**Title**
Bayesian Gaussian Mixture Models for High Density Genotyping Arrays



**Authors**
Sabatti, Chiara
Lange, Kenneth

**Publication Date**
2005-04-01

# Bayesian Gaussian Mixture Models for High Density Genotyping Arrays

Chiara Sabatti[1] and Kenneth Lange[1,2]

[1] Departments of Human Genetics and Statistics, UCLA, Los Angeles CA 90095-7088,

[2] Department of Biomathematics, UCLA, Los Angeles CA 90095-1766,

UCLA Statistics Department Preprint # 421

April 2005

**Running head** Gaussian Models for Affymetrix Chips

**Keywords** Penetrance function; pedigrees; empirical Bayes; model selection.

**Corresponding author** Chiara Sabatti

Department of Human Genetics

UCLA School of Medicine

695 Charles E. Young Drive South

Los Angeles, California 90095-7088 (USA)

FAX: (310) 794-5446

Phone: (310) 794-9567

e-mail: csabatti@mednet.ucla.edu

**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 a SNP fairly well given a large enough sample of unrelated individuals or a training sample of known genotypes. The present paper describes our attempt to improve genotype calling by constructing Gaussian penetrance 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, Gaussian penetrance 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 BIC criterion, we select the best combination of symmetry assumptions. Compared to the genotype calling results obtained from Affymetrix's software, we are able to reduce the number of no-calls substantially and quantify the level of confidence in all calls. Once pedigree analysis software can accommodate soft penetrances, we can expect to see more reliable association and linkage studies with less wasted genotyping data.

# 1 Introduction

Technologies for high throughput genotyping have undergone rapid development in the last half 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. One of the most promising new methods for SNP (single nucleotide polymor-

phism) 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 both also involve DNA complementary hybridization. The great interest in these chips is demonstrated by the number of publications comparing the success of genomewide scans conducted with microsatellites and with Affymetrix chips (Hoqhe *et al.*, 2003; Middleton *et al.*, 2004) and by the number of statistical methods and software the chips have inspired (Hao *et al.*, 2004a; Hao *et al.*, 2004b; Lin *et al.*, 2004; Tebbut *et al.*, 2005; Leykin *et al.*, 2005). 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 has been pursued within the company, but it has not yet been publicly documented (personal communications and Di *et al.*, 2003). Other researchers have also proposed calling procedures for high throughput SNP genotyping applicable to the Affymetrix chip (Fujisawa *et al.*, 2004). Given the nature of the data, it is hardly surprising that all of the calling algorithms proposed involve clustering. Within this broad setting there are many differences in computational detail and training data.

The goal in the current paper is to present a genotype calling procedure based on Gaussian mixture models. Our procedure is model based and adapts well to genotyping data gathered on human pedigrees. The dependencies occurring between relatives are both a curse and a blessing. On the one hand, they complicate statistical analysis and the identification of genotype clusters. On the other hand, the genetic constraints of Mendelian inheritance inform and correct the calling process. In other words, pedigree structure is a form of prior information capable of guiding genotype calls. Another form of prior information for a given SNP is the evidence from other SNPs. The Affymetrix chips query tens of thousands of SNPs simultaneously. This collective evidence furnishes the raw material for our empirical Bayesian analysis. Imposition of empirical priors stabilizes maximum likelihood estimation and compensates for reduced information at less poly-

morphic SNPs. Finally, model based analysis lends itself to probabilistic assignment of genotypes. In practice, most calling procedures refrain from making the hardest calls near cluster boundaries. The resulting missing data contribute nothing to linkage and association studies. Soft calls based on posterior probabilities partially salvage this lost information.

In the long term, we hope to use the Gaussian mixture model to investigate the operating characteristics of the Affymetrix chips. Our interest extends beyond simple genotyping. These chips may well prove helpful in inferring loss and duplication of chromosome segments. Excessive homozygosity is one indicator of segment loss; low hybridization levels are another. With segment duplication the opposite patterns prevail. Unfortunately, the use of quantitative intensity values is problematic unless one has a reference model for these values. Our study of chip operating characteristics leads to such models. Finally, the statistical inferences possible with a good model will surely find applications in the further development of the chips and the choice of appropriate DNA probe sets.

The remainder of this paper is organized as follows. In section 2 we present a general model for on-line clustering and genotype calling based on recorded hybridization signals. This model is potentially applicable to a variety of genotyping schemes and is not restricted to the Affymetrix chips. In section 3 we give a brief description of the technology underlying the Affymetrix 10k array and present the results of an exploratory data analysis that motivate our model. We also present a sequence of symmetry hypotheses that are biologically plausible and define interesting parameter constraints. In section 4 we illustrate how the different symmetry models can be parameterized; in section 5 we outline how to assign prior probabilities and how to obtain maximum a-posteriori estimates. Section 6 describes the model selection procedure for individual SNPs and the results of our analysis on a set of 63 genotyped individuals spred over eight pedigrees. We conclude with a brief discussion of the virtues and limitations of Gaussian mixture models.

