{ij} within the jth region, which equals the fractional rank of $1/x_{ij}$, reversing the x_{ij} order. If nearly every P_{ij} is unique within region j, the mean approaches 1/2, and the variance approaches 1/12 as $r_j \rightarrow \infty$, corresponding to the uniform distribution that is a special case of a beta distribution. Then, a conventional estimate[15] of P_{ij} is

$$\frac{r_{ij}-\frac{1}{3}}{r_j+\frac{1}{3}}\ ,$$

from which χ^2_{ij2} or χ^2_{ij3} is calculated by g01fcc of the National Algorithm Group (NAG), and the variance V_{ij} is estimated as

$$\frac{x_{ij}}{\chi^2_{ij}}\ ,$$

which is not monotonic on x_{ij}, χ^2_{ij}, or P_{ij}. There is no loss under H_0, which relies exclusively on P_{ij}.

The situation is different for a single sample in a region, j, where H_1 may be true, which gives only a single value x_j for each analysis in that region. We estimate its error variance from replicates under H_0 by fitting the regression model $\ln V_{ij} = b_1 + b_2 \ln x_{ij}$ in a subregion centered as far as possible on the H_1 value of x_j, with up to 20 values of x_{ij} on each side. V_j under H_1 is estimated as $\exp(b_1 + b_2 \ln x_j)$, which gives $\chi_j^2 = x_j/V_j$ for χ_2^2 or χ_3^2, with corresponding P_j from the g01ecc NAG subroutine. By estimation of V_j from replicates under H_0, any effect of autocorrelation is avoided.

As the first step toward genome scanning, a program called CHROMSCAN was constructed to perform the operations described above for tracts, chromosomes, or a whole genome (A.C., unpublished data). The region can be any length, with a 10 LDU default. The minimal number of SNPs is also specified, with a default of 30. Any number of replicates within a region can be chosen for evaluation of real data that contribute a single estimate to each region. Significance tests for association are based entirely on P, but the error variance V_j is used to compute an SE for S_j. Starting at converged values and with use of exact derivatives, simultaneous estimates of S_j and nuisance parameters that may include M_j and/or L_j provide an information matrix that is inverted to give the nominal variance K_{SS}^{-1}. Then, the information K_j about S_j is

$$\frac{1/K_{SS}^{-1}}{V_j/df}\ ,$$

where df = 2 or 3 according to the model, and the corresponding SE is[2]

$$\mathrm{SE}(\mathrm{S_j}) = \sqrt{1/\mathrm{K_j}}\ .$$

Its reliability under autocorrelation may be confirmed when the true value of S is known simply by fitting the nuisance parameters M and/or L at that value. Then the difference in χ^2 evaluated at $\mathrm{V_i}$ is χ^2_{1j}, and

$$\mathrm{K_j} \approx \frac{\chi^2_{1j}}{(\hat{\mathrm{S}}_j - \mathrm{S_j})^2}\ .$$

Estimating the FDR

The concept of an FDR was introduced to map major genes by polymorphisms at single loci, which led to the conclusion that a LOD of at least 3 is required to assure an FDR <0.05.[9] Forty years later, an analogy with Brownian motion extended this argument to a genome scan by linkage, with the conclusion that only a modest increase of the critical LOD (to 3.3) gives the same FDR with a major locus, increasing for complex inheritance to 3.6 for affected sib pairs and to 3.8 for affected second cousins.[16] Association mapping presents a different problem, since relatives of probands are usually excluded to generate large samples without requiring pedigree information. This sampling strategy is efficient for either a cohort or cases and controls and is especially useful for diseases of late onset, where DNA from one or both parents is often not available. Here we extend the FDR theory to association mapping and apply it to composite likelihood for unrelated individuals. In principle, our approach is applicable to any sampling strategy, even if standard FDR methods fail because the distribution of the nominal significance under the null hypothesis $\mathrm{H_0}$ is not uniform and/or the prior probability of $\mathrm{H_0}$ may approach 1. We assume that simulation as shown in the "Control of the Type I Error" section creates a uniform distribution under $\mathrm{H_0}$ and that the prior probability of $\mathrm{H_0}$ approaches 1. The material to be analyzed may be a tract, a chromosome, a set of chromosomes, or a genome. However defined, it is divided into nonoverlapping regions. When values of χ^2_1 are pooled under the assumption of a Poisson distribution of extreme significance levels, the critical significance level $P = \alpha$ satisfies $\alpha\mathrm{N} = .05$, where N in the absence of compelling evidence of a causal locus in a defined subset of regions is the number of regions in a genome scan ($\alpha \rightarrow 0$, $\mathrm{N} \rightarrow \infty$). The FDR corresponding to α is

