Published in final edited form as:

J Am Stat Assoc. 2008 March 1; 103(481): 89–100. doi:10.1198/016214507000000338..

Bayesian Gaussian Mixture Models for High-Density Genotyping Arrays

Chiara Sabatti[Associate Professor]

Departments of Human Genetics and Statistics, University of California, Los Angeles, CA 90095

Kenneth Lange[Professor and Chair]

Departments of Human Genetics, Biomathematics, and Statistics, University of California, Los Angeles, CA 90095

Abstract

Affymetrix's SNP (single-nucleotide polymorphism) genotyping chips have increased the scope and decreased the cost of gene-mapping studies. Because each SNP is queried by multiple DNA probes, the chips present interesting challenges in genotype calling. Traditional clustering methods distinguish the three genotypes of an SNP fairly well given a large enough sample of unrelated individuals or a training sample of known genotypes. This article describes our attempt to improve genotype calling by constructing Gaussian mixture models with empirically derived priors. The priors stabilize parameter estimation and borrow information collectively gathered on tens of thousands of SNPs. When data from related family members are available, our models capture the correlations in signals between relatives. With these advantages in mind, we apply the models to Affymetrix probe intensity data on 10,000 SNPs gathered on 63 genotyped individuals spread over eight pedigrees. We integrate the genotype-calling model with pedigree analysis and examine a sequence of symmetry hypotheses involving the correlated probe signals. The symmetry hypotheses raise novel mathematical issues of parameterization. Using the Bayesian information criterion, we select the best combination of symmetry assumptions. Compared to Affymetrix's software, our model leads to a reduction in no-calls with little sacrifice in overall calling accuracy.

Keywords

Empirical Bayes; Model selection; Pedigrees; Penetrance function

1. GENOTYPING ARRAYS

Technologies for high-throughput genotyping have undergone rapid development in the last decade. These technologies hold great promise for association mapping of complex disease genes, understanding of population stratification, study of chromosome-segment duplication and loss, and mutation detection.

This article focuses on genotype calling for single-nucleotide polymorphisms (SNPs). These sites of natural variation occur about every 1,000 base pairs (bp) along the human genome. Typically, only two of the four possible bases at an SNP are present in a population. If we denote the two alleles at an SNP by the letters a and b , then each person has one of three possible genotypes a/a , a/b , or b/b at the SNP, reflecting the alleles passed by the person's mother and father.

One of the most promising new methods for SNP genotyping relies on DNA chips similar to gene expression arrays (Pastinen et al. 2000). The genotyping chips produced commercially by Affymetrix (Liu et al. 2003) have recently experienced a huge surge in demand. Affymetrix's expression arrays and SNP genotyping arrays are constructed by the same photochemistry process. They also both involve DNA complementary hybridization. In this article we focus on the first- and second-generation Affymetrix 10k genotyping arrays with roughly 10,000 SNPs, providing genome coverage at an average distance between SNPs of 300,000 bp. Based on the same technology, the company has recently increased chip density to the point where 250,000 SNPs can be assayed simultaneously. Each SNP on these chips is assessed by 40 probes, each 25 bases long. Of the 40 probes, 20 are match probes that perfectly hybridize with one of the two alleles, and 20 are mismatch probes intended to measure the level of cross-hybridization. Among the 20 match probes, 10 probes are complementary to allele *a*, and 10 probes are complementary to allele *b*. Each set of 10 match probes is further subdivided into two subsets of 5 probes; one subset is complementary to the sense strand and the other subset to the antisense strand of the DNA molecule. This leads to four probe subsets: sense (*s*) *a*, antisense (*t*) *a*, sense *b*, and antisense *b*. The five probes of each subset differ in the position of the polymorphic base among the 25 bases. For some probes, the polymorphic position is central; for others, it is shifted left or right by one to three bases. Each mismatch probe is paired with one match probe and differs from it at the base in the central position of the oligonucleotide.

Genotyping proceeds by a sequence of steps: (1) The sample DNA is broken into small fragments surrounding each assayed SNP, (2) the fragments are amplified by the polymerase chain reaction (PCR), (3) fragment copies (probes) are labeled with dye molecules to distinguish the two alleles, (4) probes are hybridized to the array, and (5) the intensity of the fluorescent signal at each spot is measured. The information on an SNP genotype is contained in the relative intensities of the probes for the *a* and *b* alleles. In the presence of a homozygous *a/a* genotype, the *a* probes usually show a much stronger signal than the *b* probes. The opposite behavior occurs in the presence of a homozygous *b/b* genotype, and roughly equal intensity values occur for the heterozygous *a/b* genotype.

Together with its arrays and hybridization protocols, Affymetrix provides software for genotype calling. Details of their first calling algorithm appear in Liu et al. (2003). Further algorithm research on this and related technology has been pursued within the company, but calling protocols have only been partially documented for the public (personal communications and Di, Webster, Bartell, and Kulp 2003; Di et al. 2005). The great interest in the chips has prompted several publications comparing the utility of microsatellite and SNP genotyping in genomewide scans (Hoque, Lee, Cairns, Shoenberg, and Sidransky 2003; Middleton et al. 2004) and has inspired new statistical methods and corresponding software (Fujisawa et al. 2004; Hao, Li, Rosenow, and Wong 2004a,b; Lin et al. 2004; Leykin et al. 2005; Tebbutt et al. 2005; Rabbee and Speed 2006).

At the same time, it has become clear that Affymetrix's technology can also be profitably used for surveying copy number variation and, hence, detecting deleted or duplicated chromosomal segments. For this purpose, attention naturally shifts from the relative intensity of the allelic probes at each SNP to the absolute intensity of many probes across an entire chromosome region. If the region is deleted in one of the two homologous chromosomes carried by a person, then the average intensity of the probes within the region should be half that of other regions. The algorithms currently available to detect copy number variation capitalize on this fact, but it is largely ignored in genotype calling. It seems a shame to discard useful information.

The goal of this article is to propose a model for intensity values of allelic probes that can be used for both genotyping and detecting copy number variation. Our model differs from other models in four ways: (1) It is based on absolute rather than relative intensity values, (2) it relies only on the sample under study without reference to a training sample inaccessible to the ordinary user, (3) it makes correct and advantageous use of pedigree information whenever available, and (4) it provides soft probabilistic calls rather than hard qualitative calls for genotypes.

2. DATA PREPROCESSING AND SUMMARY STATISTICS

To motivate our model, illustrate our data preprocessing software, and explain our dimension reduction choices, we start with an exploratory analysis of the probe intensity data supplied by our UCLA colleagues Baloh, Jen, and Nelson. The data come from a linkage study of diseases of the vestibular system conducted at UCLA. In the course of their study, these investigators genotyped 63 people from 8 families using the first-generation Affymetrix 10k array. Although we do not have independent confirmation of the genotyping results, the same DNA samples were genotyped using a standard set of 400 microsatellites in a separate genome scan. The no-call rates and low prevalence of Mendelian errors in that scan convinced us that the overall DNA quality of the samples is acceptable. However, the Affymetrix genotyping experiment resulted in a high number of no-calls, motivating us to take a closer look at the data.

As described in the previous section, an Affymetrix chip queries each SNP with 40 probes. As with expression arrays, the raw data need to be transformed prior to further statistical analysis (Li and Wong 2001; Irizarry et al. 2002; Yang et al. 2002). Transformation serves to summarize the information coming from the different probes, to correct for nonspecific hybridization and to normalize intensity values across chips. We consider each purpose in turn.

To explore the relationships between the 40 probes, we considered the correlations across individuals of three different measurement groups: perfect-match (PM) probes, mismatch (MM) probes, and the differences (PM-MM) between them. It is well known that expression measurements are very skewed, and a tail-taming transformation is needed when looking at correlations. Often a log transformation is applied. In our experimental setup, the signal from a mismatch probe is expected to be comparable to the signal from a perfect-match probe in the absence of the shared allele in the sample. In this situation, we expect many probe pair differences (PM-MM) to be negative. To handle negative differences, we replaced the log transformation by the transformation $f(x) = \text{sign}(x)[\text{abs}(x)]^{1/4}$.