# 2 Empirical Bayes Gaussian Models

In high throughput SNP genotyping, we seek to identify genotype clusters on the basis of quantitative measures of allelic abundance. Consider a biallelic marker with alleles $a$ and $b$. For each individual, we observe some measure $y_a$ of the presence of $a$ in his/her tissue sample. We likewise observe $y_b$ for allele $b$. More generally, $y_a$ and $y_b$ can be multivariate vectors. In our analysis of the Affymetrix data, they are, in fact, multivariate, but for simplicity we now consider them as univariate. Individuals with genotype $a/a$ are expected to have high values of $y_a$ and low values of $y_b$. The opposite is true for individuals with genotype $b/b$. Individuals with genotype $a/b$ are intermediate. If sufficiently many individuals are typed, then three clusters should emerge from the data and lead to proper assignment of $y$ values to genotypes. Unfortunately, what constitutes a high or low value of $y_a$ or $y_b$ is unknown a-priori. Due to differences in experimental conditions, there is variability in average signal intensity across chips. This variability manifests itself as a spurious correlation between $y_a$ and $y_b$ values and tends to mask genotypes. Thus, raw measurements need to be transformed before clustering. These transformations are technology specific, so we leave them unspecified for the moment and discuss them later when we analyze the Affymetrix data.

Given observations on multiple individuals, a genotype calling procedure involves two steps: 1) identification of clusters, and 2) assignment of genotyped individuals to clusters. In supervised clustering, step 1) is carried out separately from step 2) using individuals of known genotype. This is the case with the genotype calling software provided by Affymetrix (Liu *et al.*, 2003). In step 2) new genotypes are assigned to the nearest predefined cluster. Supervised clustering has the advantage that genotype calls are computationally quick and anchored by known genotypes. However, it has the disadvantage that it does not adapt to variation in laboratory procedures. Unsupervised clustering performs both steps 1) and 2) on the same data without resorting to a training sample (Fujisawa *et al.*, 2004). It is the reasonable choice for an adaptive procedure since few laboratories can bear the cost of sequencing sample individuals to ascertain their genotypes unambiguously.

Unsupervised clustering is also more scientifically open. With canned software the rationale for no calls is hidden from view, and attempts to improve the calling procedure are frustrated.

However, unsupervised clustering presents its own challenges. Some genotypes are absent in small samples simply by chance. This confuses rigid clustering procedures. Genotyping related family members diminishes the uncertainty of many genotypes, but the allelic abundance measures are now correlated across individuals. Proponents of unsupervised clustering prefer parametric methods based on mixtures of Gaussians (Fujisawa *et al.*, 2004) since coupling these models with complexity penalties performs well in selecting the number of clusters. Balanced against this advantage, classical clustering methods are somewhat more robust to departures from independence. Because our clustering models respect pedigree structure, they overcome the sensitivity of the Gaussian models to dependent observations. Furthermore, Gaussian models permit us to incorporate empirical priors that borrow information across SNPs. This helps in estimating the parameters for all three clusters even when sample heterozygosity is low.

The Gaussian mixture model relies on the notion of penetrance, which is the likelihood of a phenotype such as $y_a$ or $y_b$ given an underlying genotype. Let $y$ be the vector of measurements for one individual and one SNP. Marginally, $y$ comes from a mixture of multivariate Gaussians,

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

In other words, conditional on genotype $g$, the measurements $y$ follow a Gaussian distribution with genotype specific means $\mu^{a/a}$, $\mu^{a/b}$, and $\mu^{b/b}$, and genotype specific variances $\Omega^{a/a}$, $\Omega^{a/b}$, and $\Omega^{b/b}$. The genotype-specific means satisfy the natural constraints

$$\mu_a^{a/a} \;\geq\; \mu_a^{a/b} \;\geq\; \mu_a^{b/b}\,, \qquad \mu_b^{a/a} \;\leq\; \mu_b^{a/b} \;\leq\; \mu_b^{b/b}. \tag{2}$$

Enforcing these inequalities stabilizes parameter estimation. When $y_a$ and $y_b$ are multivariate, the inequality constraints hold component by component. Under Hardy-Weinberg equilibrium, the genotype frequencies for a noninbred person satisfy $p^{a/a} = (p^a)^2$, $p^{a/b} = 2p^ap^b$, 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 $b$. For a pedigree of $r$ related individuals, the collective measurements $y_1, \ldots, y_r$ have joint likelihood

$$\begin{aligned} L(y_1, \ldots, y_r) &= \sum_{g_1} \cdots \sum_{g_r} \Pr(g_1, \ldots g_r) \prod_j \Pr(y_j|g_j) \\ &= \sum_{g_1} \cdots \sum_{g_r} \prod_i \text{Prior}(g_i) \prod_{k,l,m} \text{Tran}(g_m|g_k, g_l) \prod_j \text{Pen}(y_j|g_j). \end{aligned} \tag{3}$$

Here the index $i$ ranges over all pedigree founders, and the triple of indices $(k, l, m)$ ranges over all parent-offspring trios. The prior function in the likelihood incorporates the Hardy-Weinberg frequencies, and the transmission function incorporates the usual Mendel transmission probabilities of 0, $\frac{1}{2}$, and 1. The penetrance function is Gaussian in the current setting (Lange, 2002).