$$\frac{\alpha(1-\phi)}{\alpha(1-\phi)+\Omega\phi} = \frac{\alpha}{\alpha+\Omega\phi/(1-\phi)}\ ,$$

where Ω is the power to reach significance under $\mathrm{H_1}$ and ϕ is the probability that $\mathrm{H_1}$ is true whether significant or not. In genome scans ϕ and $\Omega\phi$ are of order 1/N or less, which makes this estimate impractical. Fortunately, we have independent estimates of the numerator and denominator, as illustrated in the "Results" section. Their ratio is a good estimate of the FDR if a uniform distribution holds under $\mathrm{H_0}$ and the number of $\mathrm{H_1}$ regions is large enough to determine Ω.

Presentation of Evidence

Stage 1 in association mapping is currently defined as "a scan of one or more nonoverlapping regions, including but not limited to regions suggested by linkage, function, or cytogenetics." The evidence presented for stage 1 sets a pattern and limit for subsequent meta-analysis, so its presentation is critical. Location would ideally be expressed in LDUs, which, in the few published trials, have been shown to be more efficient for association mapping than for physical units (kb).[5,13] However, the latter have approached stability with revisions that are increasingly minor, infrequent, and generally accepted. LD maps, on the contrary, are more recent, with rapidly increasing density and potential changes in types and allele frequencies of markers. Although public (Genetic Epidemiology Group), these maps are not part of the HapMap database[17] and are not exploited by all investigators. Therefore, until consensus is reached, it is necessary to convert evidence from LD maps into more stable but reportedly less efficient locations and SEs (in kb) that the CHROMSCAN program also provides.

In addition to extensive detail about reference maps (kb and LDU), number of regions studied, the basis for their choice, and the markers and methods of analysis, we assume a file with one record for each region within a given chromosome and the critical variables, where location is expressed as v.u for v kb and u additional bp (table 2). The primary and derived locations in both maps must be clearly indicated. The SE (expected to be consistently smaller than the estimate from a kb map) is the product of its estimate from the covariance matrix for composite likelihood and kb/LDU, where kb denotes the physical interval corresponding to an appropriate interval in LDU. Tentatively, we assume that the 95% confidence interval as ±1.96 SE is appropriate as a compromise between an interval so small that LDU = 0 and one so large that blocks and steps outside the confidence interval are included. This structure allows regions to be sorted by any variable. For publication, only the smallest *P* values for composite likelihood would usually be given, together with larger values that relate to other claims. However, selective reporting of a region biases meta-analysis.

Given s independent and unbiased samples for a given region,

Table 2. Essential Data for Meta-Analysis of a Region

Variable	Value	Kb and LDU
Source	ID	…
Chromosome	1…22, X, or Y	…
No. of SNPs	$\mathrm{m_j}$	…
First location in region	…	*
Last location in region	…	*
Composite likelihood[a]:	…	*
P	…	*
Estimated location (S)	…	*
SE	…	*
Information (K)	…	*
Most significant marker:	*	*
Nominal χ^2_1	*	…

NOTE.—An asterisk (*) indicates that the data are often incomplete.