Preliminary examination of the data revealed that correlations vary by genotype in sampled individuals: When two individuals share an allele, the probes corresponding to the allele are registering the same signal and tend to be more correlated than in the absence of the allele, where their luminosities derive from cross-hybridization. To take into account this effect, we stratified correlations according to the genotypes of the individuals as predicted by the Affymetrix software. Figure 1 presents average correlations for the PM-MM signal over 500 SNPs selected for high heterozygosity. The probes are grouped in sense *a* (*sa*), antisense *a* (*ta*), sense *b* (*sb*), and antisense *b* (*tb*) categories. The correlation between probes for an allele is much higher when the allele is present than when it is absent. Probes within the same strand and allele group tend to be more correlated, and two signals representing the same allele but different strands or vice versa tend to be more correlated than signals agreeing in neither allele nor strand. The corresponding picture for the PM probes is qualitatively very similar. The correlation behavior of MM probes mimics the behavior observed for PM *a* probes with *b/b* genotypes and PM *b* probes with *a/a* genotypes. Given

that probes for the same allele and the same strand behave similarly, we decided to work with a four-dimensional signal summarizing the information from the four probe groups sa , ta , sb , and tb . This decision entails some loss of information, but it has clear benefits in reducing the dimensionality of the problem and is less radical than Affymetrix's decision to summarize the information from all probes in one statistic.

The role of mismatch probes is still ambiguous. Experience with gene expression arrays suggests that relying on a single MM probe can be problematic. Indeed, some of the most recent software ignores MM probes altogether. Although MM probes are certainly noisy, they do serve as useful benchmarks. With the hope of regularizing their signal, we took the MM average within a probe group before subtracting it from each of the corresponding PM probes. After this subtraction, we applied the tail-taming transformation described previously and averaged the resulting values within each probe group, giving the four-dimensional vector $y = (y_{sa}, y_{ta}, y_{sb}, y_{tb})$ just mentioned. All subsequent analysis will be based on the summary vector y . Study of PM signals without MM corrections leads to similar results.

Figure 2 illustrates these y values for a fairly typical SNP. Note that in this article each SNP has a name starting with the letter r , followed by a number describing its genomic position. The three clusters corresponding to the three genotypes are clearly identifiable. There is also an obvious positive correlation between probes regardless of genotype, reflecting overall luminosity differences between chips. This effect is often eliminated in expression studies by normalizing the signals across arrays, for example, with quantile–quantile matching (Irizarry et al. 2002). In genotyping assays, the Affymetrix genotype-calling procedure operates on the transformed variable $y'_a = y_a / (y_a + y_b)$. Fujisawa et al. (2004) instead transformed (y_a, y_b) to polar coordinates and used only the polar angle for clustering-classification purposes. It seems to us that quantile–quantile matching and ratio and angular transforms all risk losing too much information, particularly when there are other goals beside genotype calling, and it is important to keep a record of the intensity values. We have experimented with other normalization solutions, but finally opted to do without them. The correlations apparent in Figure 2 can be adequately taken into account in the modeling stage, as the next section makes evident.

Although data preprocessing is not the focus of this article, proper transformation is obviously crucial to the study of array data. We invite interested readers to consult the work of Rabbee and Speed (2006) for another perspective on preprocessing, despite their somewhat different modeling conclusions. In collaboration with Steve Erickson, we are also investigating these matters in more detail in connection with predicting copy number variation.

3. A GAUSSIAN MIXTURE MODEL

It is natural to resort to a mixture model with different distributions for each genotype and mixing proportions equal to the population frequencies of the genotypes. For modeling ease, we take the component distributions to be Gaussian. Fortunately, the data conform reasonably well to this assumption. If y is the vector of measurements for a random individual at a single SNP, then the distribution of y can be represented symbolically as

$$y \sim p^{a/a} \mathcal{N}(\mu^{a/a}, \Omega^{a/a}) + p^{a/b} \mathcal{N}(\mu^{a/b}, \Omega^{b/b}) + p^{b/b} \mathcal{N}(\mu^{b/b}, \Omega^{b/b}). \quad (1)$$

Here superscripts denote genotypes; later on subscripts will denote alleles (a and b) or probe groups (sa , ta , sb , and tb). The three Gaussian components in the distribution (1) carry

genotype-specific means $\mu^{a/a}$, $\mu^{a/b}$, and $\mu^{b/b}$ and genotypespecific variances $\Omega^{a/a}$, $\Omega^{a/b}$, and $\Omega^{b/b}$. The genotype-specific means satisfy the natural constraints

$$\mu_a^{a/a} \geq \mu_a^{a/b} \geq \mu_a^{b/b}, \quad \mu_b^{a/a} \leq \mu_b^{a/b} \leq \mu_b^{b/b} \quad (2)$$

because, for example, average values for an a probe should reflect the number of copies of the a allele present in a genotype. Under Hardy–Weinberg equilibrium, the genotype frequencies, for a noninbred person are $p^{a/a} = (p_a)^2$, $p^{a/b} = 2p_a p_b$, and $p^{b/b} = (p_b)^2$, where p_a and $p_b = 1 - p_a$ are the population frequencies of the two alleles a and b .

In the case of measurements on n unrelated people, the overall likelihood of the data factors as n separate densities of the form (1). In linkage studies, however, the measurements y are no longer independent because genotyped individuals are members of families. The resulting dependencies in effect change the admixture coefficients for each person. The correct way to handle correlated observations is to evaluate the likelihood of each family as a unit. If we let $L(y|g)$ denote the likelihood of measurement y conditional on genotype g , then the joint likelihood of the collective measurements $y_1, \dots, y_r$ from a pedigree of r related individuals is

$$\begin{aligned} L(y_1, \dots, y_r) &= \sum_{g_1, \dots, g_r} \Pr(g_1, \dots, g_r) \prod_{j=1}^r L(y_j | g_j) \\ &= \sum_{g_1} \dots \sum_{g_r} \prod_i \Pr(g_i) \prod_{k,l,m} \text{Tran}(g_m | g_k, g_l) \times \prod_j L(y_j | g_j). \end{aligned} \quad (3)$$

Here the index i ranges over all pedigree founders, the index j over all pedigree members, and the triple index (k, l, m) over all parent–offspring trios. In statistical genetics, $L(y|g)$ is called a penetrance. The prior function $\Pr(g_i)$ in the likelihood captures the Hardy–Weinberg frequency of genotype g_i . The transmission function $\text{Tran}(g_m | g_k, g_l)$ incorporates Mendel's first law of inheritance; under the commonsense rules of the law,

$$\text{Tran}(a/b | a/a, b/b) = 1, \quad \text{Tran}(a/b | a/a, a/b) = \frac{1}{2},$$

for instance. Although a great deal of effort (Lange 2002) has gone into devising fast algorithms for evaluating the pedigree likelihood (3), we will take these advances for granted. Once the software has been written, implementing different models is largely a matter of coding the penetrance function $L(y|g)$.

Our general Gaussian model, then, requires 12 mean parameters (4 for each genotype) and 30 variance parameters (10 for each genotype). These are large numbers, and if our interest were only in genotype calling, we would be inclined to summarize the data further. We have resisted this temptation because it compromises our chance to better comprehend the operating characteristics of the Affymetrix chip. Instead, we prefer to investigate certain biologically plausible symmetry hypotheses that lead to a reduction in the number of parameters.