Our choice of the functional form of the Bayesian priors is dictated by considerations of parsimony, flexibility, and computability. We assume independent priors on the allele frequency $p^a$, the genotype-specific mean vectors $\mu$, and the genotype specific variance matrices $\Omega$. We take a beta prior on $p^a$ with hyperparameters $(\alpha, \beta)$, a Gaussian prior on each mean $\mu^{l/k}$ with hyperparameters $\theta^{l/k}$ and $\Sigma^{l/k}$, and an inverse Wishart prior on each variance matrix $\Omega^{l/k}$ with hyperparameters $\Lambda^{k/l}$ and $\nu^{k/l}$. To select the hyperparameters of the prior distributions, we proceed in two steps. We first estimate all ordinary parameters for a large subset of SNPs with high heterozygosities. Estimation is done via our pedigree analysis program Mendel, which has embedded in it a quasi-Newton optimization engine. Mendel is freely available on the UCLA Human Genetics website (Lange *et al.*, 2001). Treating the maximum likelihood estimates as realizations from the priors, we then estimate the hyperparameters by the method of moments as described in the Appendix.

Once the hyperparameters have been estimated, we re-estimate the ordinary parameters for each SNP, using the empirically derived priors to steer the maximization. With these maximum a-posteriori estimates in hand, we are ready to evaluate the penetrances and posterior probabilities of the three possible genotypes for each person $i$ and each SNP. Penetrances are easy to compute because they are the genotype-specific likelihoods as determined by the model. We distinguish two kinds of posterior genotype probabilities, depending on the availability and use of pedigree infor-

mation. A *pedigree* based posterior probability summarizes the information gathered on the entire pedigree containing person $i$. To compute such a probability, one must evaluate three separate likelihoods, fixing a different genotype for $i$ in each case. The sum of these restricted likelihoods is the pedigree likelihood (3), which serves as the normalizing constant in Bayes rule. If $i$ is an isolated individual, or if we decide to ignore pedigree information, then the posterior probability of a SNP genotype is proportional to the product of the population genotype frequency and the penetrance for that genotype; the likelihood (1) is the normalizing constant. We refer to these posterior probabilities as *individual* based. In our view, calling software should report penetrances and both posterior probabilities. Penetrances are pertinent when a SNP plays a part in a more complicated gene mapping computation. Posterior probabilities help in assessing the reliability of genotyping.

# 3 Affymetrix 10K Arrays

The first and second-generation Affymetrix 10k arrays genotype roughly 10,000 SNPs, providing genome coverage with an average distance between SNPs of 0.3 centiMorgans (cM). The company has recently increased chip density to the point where 100,000 SNPs can be assayed simultaneously. On these chips each SNP 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 polymorphic base. Given this technical set

up, it is natural to assume that the signals are more homogeneous within probe types than across probe types. Hence, it is reasonable to consider $y$ a four-dimensional vector by averaging within probe type. This decision has the obvious advantage of reducing the dimensionality of our models.

The data discussed in this paper come from a linkage study of diseases of the vestibular system conducted by our UCLA colleagues Baloh, Jen, and Nelson. In the course of their investigations, they genotyped 63 people from eight 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 convince us that the overall DNA quality of the samples is acceptable. It is well known that intensity data from hybridization arrays needs to be preprocessed in order to be meaningfully analyzed (Li and Wong, 2001; Yang *et al.*, 2002; Irizarry *et al.*, 2002). Initial exploration of the data suggested to us that a background correction and a transformation to normality would be adequate for further analysis purposes. In particular, we subtracted from each perfect match value the corresponding mismatch value and then subjected these background corrected intensities to the tail taming transformation $f(x) = \text{sign}(x)[\text{abs}(x)]^{\frac{1}{4}}$.

To gain a feel for the hybridization signals, we made genotype assignments for 500 randomly selected SNPs using the Affymetrix software. This allowed us to compute genotype specific means and correlations for the 20 different match intensities. Figure 1 displays the average correlations among the 20 measurements for each of the three genotypes. Inspection of the diagonal blocks of the figure makes it clear that the signals within a four-probe subset are highly correlated when the probes query a single allele present in the genotype. Two signals representing the same allele but different strands or vice versa also tend to be more highly correlated than signals agreeing in neither allele nor strand. Based on these observations, we concluded that intensities could be modeled in a 4-dimensional space rather than in a 20-dimensional space. We considered using both the median and the mean to summarize the five measurements within each probe subset and finally opted for the mean.

Figure 2 illustrates the mean values for a fairly typical SNP. Note that, in the present paper, each SNP has a name starting with the letter "r" and 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. In gene expression experiments, quantile-quantile matching is used to correct for luminosity differences. Global normalization can also be achieved by subtracting the average log intensity from the log intensity for each SNP measurement (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 transform $(y_a, y_b)$ to polar coordinates and use 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. Subtracting the overall chip-intensity from each SNP measurement is preferable, but subtracting mismatch intensities partially achieves the same end in a SNP-specific way. The remaining correlations are adequately modeled by a judicious choice of the covariance matrices $\Omega^{k/l}$. Thus, we have chosen the conservative course of averaging the original signals corrected for mismatches and transformed according to the fourth root function mentioned earlier. This decision has the advantage of retaining the scale of the SNP measurements in loss of heterozygosity studies.