[a] Linkage data should be expressed in this form, but SEs are less reliable and often cover multiple LDU regions; usually S, SE, and K are not estimated.

and with the assumption that the same kb map was used for all samples, the simplest meta-analysis gives

$$\bar{S} = \frac{\sum_{k=1}^{s} S_k K_k}{\sum K_k} ,$$

with nominal variance $1/\Sigma K_k$. Variation in allele frequencies, ascertainment, or other factors may inflate

$$\chi^2_{s-1} = \Sigma K_k S_k^2 - \frac{(\Sigma K_k S_k)^2}{\Sigma K_k} .$$

Interpretation becomes difficult if s is small, unless the sources of variation are controlled. If s were very large, the SE would be

$$SE = \sqrt{\left(\frac{1}{\sum K_k}\right)\left(\frac{\chi^2_{s-1}}{s-1}\right)} .$$

In general,

$$\sqrt{\frac{\chi^2_{s-1}}{s-1}}$$

may be taken as 1 if nonsignificant and, otherwise, as t with $s-1$ df in computing a confidence interval. For example, a nominal 95% confidence limit is

$$\bar{S} \pm t\sqrt{\frac{1}{\sum K}} ,$$

where t is 12.706 for df $= 1$ and 1.960 for df $= \infty$. However, the block-and-step structure of LD makes any interval in kb approximate. Inclusion of LD maps in the HapMap database would increase its use for association mapping.

To apply this logic to a genome scan of N regions, suppose that a subset with $P < \alpha$ is selected for further study in an independent sample of greater size and with denser markers. For each of the selected regions, pooling the information from the two samples with point estimates S_1 and S_2 and information K_1 and K_2, respectively, and with the assumption of homogeneity, the FDR can be calculated as for a single stage 1 sample from N regions. Within stage 1, the principal uncertainties involve regional definition and number of replicates (r). We tentatively assume that $r = 1{,}000$ is adequate to estimate P for H_1. For greater precision with extreme H_1 probabilities, the corresponding regions may be simulated under H_0 with at least $10/P$ replicates.

Results

An Illustrative Example

The CHROMSCAN program is being extended and refined as it is applied to several association studies, the publication of which is necessarily delayed by consortium agreements. We have therefore taken as an illustrative example the U.K. case/control sample of the International Type 2 Diabetes (T2D) 1q Consortium,[18] with affection status replaced by a random SNP for each region, deleting all information about the actual disease and reducing by 1 the number of predictive SNPs in that region. Here, we analyze the 39 regions in chromosome 1q21-24 over a 21,347-kb interval on the National Center for Biotechnology Information build 35/University of California–Santa Cruz March 2004 sequence, which was used to create a genomewide LDU map from the northwestern European (CEU) sample in International HapMap phase 2, with the exclusion of SNPs with minor-allele frequencies (MAFs) <0.05 or χ^2 for the Hardy-Weinberg test >10. The data consist of 447 controls and 443 cases. Whereas this CEU sample of 60 is smaller than the U.K. control sample, the number of SNPs is vastly greater and possibly gives a more accurate LDU map. The same exclusions were made from the U.K. sample, leaving 3,547 SNPs in the reference interval, of which 296 were not identified in the HapMap sample and were therefore interpolated from kb to LDU location.[19] The average density was 1 SNP per 6 kb, and the map length was 458 LDU. One SNP was randomly drawn from each region, as defined by the CHROMSCAN default under H_0 (at least 30 SNPs and 10 LDU), which retained these regions even when replacement of affection status by a random SNP reduced the number of predictive SNPs to 29. To dichotomize affection status under H_1, we pooled the rare homozygote with the heterozygote as "affected" and the other homozygote as controls. To simulate H_0, the H_1 observations from a region were shuffled and assigned at random to generate 1,000 replicates with the same frequencies of cases and controls and the same number of SNPs remaining in the jth region. These samples are sufficient to test significance of association under H_0 and relative power of AC and AD metrics under H_1. Complexities of meta-analysis with heterogeneous data that may include single SNPs, haplotypes, probability products, linkage estimates, and coalescent assay raise problems that will take longer to solve.