3.1 Symmetries

Considering the nature of the experimental data y , three symmetry hypotheses are plausible. These involve genotype additivity, allelic symmetry, and strand symmetry. Under genotype additivity, we model the signal y as the sum of two independent contributions, one

corresponding to the maternal allele and the other to the paternal allele. The genotype additivity hypothesis requires that the heterozygous genotype be intermediate in the sense that

$$\mu^{a/b} = \frac{1}{2}(\mu^{a/a} + \mu^{b/b}), \quad \Omega^{a/b} = \frac{1}{2}(\Omega^{a/a} + \Omega^{b/b}). \quad (4)$$

The allelic symmetry hypothesis says that allele labels do not matter; all that matters is the presence or absence of an allele. For example, allelic symmetry dictates the constraint

$\mu_{sa}^{a/a} = \mu_{sb}^{b/b}$. Formally, we can express allelic symmetry via the permutation matrix

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad (5)$$

with the mean constraints

$$\mu^{b/b} = P\mu^{a/a}, \quad \mu^{a/b} = P\mu^{a/b}, \quad (6)$$

and the variance constraints

$$\Omega^{b/b} = P\Omega^{a/a}P, \quad \Omega^{a/b} = P\Omega^{a/b}P. \quad (7)$$

The strand symmetry hypothesis says that measurements from sense and antisense probes for the same allele have the same distribution. Formally, if Q is the permutation matrix

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad (8)$$

then we have the mean constraints

$$\begin{aligned} \mu^{a/a} &= Q\mu^{a/a}, \\ \mu^{a/b} &= Q\mu^{a/b}, \\ \mu^{b/b} &= Q\mu^{b/b}, \end{aligned} \quad (9)$$

and the variance constraints

$$\begin{aligned} \Omega^{a/a} &= Q\Omega^{a/a}Q, \\ \Omega^{a/b} &= Q\Omega^{a/b}Q, \\ \Omega^{b/b} &= Q\Omega^{b/b}Q, \end{aligned} \quad (10)$$

In our subsequent analysis, we consider all $64 = 2^3 \times 2^3$ possible combinations of these symmetry hypotheses on the mean and variance parameters. Table 1 summarizes the various combinations and the number of free parameters each entails. We also consider a variance

model (number 0) that postulates a shared variance across all three genotypes. This model has 10 variance parameters. Later we describe parameterizations of the various models that permit straightforward maximum likelihood estimation under simple parameter constraints.

3.2 Genotype Calling

Before tackling problems of statistical inference, let us clarify how the proposed model can be used for genotyping. In the mixture model, genotyping operates via Bayes rule. For a random person, the prior probability of a genotype g is simply the Hardy–Weinberg frequency $\Pr(g)$. The posterior probability $\Pr(g|y)$ of g given array data y is proportional to the product $L(y|g) \Pr(g)$ of the penetrance and the prior. We will refer to $\Pr(g|y)$ as a conditional genotype probability. One possibility is to make hard calls that choose the genotype maximizing $\Pr(g|y)$ and ignore all others. When reliable, hard calls are preferred in linkage analysis because they lighten the computational load in computing pedigree likelihoods. If computational efficiency were not an issue, then the logical procedure would be to make soft calls and incorporate marker penetrances $L(y|g)$ directly into the pedigree likelihoods.

When we are dealing with families, we can compute pedigree-based conditional probabilities rather than individual-based conditional probabilities. Given the observations $y_1, \dots, y_r$ from a pedigree, the conditional probability that person m has genotype g_m is proportional to the multiple sum

$$\Pr(g_m|y_1, \dots, y_r) \propto \sum_{g_1, \dots, g_{m-1}, g_{m+1}, \dots, g_r} \prod_i \Pr(g_i) \times \prod_{k,l,m} \text{Tran}(g_n|g_k, g_l) \prod_j L(y_j|g_j), \quad (11)$$

with m 's genotype fixed at g_m . According to Bayes rule, if we divide the right side of (11) by the likelihood $L(y_1, \dots, y_r)$, then we recover $\Pr(g_m|y_1, \dots, y_r)$. We will use both individual- and pedigree-based conditional probabilities to call genotypes. To compute either kind of conditional probability, we must substitute estimated parameters for true parameters. Thus, it is important to have high-quality estimates.

4. INFERENCE I: MAXIMUM LIKELIHOOD ESTIMATION

To estimate the unknown mean and variance parameters of our model, we will ultimately resort to an empirical Bayes procedure that estimates parameters by maximizing the posterior density. In our model, maximum a posteriori estimation is very similar to maximum likelihood estimation. For ease of exposition, therefore, we first discuss maximum likelihood estimation. This order of exposition has the added advantage of allowing us to introduce variance parameterizations that facilitate estimation and guide the choice of prior distributions. These parameterizations may be of independent interest to some readers.

To perform likelihood maximization, we use the program Search embedded in Mendel (Lange et al. 2001). Search carries out recursive quadratic programming with quasi-Newton updates to the approximate Hessian of the log-likelihood. Mendel is freely available on the UCLA Human Genetics website. In any optimization problem, it helps to choose a parameterization involving the simplest possible constraints. This desire motivates the common practice of estimating the Cholesky decomposition (Golub and van Loan 1989) of a variance matrix rather than the matrix itself. Thus, rather than having to check that a variance matrix is positive definite, one simply checks that the diagonal entries of the Cholesky decomposition are positive. Mendel accepts lower and upper bounds on parameters in addition to linear constraints. Our symmetry models induce special structures

on the variance matrix that may conflict with the Cholesky decomposition. The additivity hypothesis poses no difficulty, so we now deal with the other cases.

In the case of allelic symmetry, the relation (7) between $\Omega^{b/b}$ and $\Omega^{a/a}$ can be stated in block matrix form as

$$\Omega^{a/a} = \begin{pmatrix} C & D \\ D' & E \end{pmatrix}, \quad \Omega^{b/b} = \begin{pmatrix} E & D' \\ D & C \end{pmatrix}.$$

If we parameterize $\Omega^{a/a}$ by its Cholesky decomposition, then it is trivial to recover $\Omega^{b/b}$. The symmetry constraint on $\Omega^{a/b}$ is

$$\Omega^{a/b} = \begin{pmatrix} C & D \\ D & C \end{pmatrix} \quad (12)$$

for 2×2 symmetric matrices C and D . The evident structure in this matrix clashes with a full Cholesky decomposition, so we must consider a more delicate parameterization.

We first use the Cholesky decomposition F of the positive-definite block C of $\Omega^{a/b}$ and write

$$\begin{pmatrix} C & D \\ D & C \end{pmatrix} = \begin{pmatrix} FF' & FGF' \\ FGF' & FF' \end{pmatrix}$$

for an unknown symmetric matrix G . Because the matrix $\Omega^{a/b}$ is positive definite, we can sweep on its upper left block FF' and conclude that the resulting lower right block

$$\begin{aligned} FF' - FGF'(FF')^{-1}FGF' &= FF' - FGGF' \\ &= F(I - G^2)F' \end{aligned}$$

is positive definite (Lange 1999). This is possible only if all eigenvalues of the symmetric matrix G are strictly less than 1 in absolute value. Hence, we have reduced the problem of parameterizing an arbitrary positive-definite matrix of the form (12) to the problem of parameterizing an arbitrary symmetric matrix G with eigenvalues on the interval $(-1, 1)$.

The further transformation $H = \frac{1}{2}G + \frac{1}{2}I$ shows that it suffices to parameterize an arbitrary symmetric matrix H with eigenvalues on the interval $(0, 1)$. Such a matrix H can always be represented as the matrix exponential

$$e^K = \sum_{n=0}^{\infty} \frac{1}{n!} K^n$$

of a negative-definite matrix K . Indeed, one can easily check that e^K is symmetric when K is symmetric and that $e^\lambda \in (0, 1)$ is an eigenvalue of e^K when $\lambda < 0$ is an eigenvalue of K . Conversely, if H is positive definite with all eigenvalues in $(0, 1)$, then the convergent series

$$K = \ln H = \ln [I - (I - H)] = - \sum_{k=1}^{\infty} \frac{1}{k} (I - H)^k$$

provides the unique K with $e^K = H$. To parameterize K , we choose the Cholesky decomposition L of $-K$. In summary, we can represent

$$\begin{pmatrix} C & D \\ D & C \end{pmatrix} = \begin{pmatrix} FF^t & F(2e^{-LL^t} - I)F^t \\ F(2e^{-LL^t} - I)F^t & FF^t \end{pmatrix} \quad (13)$$

using arbitrary Cholesky decompositions F and L .