In view of these exploratory conclusions, we decided to summarize the intensities of the Affymetrix 10k array by four statistics that we label $y_{sa}$, $y_{ta}$, $y_{sb}$, and $y_{tb}$. Each is the average of five measurements from one allele and one strand, corrected for background luminosity and transformed as previously described. The average of the 5 mismatch signals for a probe group forms its background. Our general Gaussian model for these four statistics 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 compre-

hend 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. These hypotheses involve genotype additivity, allelic symmetry, and strand symmetry.

The genotype additivity hypothesis requires that the heterozygous genotype to be intermediate in the sense that

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

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

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

as the mean constraints

$$\mu^{b/b} = P\mu^{a/a}, \qquad \mu^{a/b} = P\mu^{a/b}, \tag{5}$$

and variance constraints

$$\Omega^{b/b} = P\Omega^{a/a}P, \qquad \Omega^{a/b} = P\Omega^{a/b}P. \tag{6}$$

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

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

then we have mean constraints

$$\mu^{a/a} = Q\mu^{a/a}, \qquad \mu^{a/b} = Q\mu^{a/b}, \qquad \mu^{b/b} = Q\mu^{b/b}, \tag{7}$$

and the variance constraints

$$\Omega^{a/a} = Q\Omega^{a/a}Q, \qquad \Omega^{a/b} = Q\Omega^{a/b}Q, \qquad \Omega^{b/b} = Q\Omega^{b/b}Q. \tag{8}$$

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. In the next section we describe parameterizations of the various models that permit straightforward maximum likelihood estimation under simple parameter constraints. We also propose modifications of the priors consistent with the symmetry hypotheses and show how to perform maximum a posteriori estimation. In later sections we discuss our strategy for model selection and summarize our model fitting conclusions.

# 4 Parameterization and Constraint Simplification

To carry out maximum likelihood or maximum a posteriori estimation, it helps to choose parameters with 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. With some of the structured variance matrices mandated by our symmetry models, naive application of the Cholesky maneuver fails. We now discuss how to salvage the situation.

The additivity hypothesis poses no difficulty. In the case of allelic symmetry, the relation between $\Omega^{b/b}$ and $\Omega^{a/a}$ can be stated in block matrix form as

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

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

$$\Omega^{a/b} = \begin{pmatrix} C & D \\ D & C \end{pmatrix} \tag{9}$$

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

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

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

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

$$FF^t - FGF^t(FF^t)^{-1}FGF^t = FF^t - FGGF^t = F(I - G^2)F^t$$

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 (9) 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

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

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

$$\begin{pmatrix} C & D \\ D & C \end{pmatrix} = \begin{pmatrix} FF^t & F(2e^{-LL^t} - I)F^t \\ F(2e^{-LL^t} - I)F^t & FF^t \end{pmatrix} \tag{10}$$

using arbitrary Cholesky decompositions $F$ and $L$.

Turning now to strand symmetry and letting

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

we note that the $2 \times 2$ blocks $C$, $D$, and $E$ are all symmetric matrices of the form

$$\begin{pmatrix} u & v \\ v & u \end{pmatrix}$$

in analogy to the block decomposition (9). Similar representations hold for $\Omega^{a/b}$ and $\Omega^{b/b}$. In practice, we prefer to deal with representations of the form (9). To convert between the two types of representations, we define the symmetric permutation matrix

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

and verify that $PR = RQ$. If we now set $\Sigma = R\Omega^{a/a}R$, then

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

This symmetry relation allows us to parameterize $\Sigma$ as in equation (9). 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

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

imply

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

The identities

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

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 for means is equivalent to the parameterization

$$\mu^{a/a} = \begin{pmatrix} \nu \\ \nu \\ \omega \\ \omega \end{pmatrix}, \quad \mu^{a/b} = \begin{pmatrix} \theta \\ \theta \\ \theta \\ \theta \end{pmatrix}, \quad \mu^{b/b} = \begin{pmatrix} \omega \\ \omega \\ \nu \\ \nu \end{pmatrix}.$$

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

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

$$
\begin{aligned}
&= PQ\Omega^{a/a}QP \\
&= P\Omega^{a/a}P \\
&= \Omega^{b/b}.
\end{aligned}
$$

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 assert that

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

This follows directly from our previous discussion. 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 $\lambda_i$. Exponentiation takes a real eigenvalue $\lambda_i$ into a positive eigenvalue $e^{\lambda_i}$. Since 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 Priors

Following the outline in Section 2, we now specify prior distributions on the parameters. Our choices respect the symmetry constraints and exploit independent priors whenever reasonable. We have already assumed independent priors across mean and variance parameters.

The mean parameter vector $\mu$ 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 constaint 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 one gives the mean model number defined in Table 1, column two displays the relevant Gaussian priors, and column three recalls how dependent parameters are related to free parameters. Owing to the symmetry constraints, some of the means $\mu^{k/l}$ have just one or two free parameters. The notation $\mu^{k/l}_{[1]}$ in the table indicates entry 1 of the vector $\mu^{k/l}$, the only free parameter for that vector when the subscript appears. Similarly, $\mu^{k/l}_{[1,3]}$ indicate 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, 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.