Our first application of these simulated data was to examine their operating characteristic, which depends on the probability ϕ that H_1 is true for a random region in a genome scan. In the current HapMap build, there are ~65,000 LDUs (Genetic Epidemiology Group). For a minimal region of 10 LDUs and a marker density sufficiently high to give at least 30 markers per region, $\phi = .001$ corresponds to 6.5 causal markers per genome, whether significant or not. When idiomorphs (i.e., diallelic markers with small MAF) that can be identified more efficiently in families[20,21] are excluded, this ϕ is defensible for association mapping in unrelated individuals. Pooling a single H_1 sample with 1,000 shuffled replicates in each region, the value of $\phi/(1-\phi)$ is conveniently 0.001. The corresponding FDR is $1/(1 + .001\Omega/\alpha)$. Therefore, if we take

$$\alpha = \frac{.05}{N} = 7.7 \times 10^{-6} ,$$

a power of $\Omega > .15$ is sufficient to keep the FDR <0.05. This is satisfied in these simulations, but real data require either a much larger sample or meta-analysis of multiple samples.

To compare performance of the two analyses under H_1, we transformed χ^2_3 for AD to χ^2_2 with the same P value (NAG g0lfcc), comparable to AC with 2 df. Variables corresponding to ln (AC/AD) were submitted to a t test with 38 df, with similar results (table 3). AC is significantly more powerful, largely but perhaps not entirely because of a greater numerator, since the denominator is suggestively but not significantly smaller than that for AD. Similar results were obtained with the skewed difference instead of the logarithm of the ratio. On the contrary, the absolute deviation of estimated S from its true location favors AD. These apparently discrepant results may reflect the conversion from 3 to 2 df, which gives equal weight to the weakly determined parameter L and the strongly determined estimate of S. Tentatively, we conclude that AD is the more reliable test.

Stepwise regression of the mean location error for (AC + AD)/2 or AC and AD separately on variables representing noncentrality of the causal SNP within a region, kb/LDU, SNP density, blocks, steps, and MAF revealed no predictor of location error at the 5% significance level, but MAF was suggestive (P = .0524). Other studies have supported decline of power with the MAF.[22,18] On these simulated data, the CHROMSCAN logic performed well. Applications to real data and comparisons with other methods are being examined.

Discussion

Recent publications have described the high FDRs of current functional studies and weakly parametric linkage analysis, with little promise for meta-analysis.[23,24] On the contrary, two developments favor association mapping in the post-HapMap era: chips for ⩾500,000 SNPs are becoming affordable and are being applied to large samples that give a low FDR with high power. The mountains of association data that are now inevitable create an urgent need to develop and test appropriate methods of analysis and the databases required for efficient and reliable meta-analysis. On present evidence, it seems that composite likelihood, including but perhaps not limited to the Malecot model, is unique in providing both a point estimate of location S and its SE without bias by autocorrelation and, therefore, is the only demonstrated basis for valid meta-analysis. The most significant single SNP in a region with m markers provides nominal significance $P = \text{Prob}(\chi^2_1 > \max \chi^2_1)$ and regional Bonferroni correction $P_c = 1 - (1 - P)^m$. This becomes prohibitively conservative when extended to N regions, given a strong likelihood that m SNPs in a region with a causal marker do not include that marker.[25] Maximal χ^2_1 for single SNPs provides no SE for location and, therefore, no foundation for inferring the distance between a predictive marker and a causal one. Haplotypes have uncontrolled variability in block definition and length, number of markers, and haplotype grouping. Most steps in an LD map interrupt only a minority of haplotypes, which makes their definition arbitrary.[26]

Table 3. Comparison of AC and AD Models within Region under H_1

Variable Y	Mean	SE	$F_{1,38}$	P	e^y
$\ln \chi^2_{AC} - \ln \chi^2_{AD}$	.0829	.0311	7.106	.0112	1.086
$\ln V_{AC} - \ln V_{AD}$	−.0706	.0359	3.872	.0567	.932
$\ln \Lambda_{AC} - \ln \Lambda_{AD}$	.0123	.0067	3.336	.0756	1.012
$\ln \lvert \hat{S} - S \rvert_{AC} - \ln \lvert \hat{S} - S \rvert_{AD}$	.3481	.1343	6.715	.0135	1.416

NOTE.—$\ln \chi^2_{AC} - \ln \chi^2_{AD}$ was transformed to 2 df for comparability with χ^2_{AC}.