Turning now to strand symmetry, (10) compels the 2×2 blocks C , D , and E of the matrix

$$\Omega^{a/a} \begin{pmatrix} C & D^t \\ D & C \end{pmatrix}$$

to be symmetric with equal diagonal entries, in analogy to the block decomposition (12). Similar representations hold for $\Omega^{a/b}$ and $\Omega^{b/b}$. In practice, we prefer to deal with representations of the form (12). To convert between the two types of representations, we define the symmetric permutation matrix

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is then easy to check the identity $PR = RQ$ connecting R to the permutation matrices P and Q displayed in (5) and (8). If we now set $\Sigma = R\Omega^{a/a}R$, then

$$P\Sigma P = PR\Omega^{a/a}RP = RQ\Omega^{a/a}QR = R\Omega^{a/a}R = \Sigma.$$

The symmetry relation $P\Sigma P = \Sigma$ allows us to parameterize Σ as in (12). It is straightforward to recover $\Omega^{a/a} = R\Sigma R$ based on the identities $\Sigma = R\Omega^{a/a}R$ and $R^2 = I$. Similar considerations apply to $\Omega^{a/b}$ and $\Omega^{b/b}$.

If an additive model has either allelic symmetry or strand symmetry, then corresponding symmetry constraints on the heterozygote mean are automatically satisfied. For example, in the case of allelic symmetry, the identities

$$\mu^{a/b} = \frac{1}{2} (\mu^{a/a} + \mu^{b/b}), \quad \mu^{b/b} = P\mu^{a/a}, \quad P^2 = I$$

imply

$$\begin{aligned} P\mu^{a/b} &= \frac{1}{2}P(\mu^{a/a} + P\mu^{a/a}) = \frac{1}{2}(P\mu^{a/a} + P^2\mu^{a/a}) \\ &= \frac{1}{2}(P\mu^{a/a} + \mu^{a/a}) = \mu^{a/b}. \end{aligned}$$

The identities

$$\Omega^{a/b} = \frac{1}{2}(\Omega^{a/a} + \Omega^{b/b}), \quad \Omega^{b/b} = P\Omega^{a/a}P, \quad P^2 = I$$

likewise imply the variance identity $\Omega^{a/b} = P\Omega^{a/b}P$. Similar properties hold for strand symmetry.

The combination of allelic symmetry and strand symmetry is equally interesting. The obvious parameterization of means are $\mu^{a/a} = (v, v, \omega, \omega)^t$, $\mu^{a/b} = (\theta, \theta, \theta, \theta)^t$, and $\mu^{b/b} = (\omega, \omega, v, v)^t$. Variances are more complicated. It suffices to parameterize $\Omega^{a/a}$ and $\Omega^{a/b}$ and express $\Omega^{b/b}$ as $P\Omega^{a/a}P$. This assertion depends on $\Omega^{b/b}$ satisfying the symmetry relation $\Omega^{b/b} = Q\Omega^{b/b}Q$. Given that $\Omega^{a/a} = Q\Omega^{a/a}Q$, the easily checked identity $PQ = QP$ implies

$$Q\Omega^{b/b}Q = QP\Omega^{a/a}PQ = PQ\Omega^{a/a}QP = P\Omega^{a/a}P = \Omega^{b/b}.$$

We have already dealt with $\Omega^{a/a}$ subject to the symmetry constraint $\Omega^{a/a} = Q\Omega^{a/a}Q$. Turning to the parameterization of $\Omega^{a/b}$ subject to the two constraints $\Omega^{a/b} = P\Omega^{a/b}P$ and $\Omega^{a/b} = Q\Omega^{a/b}Q$, we note that our previous discussion immediately yields the representation

$$\Omega^{a/b} = \begin{pmatrix} \begin{pmatrix} u & v \\ v & u \end{pmatrix} & \begin{pmatrix} w & x \\ x & w \end{pmatrix} \\ \begin{pmatrix} w & x \\ x & w \end{pmatrix} & \begin{pmatrix} u & v \\ v & u \end{pmatrix} \end{pmatrix}$$

Fortunately, we can explicitly diagonalize a matrix of this form using the columns of the orthogonal matrix

$$O = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

as eigenvectors. In view of this fact, we represent $\Omega^{a/b}$ as

$$\Omega^{a/b} = O \begin{pmatrix} e^{\lambda_1} & 0 & 0 & 0 \\ 0 & e^{\lambda_2} & 0 & 0 \\ 0 & 0 & e^{\lambda_3} & 0 \\ 0 & 0 & 0 & e^{\lambda_4} \end{pmatrix} O^t$$

by exponentiating an arbitrary diagonal matrix with i th diagonal entry λ_i . Exponentiation takes a real eigenvalue λ_i into a positive eigenvalue e^{λ_i} . Because a symmetric matrix is

positive definite if and only if all of its eigenvalues are positive, our representation ranges over all positive-definite matrices fulfilling the symmetry constraints.

5. INFERENCE II: BAYESIAN MODELS

Maximum likelihood estimation as described in the previous section leads to satisfactory results on the 500-SNP dataset presented in Figure 2. Some of these results for a typical SNP are illustrated in Figure 4, but we defer further discussion to Section 6. Relying on maximum likelihood can lead to trouble, however. Our model assumes three mixture components, corresponding to the three different genotypes. In a limited sample, it is quite possible for only one or two of the genotypes to actually occur for a given SNP. In such cases, fitting a three-component mixture is bound to encounter difficulties. Even when all three genotypes are present, the model requires estimating a large number of parameters. Overparameterization is ameliorated but not eliminated by imposing symmetry restrictions.

Both absent genotypes and overparameterization can be handled by a good Bayesian model. If no genotypes occur for a given component, then the prior distributions on the center and dispersion of the component will entirely describe it and should have little impact on parameter estimates for other components. Adequate prior information also makes it easier to estimate a large number of parameters. Although a Bayesian model can resolve these impasses, its practical success will depend on the validity of the priors imposed. Fortunately, as so often happens with genomics problems, we are blessed with an abundance of data. The Affymetrix chips query tens of thousands of SNPs simultaneously. This collective evidence furnishes the raw material for an empirical Bayesian analysis. The key to choosing good priors is to base them on hyperparameters that can be estimated from the data collectively.

Let us now define the functional form of the priors, which is dictated by considerations of parsimony, flexibility, and computability. We assume independent priors on the allele frequency p_a , the genotype-specific mean vectors μ , and the genotype-specific variance matrices Ω . We take a beta prior on p_a with hyperparameters (α, β) . Generally speaking, we choose a Gaussian prior on each mean $\mu^{k/l}$ and an inverse Wishart (IW) or a Wishart prior on each variance matrix $\Omega^{k/l}$. The different symmetry hypotheses dictate constraints on the parameters that need to be incorporated into the priors. We will illustrate how this works in detail.

The mean parameter vector μ is partitioned into the components $\mu^{a/a}$, $\mu^{a/b}$, and $\mu^{b/b}$ pertinent to each genotype. On each of these four-dimensional vectors, we assume a Gaussian prior with mean vector $\theta^{k/l}$ and variance matrix $\Sigma^{k/l}$. Unless a symmetry constraint forces us to assume otherwise, we take these priors to be independent. In reality, the mean parameters are never fully independent because of the inequality constraints (2). These constraints are embedded in our maximization routine and, hence, implicitly in our priors. Table 2 summarizes the mean priors. Column 1 gives the mean model number defined in Table 1, column 2 displays the relevant Gaussian priors, and column 3 recalls how dependent parameters are related to free parameters. Because of symmetry constraints, some of the

means $\mu^{k/l}$ have just one or two free parameters. The notation $\mu_{[1]}^{k/l}$ in the table indicates entry 1 of the vector $\mu^{k/l}$, the only free parameter for that vector when the subscript appears.