We impose on the variance matrices $\Omega^{a/a}$, $\Omega^{a/b}$, and $\Omega^{b/b}$ independent inverse Wishart (IW) priors 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 particluar case, we impose a constrained IW distribution on $\Omega^{a/b}$ with density proportional to

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

where $\nu_{a/b}$ denotes the IW degrees of freedom, $\Lambda^{a/b}$ the IW scale matrix, and $d = 4$ the dimension of the matrix argument $\Omega$. In principle, we could approximate the normalizing constant by an 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 one gives the variance model number defined in Table 1, column two displays the relevant independent IW priors, column three lists the constraints on the constrained IW densities, and column four 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 equation (10). Whenever such a parameterization is needed, we assume that $FF^t$ and $LL^t$ follow independent Wishart priors. It is known (Anderson, theorem 7.2.1, 1984) that when a $d \times d$ matrix $FF^t$ follows a Wishart distribution with scale matrix $\Lambda$ and degrees of freedom $\nu$, its Cholesky decomposition $F = (f_{ij})$ has density

$$f(F) \;=\; \frac{\prod_{i=1}^d f_{ii}^{\nu-i}\exp\{-\text{tr}(\Lambda^{-1}FF^t)/2\}}{|\Lambda|^{\nu/2}2^{d(\nu-2)/2}\pi^{d(d-1)/4}\prod_{i=1}^d\Gamma(\frac{\nu+1-i}{2})}\,. \tag{11}$$

Here each each $f_{ii} > 0$, and in our case $d = 2$. Table 4 provides a detailed list of the various Wishart priors. 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 (10). 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 $e^{\lambda_i}$, consistent with the marginal distributions on the diagonal entries in the density (11).

# 6 Analysis of the Affymetrix Data

We now turn to the analysis of the intensity data from our sample of 63 individuals. Our initial goal is to select the best of the $72 = 8 \times 9$ possible symmetry models. Our strategy is motivated by two considerations. On the one hand, we prefer that the same model should hold for all SNPs. After all, each SNP is assayed with the same technology and presumably involves the same symmetries. On the other hand, we are committed to a Bayesian analysis with model selection using the Bayesian information criterion (BIC) described by Schwarz (1978). This allows us to carry out model selection using maximum likelihood with no prior specified. Given the computation demands of pedigree analysis, the number of models, and the sheer number of SNPs, we limit model selection to 500 randomly chosen SNPs having at least five Affymetrix identified representatives for each genotype. This minimum representation rule avoids model selection biases caused by poor fitting of mean and variance components, a problem exacerbated by low sample heterozygosity.

Figure 3 graphically illustrates the BIC

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

for each of the 72 models. Here $L$ denotes a likelihood and $\mathcal{M}$ a particular model having $|\mathcal{M}|$ parameters. A glance at the right half of the figure (mean models 5-8) shows the implausibility of

the strand symmetry hypothesis for mean parameters. 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 hypothesis immediately ruled out is additivity for the variances (even-numbered variance models). The maximum average BIC occurs for the combination of mean model 1 and variance model 3. This model has 12 free mean parameters and, owing to the hypothesis of allelic symmetry on the variance matrices, 16 variance parameters. The second best BIC adds strand symmetry to allelic symmetry on the variance matrices, for a total of 10 variance parameters. The mean inequality constraints (2) were enforced during maximization for all 72 models. When we repeat the same analysis using globally normalized luminosity values, we come to the same qualitative conclusions.

We also checked the adequacy of the "one model for all SNPs" hypothesis. This check is achieved by recording the BICs for each SNP and clustering SNPs on the basis of theirl BIC vectors over the 72 models. We use the agglomerative hierachical clustering procedure `hclust` implemented in R, with distances between SNPs defined by correlation. The results are shown in Figure 4, where each SNP corresponds to a row in the image, and each model to a column. Here the SNPs are ordered so that the branches of the invisible cluster tree do not cross. Models are ordered by their average BIC values across all SNPs. While it is possible to identify some subgroups in the data, these do not lend much support to models other than 1-3 or 1-7, in corroboration of our "one model for all SNPs" hypothesis.

Comparison of our genotype calls with those of Affymetrix provide another helpful check in model selection. In making comparisons, we do not mean to imply that the Affymetrix procedure is the gold standard. But with a published error rate of 0.1% in ideal circumstances (Hao *et al.*, 2004), it does offer a benchmark for improvement and discrimination of models. One natural statistic for comparison is the average number of discordant calls model by model across all 500 SNPs. This statistic, as presented in Figure 5, favors the combination of mean model 1 and variance model 7 (allelic and stand symmetry). Adopting variance model 3 (allelic symmetry alone) increases the

discordance rate from 1.5% to 2%. Let us stress that these comparisons involve posterior genotype probabilities based on individuals and evaluated at the maximum likelihood estimates.

A more detailed comparison of the clusters defined by the methods is also revealing. The data in Figure 6 on SNP r1014 recapitulates Figure 2, but now with a focus on models 1-1, 1-3, and 1-7. We have already remarked on the well differentiated clusters. What is surprising is the number of no calls by Affymetrix. In some cases, SNP intensities fall very near the center of a cluster and still are not called by Affymetrix. In cases where chip luminosity values are exceptionally high and signals occur far from cluster centers, the reluctance of the Affymetrix software to make a call is more understandable. Certainly a no call is preferable to a bad call. In our genotype calls, we make no attempt to screen outliers. All genotypes are called, and depending on the model, may be called differently. In practice, outliers should be eliminated at the quality control stage. Furthermore, we do not actually advocate using called genotypes or posterior genotype probabilities in pedigree analysis. Genotype-by-genotype penetrances are more pertinent because genotype frequencies are incorporated in pedigree computations when founders are visited.