Admixture mapping is the most extreme outlier among methods for association mapping. Estimates of the frequency of a particular ancestral group, the cumulative probability of recombination with a causal marker, and the allele frequencies in the two ancestral groups are used to suggest a conserved region with a disease-related marker.[27] The Hardy-Weinberg model is assumed, although it is not appropriate if mating has been assortative by race. This approach has identified a region in 8q24 that accounts for a substantial part of the prostate cancer risk in men of European and African American origin, especially the latter.[28,29] High significance was confirmed by whole-genome admixture analysis with only 1,266 SNPs, for a 95% credible region of 3.8 Mb that includes the earlier assignment. The regional evidence is therefore overwhelming, but the confidence interval is huge compared with association studies based on older populations, much higher marker density, and composite likelihood. The Bayesian logic that provides a 95% credible interval ignores autocorrelation and, like haplotypes, does not give a point estimate or an SE.[30]

It has been amply demonstrated that linkage and LD maps are more highly correlated than either is with the kb map,[19] but they differ in several respects. Linkage maps are sex-specific and reflect interference but do so inaccurately, because only the Haldane mapping function that assumes no interference is multilocus feasible, and so a rough approximation is conventionally made to the better but still approximate Kosambi function. Population differences may exist but have not been systematically sought or found. Selective sweeps are too gradual to affect linkage maps. On the contrary, LD maps do not reflect interference, because multiple recombination in intervals small enough to retain LD takes place in different generations, are not sex-specific, and show substantial effects of selective sweeps and other differences between populations accumulated over different bottlenecks, outcrosses, and generations. A population-specific scaling factor to make a linkage map approximate an LD map must come from the latter, just as a map derived from coalescent theory is currently scaled to the linkage map.

Besides these familiar methods, many others have been suggested but seldom, if ever, used. Under H_0, the fitted parameters of the beta distribution for products of nominal probabilities depart significantly from $\alpha = \beta = 1$,[31] whereas, for our estimates of P_{ij} from composite likelihood, the fit is good. Browning[30] reanalyzed data on RFLPs. Because of extensive duplication in whole genomes, few of them have been assigned to maps created by the Human Genome Project, so neither kb nor LDU mapping was attempted. Verzelli et al.[32] combined 100,000 SNPs simulated by unspecified criteria with 32 real markers, 5 of which were within the causal sequence and were therefore not used by other investigators in a proof-of-principle exercise.[33] Not surprisingly, simulation of a known location worked well when limited to markers with an associated Bayes factor >200. No attempt was made to relate this manipulation of the data to a confidence interval and information metric for meta-analysis. All methods based on an alternative to composite likelihood have high sensitivity to number and choice of markers, which leads to the prediction that their locations and SE (if provided) are less reliable.

However much current methods for composite likelihood are refined and regardless of whether these refinements are generally accepted, current designs for genetic information networks raise serious problems.[8,14] Genome scans with 500,000 or more SNPs are envisaged from samples differing in objectives, ascertainment sources, and covariates, without anticipation of how these data would be analyzed or how the results might be incorporated into meta-analysis.[17] Permissive definition and selection of a candidate region allow caveat emptor competition among composite likelihood and single SNPs, haplotypes, linkage, functional considerations, and other alternatives in unrelated individuals or families, if they provide a point estimate of location but perhaps not an SE. Advocates of single SNPs are free to favor whatever appeals to them (e.g., tag SNPs, nonsynonymous substitutions, deletions, insertions, exons, introns, promoters, RNA polymorphisms, or common or uncommon minor alleles). However, this permissiveness does not extend to meta-analysis of independent samples, which is usually required to reach a significance level sufficient to control the FDR. Efficient meta-analysis demands reliable SEs, so methods defective in that regard cannot contribute objectively to localization of a causal polymorphism within a region. Genome scanning of LD will not realize its potential until information networks and biobanks overcome barriers to meta-analysis by adopting the best methods and eliminating inappropriate ones.

Acknowledgments

This research was supported by National Institutes of Health grant GM69414. We are grateful to M. McCarthy (2005) for permission to base simulations on anonymized data from the U.K. component of the International T2D 1q Consortium, and to J. W. McDonald and D. Siegmund for helpful advice. Values of χ^2 as a function of P were computed by nag_deviates_chi_sq (g01fcc) from the NAG, whereas the reverse inference used nag_prob_chi_sq (g01ecc).