Similarly, $\mu_{[1,3]}^{k/l}$ indicates entries 1 and 3 of $\mu^{k/l}$, the only free parameters for that vector when the subscript appears. Recall that the vector $\mu^{k/l}$ has four entries, corresponding to the sense a , antisense a , sense b , and antisense b statistics, respectively. The hyperparameters $\theta^{k/l}$ and $\Sigma^{k/l}$ apply only to the free parameters and are subscripted by [1], [1, 2], or [1, 3] as needed. On the variance matrices $\Omega^{a/a}$, $\Omega^{a/b}$, and $\Omega^{b/b}$, we impose independent inverse Wishart (IW) priors with hyperparameters $\Lambda^{k/l}$ and $\nu^{k/l}$ whenever possible (Anderson 1984).

If, under a symmetry model, one of these matrices remains unconstrained, then we continue to impose an IW prior on it. This rule allows us to define priors for model 0 (a single IW prior on the common variance matrix), model 1 (independent IW priors on each of the three variance matrices), model 2 (independent IW priors on $\Omega^{a/a}$ and $\Omega^{b/b}$), and model 4 (a single IW prior on $\Omega^{a/a}$). Under the remaining models, the three variance matrices do not collapse to a smaller set of variance matrices. For example, under allelic symmetry in model 3, $\Omega^{b/b} = P\Omega^{a/a}P$, but $\Omega^{a/b} = P\Omega^{a/b}P$ has internal symmetries that must be taken into account. In this particular case, we impose a constrained IW distribution on $\Omega^{a/b}$ with density proportional to

$$f(\Omega) \propto 1_{(\Omega=P\Omega P)} |\Omega|^{-(v_{a/b}+d+1)/2} \exp\left\{-\text{tr}\left(\Lambda^{a/b}\Omega^{-1}\right)/2\right\},$$

where $v_{a/b}$ denotes the IW degrees of freedom, $\Lambda^{a/b}$ the IW scale matrix, and $d = 4$ the dimension of the matrix argument Ω . In principle, we could approximate the normalizing constant by a Markov chain Monte Carlo (MCMC) method on the slice $\Omega = P\Omega P$ in question, but there is no need to do so in maximum likelihood and maximum a posteriori estimation. Model selection is another matter, and we will take that up later. Table 3 summarizes the constrained and unconstrained IW priors. Column 1 gives the variance model number defined in Table 1, column 2 displays the relevant independent IW priors, column 3 lists the constraints on the constrained IW densities, and column 4 recalls how dependent parameters are related to free parameters.

As an alternative to employing IW priors on the variance matrices, we now discuss imposing priors on the parameters actually used in maximization. At first glance, these priors seem less natural than priors on the variance matrices, but they avoid the problem of unknown normalizing constants and are more convenient to implement in practice. For technical reasons that will soon be clear, we discuss putting Wishart priors on the variance matrices. (If IW priors are preferred, then Wishart priors can be specified on the inverse of the variance–covariance matrices.) To begin, recall how allelic and strand symmetry lead to the consideration of the positive-definite matrices displayed in (13). Whenever such a parameterization is needed, we assume that FF^t and LL^t follow independent Wishart priors. It is known (Anderson 1984, thm. 7.2.1) that when a $d \times d$ matrix FF^t follows a Wishart distribution with scale matrix Λ and degrees of freedom v , its Cholesky decomposition $F = (f_{ij})$ has density

$$f(F) = \frac{\prod_{i=1}^d f_{ii}^{v-i} \exp\left\{-\text{tr}\left(\Lambda^{-1}FF^t\right)/2\right\}}{|\Lambda|^{v/2} 2^{d(v-2)/2} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma((v+1-i)/2)}. \quad (14)$$

Here each $f_{ii} > 0$, and in our case $d = 2$. Table 4 provides a detailed list of the various Wishart priors. In the table, the notation $F \sim \text{CW}(\Lambda, v)$ indicates that F has the density (14). To parameterize a model with strand symmetry, a variance matrix $\Omega^{k/l}$ is mapped to another matrix $\Sigma^{k/l} = R\Omega^{k/l}R$ admitting the parameterization (13). To make this transformation clear, we use $V^{k/l}$ and $W^{k/l}$ rather than $F^{k/l}$ and $L^{k/l}$ to indicate the Cholesky factorizations of $\Sigma^{k/l}$. Model 7 represents an exception discussed in detail earlier. In this case, we put independent gamma priors on the e^{λ_i} , consistent with the marginal distributions on the diagonal entries in the density (14).

Now that we have described the functional form of the priors, let us consider how the hyperparameters can be specified by borrowing information across SNPs. A traditional

empirical Bayes procedure would choose the hyperparameters that maximize the predictive distribution of the data. Unfortunately, finding the predictive distribution and evaluating it for all SNPs is very demanding. As a shortcut, we note that maximum likelihood estimation works well for SNPs with high heterozygosity. This suggested to us that we perform maximum likelihood estimation on a subset of highly polymorphic SNPs as identified by the Affymetrix software. After some experimentation, we concluded that 500 SNPs with at least five samples per genotype were enough to survey the range of common parameter values. We then treated the 500 estimated parameter vectors as a random sample from the prior and estimated the hyperparameters by maximizing the prior. Let us stress that the same procedure could be applied to a larger collection of SNPs.

6. DATA ANALYSIS I: MODEL SELECTION

We now turn to the analysis of the intensity data from our sample of 63 individuals. Our first goal is the selection of the best symmetry model among the possible $72 = 8$ (mean models) $\times$ 9 (variance models). The Bayesian information criterion (BIC) described by Schwarz (1978) fits in well with our overall approach. This criterion allows us to carry out model selection using maximum likelihood with no prior specified. In our setting, the BIC is

$$\text{BIC}(\mathcal{M}) = \max_{\mu, \Omega, p} \ln \left[L(y^1, \dots, y^{63} | \mathcal{M}, \mu, \Omega, p) \right] - \frac{\ln(63) |\mathcal{M}|}{2}. \quad (15)$$

Here L denotes a likelihood and $\mathcal{M}$ a particular model having $|\mathcal{M}|$ parameters.

Our symmetry hypotheses are all motivated by characteristics of the genotyping experiments more or less constant across SNPs. Each hypothesis involves the relative behavior of the various probes for a single SNP. Given the uniformity of chip design from SNP to SNP, we can reasonably hope that the same symmetry model applies to every SNP. If this is the case, we are in a privileged position for model selection and can aggregate information across SNPs. Because the additivity and allelic symmetry hypotheses involve comparisons of all three genotypes, it is imperative to select polymorphic SNPs. Our decision to analyze 500 SNPs with at least five representatives of each genotype addresses this issue and avoids model selection biases caused by poor fitting of mean and variance components. Evaluating the BIC for all models in these representative SNPs makes it possible to select the correct combination of symmetries and to check our assumption that a single model applies to all SNPs.

Figure 3 summarizes our analysis of the 72 models in the 500 SNPs. Although the BIC values for each model followed a fairly reproducible pattern, it was possible to cluster groups of SNPs based on their BIC vectors. Hierarchical clustering reveals six SNP clusters. Plotting the average BIC value per model within each cluster allows us to take into account possible differences between SNPs as well to reduce the variability associated with analyzing one SNP at a time. Although there is some variation across SNPs, this does not seem to make much difference in the selection of the best model. Except for a small cluster of five poorly performing cases, models 1–3 and 1–7 achieve the highest BIC values in each cluster. Moreover, overall model ordering appears to be rather stable across clusters.

By assuming one model for all SNPs, Table 5 compares average BIC values for the various models over all 500 SNPs. The strand symmetry hypothesis for mean parameters appears implausible. The same hypothesis on variance parameters is credible; evidently, the decrease in maximum likelihood is compensated by the substantial reduction in the number of parameters. The other hypotheses immediately ruled out involve additivity of the means and variances; these are models with even digits. The maximum average BIC occurs for the

combination of mean model 1 and variance model 3. This model has 12 mean parameters and, given allelic symmetry on the variance matrices, 16 variance parameters. The second best model under BIC adds strand symmetry to allelic symmetry on the variance matrices, for a total of 10 variance parameters. In the following, we will work with both models 1–3 and 1–7.