Secure in the knowledge that models 1-3 and 1-7 describe the data well, we estimated the hyperparameters of the priors from the maximum likelihood estimates using the method of moments estimators described in the Appendix. Armed with the hyperparameters, we then performed maximum a posteriori estimation on the entire SNP data set. Comparison of Figures 7 and 6 shows that the differences between maximum likelihood and maximum a posteriori estimates are not substantial for the typical SNP r1014. This holds for both Wishart and inverse Wishart priors on the variance matrices under model 1-7. Figure 8 depicts SNP r1015 under model 1-3. Here the presence of three distinct clusters appears dubious, and the use of a hierarchical Bayes method substantially improves results. All possible combinations of model 1-3 and 1-7 and Wishart and inverse Wishart priors perform similarly. In the sequel, unless otherwise stated, we report results for model 1-7 with a Wishart prior.

We now turn to a more systematic comparison of the differences in genotype calls. Figure 9

gives a breakdown by SNP and person of Affymetrix's no calls in our data. Some SNPs and some individuals clearly account for a disproportionate fraction of the 13.6% of no calls. Presumably this reflects the low efficiency of some probes and the poor DNA quality of some samples. It is interesting to see how the uncertainty perceived by Affymetrix is reflected in our procedure. Since our posterior genotype probabilities $\pi_{a/a}$, $\pi_{a/b}$, and $\pi_{b/b}$ are soft and distributed over all three genotypes, it is useful to quantify their spread. One measure is the Gini purity index (Bhargave and Uppuluri, 1977)

$$\pi_{\text{purity}} \quad = \quad \pi_{a/a}^2 + \pi_{a/b}^2 + \pi_{b/b}^2,$$

which ranges from a maximum of 1 to a minimum of $\frac{1}{3}$. Figure 10 shows that the purity of Affymetrix's no calls is lower than overall purity. However, the purity of some no calls is quite high. There are three reasons for this: (1) our model is able to make definitive calls on many Affymetrix no calls, (2) the purity index is inflated because uncertainty is almost always spread over at most two neighboring genotypes and hence has minimum of $\frac{1}{2}$ rather than $\frac{1}{3}$, and (3) we do not exclude outliers, which may be far from all genotype centers but are still confidently assigned. A hard genotype call should take into account not only the most probable genotype but also the posterior probability attached to that genotype. In particular, it is wise to refrain from calling a genotype when the posterior probability of error is above a certain threshold. If we elect not to call genotypes whose maximum posterior probability falls below 99%, then our no-call rate is 11% for individual calls and 7% for pedigree calls. Not surprisingly, pedigree information makes it possible to resolve some uncertain cases. These values rise to 25% and 15%, respectively, among Affymetrix no calls. Figure 11 depicts these results as well as our most probable genotype assignments for Affymetrix's no calls. The no calls show a noticeable increase in the frequency of heterozygotes.

In the intersection of the Affymetrix called genotypes and our confidently called genotypes (posterior probability $> 0.99$) the discordance rate is 1.8%. Notice that this is close to our nominal

error rate of 1%. A detailed breakdown appears in the top half of Table 5. If we use the information in an entire pedigree in making a call, then this difference increases to 2%. These calls are referred to as pedigree calls in the table. The bottom half of the table reports the same comparisons for all of the Affymetric called genotypes. Not surprisingly, the vast majority of disagreements occur as a heterozygote-homozygote difference.

Seven samples from one of the pedigrees in the study resulted in a particularly high Affymetrix no-call rate (30%). Our collegues decided to re-genotype these samples. The re-genotyping was carried out using the second generation Affymetrix 10K chip, which for purposes of comparison shares about 900 SNPs with the first generation chip. The Affymetrix no-call rate for this new experiment was much lower (4%), suggesting better experimental results. Affymetrix's own discordance rate for the called genotypes common to both experiments was 1.3%. The discordance rate between our own calls on the first experiments with Affymetrix calls on the second experiment is 5%. This higher discordance rate is explained by our willingness to call more genotypes. Indeed, our discordance rate rises from 1.8% among the original Affymetrix called genotypes to 15% among the original no calls. This experience suggests that we screen calls that have low penetrances for all genotypes and eliminate the corresponding outliers that conform poorly to the overall model.

# 7 Discussion

The methods described here extract more information from Affymetrix SNP arrays than the company software does. Although the Affymetrix calling algorithm is accurate with high quality DNA (Hao *et al.*, 2004a), accuracy degrades and the fraction of no calls increases in less ideal circumstances. Figure 9 emphasizes this point. In our experience, good modeling and close inspection of each data set mitigate the worse declines in performance.