Appendix A

Roman Symbols

d_h	Distance (in kb) between adjacent markers h and h + 1
f	Frequency of affected diplotypes in a random sample
i	Designates a replicate in the jth region ($i = 1, \ldots, r_j$)
j	Designates a region ($j = 1, \ldots, N$)
K_ψ	Information metric for parameter ψ under H_0
K_j	Estimate of information about S_j
L	Asymptote at large distance in the Malecot model with value between 0 and 1
M	Intercept at 0 distance in the Malecot model with value between 0 and 1
m_j	Number of markers in jth region
n	Number of diallelic haplotypes for a specified diallelic pair
N	Number of regions tested
p	Subscript indicating predicted value
P	Nominal significance
P_{ij}	Significance of the ith replicate in the jth region
P_c	Significance after Bonferroni or similar correction
r_j	Number of replicates in jth region
S_j	Location of a presumptive causal marker in region j
S_k	Estimated location of a possible causal marker in sample k
S_h	Location of the hth marker
s	Number of independent samples for a given region in meta-analysis
V_{ij}	Estimate of variance x_{ij}/χ^2_{ij} for the ith replicate in region j
V_j	Mean of V_{ij} near x_j under H_0
w	Penetrance in specified homozygote
x_j	Value of x for a single sample from region j without permutation
x_{ij}	Difference for the ith replicate in the jth region between two quadratic forms, the first (A) a subhypothesis of the second (C or D)

z Attributable risk defined on diplotypes for association mapping in complex inheritance

H_0 Null hypothesis of no causal association

H_1 Alternative hypothesis of causal association

Greek Symbols

α Conservative significance level = .05/N

γ Attributable risk of a two-locus haplotype under additivity

Δ Kronecker delta that gives correct sign to derivatives of the composite likelihood: $\Delta = 1$ if $S_h > S$ and -1 otherwise

ε Scaling factor for association in a kb or LDU map

ε_h LDU:kb ratio for distance between adjacent markers h and $h+1$: the interval d_h on the kb map corresponds to $\varepsilon_h d_h$ on the LD map

Λ Quadratic form in composite likelihood $= \sum K_\psi(\hat{\psi} - \psi)^2$

ρ Association probability defined on haplotypes for major loci ($\gamma = 1$) or construction of LD maps

χ $\chi^2{}_{ij}$ is the estimate of χ^2 with 2 or 3 df derived from P_{ij}

ψ Metric (usually ρ or z) for association between a pair of diallelic markers

Ω Power to reach significance under H_1

ϕ Probability that H_1 is true whether significant or not

Most Useful Formulas

A circumflex (^) indicates a local estimate of a predicted value when both appear in the same formula. A bar above a symbol indicates a weighted mean of estimates.

For maps in LDU, with no allowance for autocorrelation and map error not directly evaluated,

$$\hat{\rho} = \frac{ad - bc}{(a+b)(b+d)} \ ,$$

$$\rho = (1 - L)Me^{-\Sigma \varepsilon_h d_h} + L \ ,$$

and

$$\Lambda = \Sigma K_\rho(\hat{\rho} - \rho)^2 \ .$$

For association mapping, with allowance for autocorrelation by permuting,

$$\hat{z} = \frac{ad - bc}{(a+b)(b+d)} \text{ for SNP} \times \text{affection} \ ,$$

$$z = (1 - L)Me^{-\varepsilon\Delta(S_h - S)} + L \ ,$$

and

$$\Lambda = \Sigma K_z(\hat{z} - z)^2 \ .$$

For meta-analysis, unlike preceding applications, observed and expected values are not paired (as in $\hat{\rho}$ and ρ); instead, many estimates are associated with a single presumptive causal site:

$$\bar{S} = \frac{\sum_{h=1}^{s} S_h K_h}{\sum_{h=1}^{s} K_h} \ ,$$

$$SE \approx \sqrt{\left(\frac{1}{\sum K_h}\right)\left(\frac{\chi^2_{s-1}}{s-1}\right)} \ ,$$

and

$$K \approx \begin{matrix} \sum K_h \quad \text{if } \chi^2_{s-1}/(s-1) \text{ is nonsignificant} \\ \sum K_h/[\chi^2_{s-1}/(s-1)] \text{ else} \end{matrix} \ .$$