The advantages of models 1–3 and 1–7 over the unrestricted model 1–1 can be appreciated in Figure 4. The top and middle rows present the output of maximum likelihood estimation for SNP r1014 under models 1–1 and 1–7. Maximum likelihood places one observation in the *a/b* cluster under model 1–1 and in the *a/a* cluster under model 1–7. Affymetrix assigns the same observation to the *a/a* cluster. The probable explanation of the difference appears in the (y_{sa}, y_{sb}) plot of the top row of the figure. Here the projected observation is closer to the *a/b* cluster. The erroneous *a/b* call is presumably a consequence of the higher variability in parameter estimates with a large unrestricted model such as 1–1.

7. DATA ANALYSIS II: GENOTYPE CALLING

Confident that models 1–3 and 1–7 describe the data well, we used the method of moments to estimate the hyperparameters of the priors from the maximum likelihood estimates. Armed with the hyperparameters, we then performed maximum a posteriori estimation on the entire 10,000-SNP dataset. Comparison of the rows of Figure 4 shows that the differences between maximum likelihood and maximum a posteriori estimates are modest for the typical SNP r1014. This holds for both Wishart and inverse Wishart priors on the variance matrices under models 1–3 and 1–7. Figure 5 depicts SNP r9659 under model 1–7 with a Wishart prior on the variance. Because only one genotype is present in the data, here the advantage of a hierarchical Bayes method is obvious. All possible combinations of models 1–3 and 1–7 and Wishart and inverse Wishart priors perform similarly. Unless stated otherwise in the rest of the article, we report results for model 1–7 with a Wishart prior.

We now turn to a systematic comparison of the differences in genotype calls between Affymetrix's procedure and ours, which is based on the posterior probabilities of the three genotypes as discussed in Section 3.2. Although the scatterplots of Figure 4 suggest that it is straightforward to assign a sample to a component, calling every single genotype may not be wise. Affymetrix decides whether or not to call a sample based on quality-control indicators defined on all 40 raw measurements gathered on an SNP. Although quality control is surely important, it is not the primary focus of this article. If we want reliable calls within the context of our model, two options are open to us: (1) Restrict calling to samples already called by Affymetrix, thereby implicitly adopting their quality-control procedures; (2) impose a threshold on the computed maximum posterior probability below which we refuse to call a sample. Given the importance of reliable calls, we set the threshold at 99%. With this threshold, our no-call rate is 11% for individual calls and 7% for pedigree calls. Not surprisingly, pedigree information helps resolve some uncertain cases. As evidence that our posterior probabilities are able to capture at least some of Affymetrix's quality-control information, if we consider only those samples not called by Affymetrix, then our no-call rates rises to 25% for individual calls and to 15% for pedigree calls.

In the intersection of the Affymetrix-called genotypes and our individually called genotypes, the discordance rate is 1.7%. This is close to the nominal error rate of 1% with a posterior probability threshold of 99%. If we use pedigree calls, the discordance rate increases to 2%. In trying to understand the discrepancies between Affymetrix's calls and our calls, we are not simply comparing two algorithms. Affymetrix takes advantage of training data unavailable to us. For this reason, it is tempting to argue that all disagreements should be settled in Affymetrix's favor. However, this conclusion ignores the advantages of a good

model that captures information across many SNPs. Although we do not have true genotypes on any of the samples studied, we do have re-genotyping data. These data suggest that our discrepancy rate with Affymetrix calls is very close to Affymetrix's effective genotyping error rate in the current pedigrees.

Because seven samples from one of the pedigrees in the study gave a particularly high Affymetrix no-call rate of 30%, our colleagues decided to re-genotype the samples. The re-genotyping was carried out with the second-generation Affymetrix 10k chip and constituted an entirely new experiment. Unfortunately, it is impossible for us to analyze the raw measurements of the second experiment because it was discarded by the technician carrying out the genotyping. In any case, Affymetrix's no-call rate for this new experiment was only 4%. These better experimental results suggest that the second genotype is the correct one. Therefore, it is interesting to compare the agreement of Affymetrix's calls in the two experiments with the agreement of our calls on the first experiment with their calls on the second experiment. A priori, one would expect higher agreement between the two Affymetrix calls if the Affymetrix algorithm is more reliable.

There were about 9,000 SNPs common to the two platforms available for comparison purposes. Affymetrix's discordance rate for the called genotypes common to both experiments was 1.3%. The discordance rate between our calls on the first experiment and Affymetrix's calls on the second experiment was 1.7%, precisely the same as the discordance rate between our calls on the first experiment and Affymetrix's calls on the first experiment. This observation together with the small difference $.4\% = 1.7\% - 1.3\%$ in discordance rates leads us to believe the two methods enjoy similar accuracy and that incorrect calls are largely a function of the limited resolution of the experimental data.

Another way of comparing the methods is to look at the no-calls in the initial Affymetrix analysis. Our methods lead to 25% no-calls in these cases. The genotypes of the remaining 75% of the cases agree 85% of the time with the Affymetrix genotypes from the second experiment. On one hand, this result suggests that Affymetrix's conservative practice of basing calls on quality-control measures is prudent. On the other hand, it shows that there is considerable room for improvement. Lacking a training sample and faced with presumably bad data, we succeed in correctly calling $64\% = 75\% \times 85\%$ of Affymetrix's no-calls.

8. DISCUSSION

Compared to Affymetrix's methods, our model-based methods give equivalent accuracy and a higher rate of genotype calling. Other advantages of our methods include: (1) no requirement for a separate training sample, (2) propagation of information from a large set of SNPs to each individual SNP, (3) ready adaption to specific experiments, (4) effective use of pedigree information penalizing Mendelian inconsistencies, and (5) availability of penetrances if soft genotype calls are sought. The explicit handling of signal intensity is another selling point of model-based methods. Genotyping chips enable more sensitive assays for variation in copy numbers and have the potential to revolutionize cytogenetics and fetal screening for chromosome abnormalities.

Further improvements in genotype calling are certainly possible. For instance, the presence of linkage disequilibrium makes information from neighboring SNPs very relevant to genotype calling. The recent article of Ayers et al. (2007) takes some tentative steps in this direction. Exploitation of linkage disequilibrium does not require pedigree data.

Our models were initially motivated by the need to analyze data obtained on the Affymetrix 10k arrays. They can, however, be applied to the next generation of arrays with much higher

density. We are currently studying microdeletions with 250k arrays in collaboration with Stan Nelson at UCLA.

The computational cost of implementing our models is linear in the number of SNPs. Except for large pedigrees with inbreeding loops, it is also linear in study sample size. Larger study samples obviously increase the precision of parameter estimates. Many features of our models generalize to other high-throughput genotyping technologies based on quantitative measurements of allele abundance. The Illumina genotyping platform is a case in point.

Finally, including symmetry relations in our models has forced us to revisit the question of how best to parameterize variance matrices with block symmetry structure. This is a generic problem, and we hope that our representations will find other statistical applications.

Acknowledgments

We thank the editor, an associate editor, and two referees for their careful suggestions. We thank professors Robert Baloh, Joanna Jen, and Stan Nelson (all of UCLA) for letting us use the data collected in their genetic study. Hane Lee carried out the genotyping, and Hui Wang helped with data formatting. Chiara Sabatti was supported in part by NSF grant DMS-02-39427, NIH/NIDOD grant DC04224, ASA/Ames grant NCC2-1364, and USPHS grant GM53275. Kenneth Lange was supported in part by USPHS grants GM53275 and MH59490.