Our analysis is based on the assumption of normality. While normality appears to be a reason-

able approximation for suitably transformed intensity values, it clearly breaks down in the presence of extreme outliers. One remedy is to eliminate outliers prior to genetic analysis using quality control techniques. Such techniques are currently under investigation. Another remedy is to replace the multivariate normal distribution in analysis with the multivariate t distribution (Lange *et al.,* 1989). Despite the fact that we have ignored outliers in the current data, our genotype calls agree well with the Affymetrix calls. In addition we are able to reduce the number of no calls significantly.

One of our interests in modeling probe measurements is their potential in loss of heterozygosity studies. Qualitatively oriented genotype calling confuses low intensities with SNP homozygosity. The symmetry relations we have introduced are biologically plausible and effectively reduce the dimension of parameter space. Investigating these relations has forced us to revisit the question of how best to parameterize variance matrices with block structure. We hope that our representations will find other statistical applications.

Our algorithms output penetrance probabilities as well as genotype calls. To our knowledge, no linkage analysis software is equipped to handle penetrances. But this situation is about to change as geneticists see the virtues of soft calls versus no calls. Current software for association mapping does handle quantitative data in case-control studies; it does not in association studies with pedigrees. We tend to be philosophic because hardware almost always leads software. No field of science or technology can be judged mature until both function at the highest level. We certainly take this lag as a spur to bring more sophisticated statistical methods and software to bear on genetic epidemiology.

# Acknowledgments

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 DMS0239427, NIH/NIDOCD grant DC04224, ASA/Ames grant NCC2-1364, and USPHS grant GM53275. Kenneth Lange was supported in part by USPHS grants GM53275 and MH59490.

# Appendix: Estimation of Hyperparameters

To estimate the hyperparameters of the beta prior for the SNP population frequency $p^a$, we use the method of moments. It is well known that the mean and variance of the beta density are

$$\begin{aligned} m &= \frac{\alpha}{\alpha+\beta} \\ v &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

These equations can be solved for $\alpha$ and $\beta$ in the form

$$\begin{aligned} \alpha &= \frac{m(m-m^2-v)}{v} \\ \beta &= \frac{(1-m)(m-m^2-v)}{v}. \end{aligned}$$

If we compute the sample mean and variance

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} \hat{p^a}_i, \qquad \hat{v} = \frac{1}{n}\sum_{i=1}^{g} n(\hat{p^a}_i - \hat{m})^2,$$

then substituting $\hat{m}$ for $m$ and $\hat{v}$ for $v$ yields the method of moment estimators $\hat{\alpha}$ and $\hat{\beta}$ of $\alpha$ and $\beta$. If a reliable estimate of $p^a$ is available from more extensive data on a particular SNP, it probably is better to fix the value of $p^a$ at that estimate during maximum a posterior estimation of the penetrance parameters and omit the corresponding beta prior on the SNP.

To estimate the hyperparameters $\theta^{k/l}$ and $\Sigma^{k/l}$ of the Gaussian prior $\mathcal{N}(\theta^{k/l}, \Sigma^{k/l})$ on $\mu^{k/l}$, we use the sample mean and the sample covariance matrix of the maximum likelihood estimates $\hat{\mu}_i^{k/l}$, $i = 1, \ldots, n$. We also use maximum likelihood estimates to select the hyperparameters of the

unconstrained IW prior on each $d \times d$ variance matrix $\Omega^{k/l}$. It is well known that the IW density

$$f_{\Lambda,k}(\Omega) \;=\; \frac{|\Lambda|^{\frac{\nu}{2}}|\Omega|^{-\frac{\nu+d+1}{2}}\exp\{-\text{tr}(\Lambda\Omega^{-1})/2\}}{2^{\nu d/2}\pi^{\frac{d(d-1)}{4}}\prod_{i=1}^{d}\Gamma(\frac{\nu+1-i}{2})}$$

has mean matrix $\mathrm{E}(\Omega) = (\nu - d - 1)^{-1}\Lambda$ for $\nu > d + 1$. Hence, the method of moments estimator of $\Lambda$ is

$$\hat{\Lambda} \;=\; \frac{\nu - d - 1}{n}\sum_{i=1}^{n}\hat{\Omega}_i. \tag{12}$$

This estimate correctly centers $\hat{\Lambda}$ but leaves estimation of $\nu$ unresolved. To avoid an overly strong prior, we prefer setting $\nu$ to its smallest admissible integer value. Of course, there are alternatives to this naive estimate. One is to substitute the method of moments estimate (12) into the likelihood for the sample $\hat{\Omega}_1, \ldots, \hat{\Omega}_n$ and maximize the resulting function of $\nu$. Alternatively, one can exploit the identity

$$\mathrm{E}(\Omega^{-1}) \;=\; \nu\Lambda^{-1} \;=\; \frac{\nu}{\nu - d - 1}\,\mathrm{E}(\Omega)^{-1}$$

and write

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}_i^{-1} \;\approx\; \frac{n\nu}{\nu - d - 1}\left(\sum_{i=1}^{n}\hat{\Omega}_i\right)^{-1}.$$

We then choose the value of $\nu$ that minimizes the absolute difference between the traces of these two matrices or the sum of squared differences of their entries. These two criteria yielded virtually identical results in the cases we explored.

To estimate the parameters of the $CW$ distributions from the sample estimates $\hat{F}_1, \ldots, \hat{F}_n$, we generate the corresponding products $\hat{F}_1\hat{F}_1^t, \ldots, \hat{F}_n\hat{F}_n^t$, treat these as independent Wishart observations, and extract the method of moments estimates as just described.