For FDR,

$$\mathrm{FDR} = \frac{\alpha}{\alpha + \Omega\phi/(1-\phi)} \ .$$

For error variance (see the “Control of Type I Error” section),

$$\begin{array}{ll} \text{replicates } (\mathrm{H}_0) & \mathrm{P}_{ij} \rightarrow \chi^2_{ij} \rightarrow \mathrm{V}_{ij} \\ \text{single sample } (\mathrm{H}_0 \text{ or } \mathrm{H}_1) & \mathrm{V}_{ij} \rightarrow \mathrm{V}_j \rightarrow \mathrm{P}_j \end{array} \ .$$

Web Resources

The URLs for data presented herein are as follows:

Genetic Epidemiology Group, http://cedar.genetics.soton.ac.uk/public_html/
International HapMap Project, http://www.hapmap.org/

References

1. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui L-C (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080
2. Collins A, Morton NE (1998) Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 95:1741–1745
3. Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y, Collins A (2001) The optimal measure of allelic association. Proc Natl Acad Sci USA 98:5217–5221
4. Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke K, Morton NE (2002) The first linkage disequilibrium maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA 99:2228–2233
5. Maniatis N, Morton NE, Gibson J, Xu C-F, Hosking LK, Collins A (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum Mol Genet 14:145–153
6. International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
7. Morton NE (2006) Fifty years of genetic epidemiology, with special reference to Japan. J Hum Genet 51:269–277
8. Kaiser J (2006) NIH goes after whole genome in search of disease genes. Science 311:933
9. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318
10. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100:9440–9445
11. Dalmasso C, Broet P, Moreau T (2004) A simple procedure for estimating the false discovery rate. Bioinformatics 21:660–668
12. Lindsay BG (1988) Composite likelihood methods. Contemp Math 80:221–239
13. Maniatis N, Collins A, Gibson J, Zhang W, Tapper W, Morton NE (2004) Positional cloning by linkage disequilibrium. Am J Hum Genet 74:846–855
14. Ioannidis JPA, Gwinn M, Little J, Higgins JPT, Bernstein JI, Boffetta P, Bondy M, Bray MS, Brenchley PE, Buffler PA, et al (2006) A road map for efficient and reliable human genome epidemiology. Nat Genet 38:3–5
15. Tukey JW (1962) The future of data analysis. Ann Math Stat 33:1–67
16. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
17. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
18. Zeggini E, Raynor W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated datasets. Nat Genet 37:1320–1322
19. Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE (2005) A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA 102:11835–11839
20. Smith CAB (1953) The detection of linkage in human genetics. J Roy Stat Soc B 15:153–192
21. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570
22. Zhang W, Collins A, Morton NE (2004) Does haplotype diversity predict power for association mapping of disease genes? Hum Genet 115:157–164
23. Todd JA (2006) Statistical false positive or true disease pathway? Nat Genet 38:731–733
24. Iwamato K, Kato T (2006) Gene expression profiling in schizophrenia and related mental disorders. Neuroscientist 12:349–361
25. Risch N, Merikangas K (1996) The future of genetic studies of complex human disease. Science 273:1516–1517
26. Daly MJ, Rioux JV, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229–232
27. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O’Brien SJ, Altshuler D, et al (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000
28. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, et al (2006) A common variant associated with prostate cancer in European and African populations. Nat Genet 38:652–658
29. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tan-

don A, Waliszewska A, Penney K, Steen RG, Ardie K, John EM, et al (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Nat Acad Sci USA 103:14068–14073

30. Browning SR (2006) Multilocus mapping using variable-length Markov chains. Am J Hum Genet 78:903–913

31. Dudbridge F, Koeleman BPC (2004) Efficient computation of significance levels from multiple association in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424–435

32. Verzilli CJ, Stallard N, Whittaker JC (2006) Bayesian graphical models for genomewide association studies. Am J Hum Genet 79:100–112

33. Hosking LK, Boyd PR, Xu CF, Nissum M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K, et al (2002) Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. Pharmacogenomics J 2:165–175