REFERENCES

- Anderson, T. An Introduction to Multivariate Statistical Analysis. Wiley; New York: 1984.
- Ayers K, Sabatti C, Lange K. A Dictionary Model for Haplo-typing, Genotype Calling, and Association Mapping. *Genetic Epidemiology*. 2007; 31:672–683. [PubMed: 17487885]
- Di X, Webster T, Bartell D, Kulp D. A Dynamic Model-Based Genotyping Algorithm for WGA. *Affymetrix invention disclosure*. 2003
- Di X, et al. Dynamic Model Based Algorithms for Screening and Genotyping Over 100 K SNPs on Oligonucleotide Microarrays. *Bioinformatics*. 2005; 21:1958–1963. [PubMed: 15657097]
- Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*. 2001; 96:1151–1160.
- Fujisawa H, Eguchi S, Ushijima M, Miyata S, Miki Y, Muto T, Matsuura M. Genotyping of Single Nucleotide Polymorphism Using Model-Based Clustering. *Bioinformatics*. 2004; 20:718–726. [PubMed: 14751987]
- Golub, GH.; van Loan, CF. *Matrix Computations*. 3rd ed.. The Johns Hopkins University Press; Baltimore: 1996.
- Hao K, Li C, Rosenow C, Wong W. Estimation of Genotype Error Rate Using Samples With Pedigree Information...An Application on GeneChip Mapping 10K Array. *Genomics*. 2004a; 84:623–630. [PubMed: 15475239]
- Hao K, Li C, Rosenow C, Wong W. Detect and Adjust for Population Stratification in Population-Based Association Study Using Genomic Control Markers: An Application of Affymetrix Genechip Human Mapping 10k Array. *European Journal of Human Genetics*. 2004b; 12:1001–1006. [PubMed: 15367915]
- Hoque M, Lee C, Cairns P, Schoenberg M, Sidransky D. Genome-Wide Genetic Characterization of Bladder Cancer: A Comparison of High-Density Single Nucleotide Polymorphism Arrays and PCR-Based Microsatellite Analysis. *Cancer Research*. 2003; 63:2216–2222. [PubMed: 12727842]
- Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*. 2002; 4:249–264. [PubMed: 12925520]
- Lange, K. *Numerical Analysis for Statisticians*. Springer-Verlag; New York: 1999.
- Lange, K. *Mathematical and Statistical Methods for Genetic Analysis*. 2nd ed.. Springer-Verlag; New York: 2002.

- Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E. Mendel Version 4.0: A Complete Package for the Exact Genetic Analysis of Discrete Traits in Pedigree and Population Data Sets. *American Journal of Human Genetics*. 2001; 69:A1886.
- Li C, Wong W. Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. *Proceeding of the National Academy Science*. 2001; 98:31–36.
- Leykin I, Hao K, Cheng J, Meyer N, Pollak M, Smith R, Wong W, Rosenow C, Li C. Comparative Linkage Analysis and Visualization of High-Density Oligonucleotide SNP Array Data. *BMC Genetics*. 2005; 6:7. [PubMed: 15713228]
- Lin M, Wei L, Sellers W, Lieberfarb M, Wong W, Li C. dChipSNP: Significance Curve and Clustering of SNP-Array-Based Loss-of-Heterozygosity Data. *Bioinformatics*. 2004; 20:1233–1240. [PubMed: 14871870]
- Liu W, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder T, Webster T, Dong S, Liu G, Jones K, Kennedy G, Kulp D. Algorithms for Large-Scale Genotyping Microarrays. *Bioinformatics*. 2003; 19:2397–2403. [PubMed: 14668223]
- Middleton F, Pato M, Gentile K, Morley C, Zhao X, Eisener A, Brown A, Petryshen T, Kirby A, Medeiros H, Carvalho H, Macedo A, Dourado A, Coelho I, Valente J, Soares M, Ferreira C, Lei M, Azevedo M, Kennedy J, Daly M, Sklar P, Pato. Genomewide Linkage Analysis of Bipolar Disorder by Use of a High-Density Single-Nucleotide-Polymorphism (SNP) Genotyping Assay: A Comparison With Microsatellite Marker Assays and Finding of Significant Linkage to Chromosome 6q22. *American Journal of Human Genetics*. 2004; 74:886–897. [PubMed: 15060841]
- Pastinen T, Raitio M, Lindroos K, Tainola P, Peltonen L, Syvänen A. A System for Specific, High-Throughput Genotyping by Allele-Specific Primer Extension on Microarrays. *Genome Research*. 2000; 10:1031–1042. [PubMed: 10899152]
- Rabbee N, Speed T. A Genotype Calling Algorithm for Affymetrix SNP Arrays. *Bioinformatics*. 2006; 22:7–12. [PubMed: 16267090]
- Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978; 6:461–464.
- Tebbutt S, Opushnyev I, Tripp B, Kassamali A, Alexander W, Anderson M. SNP Chart: An Integrated Platform for Visualization and Interpretation of Microarray Genotyping Data. *Bioinformatics*. 2005; 21:124–127. [PubMed: 15308543]
- Yang YH, Dudoit S, Luu P, Lin D, Peng V, Ngai J, Speed T. Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation. *Nucleic Acids Research*. 2002; 30:e15. [PubMed: 11842121]

Figure 1.

Gray scale image of the average correlations between SNP signals. The map involves 20 signals per SNP and averages over 500 SNPs. From left to right, the genotypes are a/a , a/b , and b/b . Lighter colors correspond to higher correlations. The signals occur in the following order: five probes for a allele, sense; five probes for a allele, antisense; five probes for b allele, sense; and five probes for b allele, antisense.

Figure 2.

Scatter plot of the four statistics for SNP r1014. From left to right, we plot the average of y_{sa} and y_{ta} versus the average of y_{sb} and y_{tb} , y_{sa} versus y_{sb} , and y_{ta} versus y_{tb} . The clusters corresponding to the three genotypes can be easily identified. Samples that are called by the Affymetrix software are indicated with filled symbols: a triangle for a/a , squares for a/b , and circle for b/b . The substantial positive correlation between a allele and b allele probes is obvious.

Figure 3.

BIC values for 72 models over 500 SNPs. Models are indicated above the graph by a two-digit number, the first digit identifying the mean model and the second digit the variance model. The y coordinate of each curve is the mean BIC value of the corresponding model for all SNPs in one of the six identified clusters. The thickness of a curve is proportional to the number of SNPs in the cluster. The two vertical lines identify models 1–3 and 1–7.

Figure 4.

Parameter estimates and genotype calls for SNP r1014. The top row depicts model 1–1, corresponding to unrestricted mean and variance parameters; the center row depicts model 1–7, corresponding to allelic and strand symmetry in the variance parameters; and the bottom row depicts model 1–7 in the Bayesian framework, with Wishart priors on the variances. The plots from left to right show the average of y_{sa} and y_{ta} versus the average of y_{sb} and y_{tb} , y_{sa} versus y_{sb} , and y_{ta} versus y_{tb} . Samples called by the Affymetrix software are indicated with filled triangles for a/a , filled diamonds for a/b , and filled circles for b/b . Samples called by our procedure are indicated by corresponding unfilled shapes. An ellipse surrounds each estimated mean, with scale determined by the estimated variance matrix.

Figure 5. Parameter estimates and genotype calls for SNP r9659. This depiction of the average of y_{sa} and y_{ta} versus the average of y_{sb} and y_{tb} relies on the plotting conventions of Figure 4.

Table 1

Models defined by combinations of symmetry hypotheses

Model characteristics	Model number							
	0	1	2	3	4	5	6	7 8
Additivity	—		*		*		*	*
Allelelic symmetry	—			*	*			*
Strand symmetry	—					*	*	*
Mean parameters	—	12	8	6	4	6	4	3 2
Variance parameters	10	30	20	16	10	18	12	10 6

Table 2

Priors for the mean parameters

Model	Priors for free mean parameters	Constraints on mean parameters
1	$\mu^{k/l} \sim N(\theta^{k/l}, \Sigma^{k/l}), k, l = \{a, b\}$	None
2	$\mu^{k/k} \sim N(\theta^{k/k}, \Sigma^{k/k}), k = \{a, b\}$	$\mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$
3	$\mu^{a/a} \sim N(\theta^{a/a}, \Sigma^{a/a}), \mu_{[1,2]}^{a/b} \sim N(\theta_{[1,2]}^{a/b}, \Sigma_{[1,2]}^{a/b})$	$\mu^{b/b} = P\mu^{a/a}, \mu_{[3,4]}^{a/b} = \mu_{[1,2]}^{a/b}$
4	$\mu^{a/a} \sim N(\theta^{a/a}, \Sigma^{a/a})$	$\mu^{b/b} = P\mu^{a/a}, \mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$
5	$\mu_{[1,3]}^{k/l} \sim N(\theta_{[1,3]}^{k/l}, \Sigma_{[1,3]}^{k/l}), k, l = \{a, b\}$	$\mu_{[2,4]}^{k/l} = \mu_{[1,3]}^{k/l}$
6	$\mu_{[1,3]}^{k/k} \sim N(\theta_{[1,3]}^{k/k}, \Sigma_{[1,3]}^{k/k}), k = \{a, b\}$	$\mu_{[2,4]}^{k/k} = \mu_{[1,3]}^{k/k}, \mu^{a/b} = (\mu^{a/a} + \mu^{b/b}) / 2$
7	$\mu_{[1]}^{k/l} \sim N(\theta_{[1]}^{k/l}, \Sigma_{[1]}^{k/l}), k, l = \{a, b\}$	$\mu_{[3]}^{k/k} = \mu_{[1]}^{l/l}, \mu_{[3]}^{a/b} = \mu_{[1]}^{a/b}, \mu_{[2,4]}^{k/l} = \mu_{[1,3]}^{k/l}$
8	$\mu_{[1]}^{k/k} \sim N(\theta_{[1]}^{k/k}, \Sigma_{[1]}^{k/k}), k = \{a, b\}$	$\mu_{[3]}^{k/k} = \mu_{[1]}^{l/l}, \mu_{[2,4]}^{k/l} = \mu_{[1,3]}^{k/k}, \mu^{a/b} = (\mu^{a/b} + \mu^{b/b}) / 2$

NOTE: The entire prior density on the mean parameters is specified by taking the product of the prior densities in the second column. Column 1 identifies the model, and column 3 recalls how the remaining parameters are determined by the free parameters.

Table 3

Constrained IW priors for the variance parameters

Inverse Wishart priors for variance parameters			
Model	Unconditional distributions	Conditioning events	Dependent matrices
0	$\Omega^{a/a} \sim \text{IW}(\Lambda^{a/a}, \mathbf{v}^{a/a})$	None	$\Omega^{b/b} = \Omega^{a/b} = \Omega^{a/a}$
1	$\Omega^{k/l} \sim \text{IW}(\Lambda^{k/l}, \mathbf{v}^{k/l}), k, l = \{a, b\}$	None	None
2	$\Omega^{k/k} \sim \text{IW}(\Lambda^{k/k}, \mathbf{v}^{k/k}), k = \{a, b\}$	None	$\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$
3	$\Omega^{a/a} \sim \text{IW}(\Lambda^{a/a}, \mathbf{v}^{a/a})$ $\Omega^{a/b} \sim \text{IW}(\Lambda^{a/b}, \mathbf{v}^{a/b})$	$\Omega^{a/b} = P\Omega^{a/b}P$	$\Omega^{b/b} = P\Omega^{a/a}P$
4	$\Omega^{a/a} \sim \text{IW}(\Lambda^{a/a}, \mathbf{v}^{a/a})$	None	$\Omega^{b/b} = P\Omega^{a/a}P$ $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$
5	$\Omega^{k/l} \sim \text{IW}(\Lambda^{k/l}, \mathbf{v}^{k/l}), k, l = \{a, b\}$	$\Omega^{a/a} = Q\Omega^{a/a}Q$ $\Omega^{a/b} = Q\Omega^{a/b}Q, \Omega^{b/b} = Q\Omega^{b/b}Q$	None
6	$\Omega^{k/k} \sim \text{IW}(\Lambda^{k/k}, \mathbf{v}^{k/k}), k = \{a, b\}$	$\Omega^{a/a} = Q\Omega^{a/a}Q, \Omega^{b/b} = Q\Omega^{b/b}Q$	$\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$
7	$\Omega^{a/a} \sim \text{IW}(\Lambda^{a/a}, \mathbf{v}^{a/a})$ $\Omega^{a/b} \sim \text{IW}(\Lambda^{a/b}, \mathbf{v}^{a/b})$	$\Omega^{a/a} = Q\Omega^{a/a}Q,$ $\Omega^{a/b}P, \Omega^{a/b} = Q\Omega^{a/b}Q$	$\Omega^{b/b} = P\Omega^{a/a}P$
8	$\Omega^{a/a} \sim \text{IW}(\Lambda^{a/a}, \mathbf{v}^{a/a})$	$\Omega^{a/a} = Q\Omega^{a/a}Q$	$\Omega^{b/b} = P\Omega^{a/a}P$ $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$

NOTE: Column 1 identifies the model. Column 2 lists the free variance matrices and the unconditional IW priors assumed for them; different free variance matrices $\Omega^{k/l}$ and $\Omega^{i/i}$ are independent. Column 3 lists the constraints on which the prior distribution has to be conditioned. Column 4 recalls how the remaining parameters are determined by the free parameters.

Table 4

Wishart prior distributions on the free variance parameters

Model	Wishart priors for free variance parameters
0	$\Omega^{a/a} \sim W(\Lambda^{a/a}, \nu^{a/a})$
1	$\Omega^{k/l} \sim W(\Lambda^{k/l}, \nu^{k/l}), k, l = \{a, b\}$
2	$\Omega^{k/k} \sim W(\Lambda^{k/k}, \nu^{k/k}), k = \{a, b\}$
3	$\Omega^{a/a} \sim W(\Lambda^{a/a}, \nu^{a/a}), F \sim CW(\Lambda_F, \nu_F), L \sim CW(\Lambda_L, \nu_L)$
4	$\Omega^{a/a} \sim W(\Lambda^{a/a}, \nu^{a/a})$
5	$V^{k/l} \sim CW(\Lambda_V^{k/l}, \nu_V^{k/l}), W^{k/l} \sim (\Lambda_W^{k/l}, \nu_W^{k/l}), k, l = \{a, b\}$
6	$V^{k/k} \sim CW(\Lambda_V^{k/k}, \nu_V^{k/k}), W^{k/k} \sim (\Lambda_W^{k/k}, \nu_W^{k/k}), k = \{a, b\}$
7	$V^{a/a} \sim CW(\Lambda_V^{a/a}, \nu_V^{a/a}), W^{a/a} \sim CW(\Lambda_W^{a/a}, \nu_W^{a/a}), e_i^\lambda \sim \text{Gamma}(a_i, b_i) \text{ independent}$
8	$V^{a/a} \sim CW(\Lambda_V^{a/a}, \nu_V^{a/a}), W^{a/a} \sim CW(\Lambda_W^{a/a}, \nu_W^{a/a})$

NOTE: Column 1 identifies the model. Column 2 specifies the independent priors. See the text and Table 3 for a description of how the entire set of parameters can be reconstructed from the free ones.

Table 5

BIC values for the different model combinations

Mean model	Variance model								
	0	1	2	3	4	5	6	7	8
1	-118.77	-108.82	-117.03	-100.31	-112.31	-107.47	-117.6	-103.88	-115.09
2	-128.13	-115.51	-124.8	-109.71	-122.48	-115.15	-126.93	-112.88	-125.1
3	-137.42	-123.91	-131.85	-118.49	-129.19	-124.57	-134.66	-121.83	-131
4	-142.82	-126.92	-136.13	-123.49	-134.23	-129.13	-139.17	-128.08	-137.79
5	-160.26	-146.65	-158.12	-144.91	-158.33	-146.34	-160	-147.22	-161.27
6	-166.18	-148.3	-162.67	-147.75	-164.59	-151.69	-165.03	-152.96	-167.07
7	-163.59	-148.85	-159.35	-150.99	-162.44	-153.78	-164.6	-155.3	-166.35
8	-168.54	-150.16	-163.88	-153.21	-167.75	-157.85	-169.04	-160.77	-171.48