# References

Anderson, T. (1984) *An Introduction to Multivariate Statistical Analysis*, Wiley.

Bhargava, T., and Uppuluri,V. , (1977) "An axiomatic derivation of the Gini's index of diversity with applications," *Metron*, 33, 41–53.

Di, X., Webster, T., Bartell, D., and Kulp, D. (2003) "A dynamic model-based genotyping algorithm for WGA", *Affymetrix invention disclosure.*

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001) "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto, T., and Matsuura, M. (2004) "Genotyping of single nucleotide polymorphism using model-based clustering," *Bioinformatics*, 20, 718–726.

Hao, K., Li, C., Rosenow, C., Wong, W. (2004a) "Estimation of genotype error rate using samples with pedigree information – an application on GeneChip Mapping 10K array," *Genomics*, 84, 623–630.

Hao, K., Li, C., Rosenow, C., Wong, W. (2004b) "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*, 12, 1001–1006.

Hoque, M., Lee, C., Cairns, P., Schoenberg, M., and Sidransky, D. (2003) "Genome-wide genetic characterization of bladder cancer: a comparison of high-density single nucleotide polymorphism arrays and PCR-based microsatellite analysis," *Cancer Research*, 63, 2216–2222.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T. (2002) "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, 4, 249–264.

Lange, K. (1999) *Numerical Analysis for Statisticians.* Springer-Verlag, New York.

Lange, K. *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J., Sobel, E. (2001) "Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets," *Amer. J. Hum. Genetics,* 69 (supplement), A1886.

Lange K., Little, R.J.A., Taylor, J.M.G. (1989) "Robust statistical modeling using the t distribution." *JASA*, 84, 881–896.

Li, C. and Wong, W. (2001) "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proc. Natl. Acad. Sci.*, 98, 31–36.

Leykin, I., Hao, K., Cheng, J., meyer, N., Pollak, M., Smith, R., Wong, W., Rosenow, C., Li, C. (2005) "Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data," *BMC Genetics* in press.

Lin, M., Wei, L., Sellers, W., Lieberfarb, M., Wong, W., and Li, C. (2004) "dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data," *Bioinformatics*, 20, 1233–1240

Liu, W., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T., Webster, T., Dong, S., Liu, G., Jones, K., Kennedy, G., and Kulp, D. (2003) "Algorithms for large-scale genotyping microarrays," *Bioinformatics*, 19, 2397–2403.

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., and Pato, C. (2004) "Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphis (SNP) genotyping assay: a comparison with microsatellite marker

assays and finding of significant linkage to chromosome 6q22," *Am. J. Hum. Genet.*, 74, 886–897.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., Tsui, K. (2001) "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data," *Journal of Computational Biology*, 8, 37–52.

Pastinen, T., Raitio, M., Lindroos, K., Tainola, P., Peltonen, L., and Syvänen, A. (2000) "A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays," *Genome Res.*, 10, 1031–1042.

Schwarz, G. (1978) "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Tebbutt, S., Opushnyev, I., Tripp, B., Kassamali, A., Alexander, W., and Anderson, M. (2005) "SNP chart: an integrated platform for visualization and interpretation of microarray genotyping data," *Bioinformatics*, 21, 124–127.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., Speed, T. (2002) "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research,* 30, No. 4 e15.

| Model Number | Additive Symmetry | Allelic Symmetry | Strand Symmetry | # Mean Parameters | # Variance Parameters |
|---|---|---|---|---|---|
| 1 | no | no | no | 12 | 30 |
| 2 | yes | no | no | 8 | 20 |
| 3 | no | yes | no | 6 | 16 |
| 4 | yes | yes | no | 4 | 10 |
| 5 | no | no | yes | 6 | 18 |
| 6 | yes | no | yes | 4 | 12 |
| 7 | no | yes | yes | 3 | 10 |
| 8 | yes | yes | yes | 2 | 6 |

Table 1: **Models Defined by Combinations of Symmetry Hypotheses**

| M | Priors for Free Mean Parameters | Constraints on Mean Parameters |
|---|---|---|
| 1 | $\mu^{k/l} \sim \mathcal{N}(\theta^{k/l}, \Sigma^{k/l}) \quad k, l = \{a, b\}$ | None |
| 2 | $\mu^{k/k} \sim \mathcal{N}(\theta^{k/k}, \Sigma^{k/k}) \quad k = \{a, b\};$ | $\mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$ |
| 3 | $\mu^{a/a} \sim \mathcal{N}(\theta^{a/a}, \Sigma^{a/a}) \;\; \mu^{a/b}_{[1,2]} \sim \mathcal{N}(\theta^{a/b}_{[1,2]}, \Sigma^{a/b}_{[1,2]})$ | $\mu^{b/b} = P\mu^{a/a} \quad \mu^{a/b}_{[3,4]} = \mu^{a/b}_{[1,2]}$ |
| 4 | $\mu^{a/a} \sim \mathcal{N}(\theta^{a/a}, \Sigma^{a/a})$ | $\mu^{b/b} = P\mu^{a/a} \;\; \mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$ |
| 5 | $\mu^{k/l}_{[1,3]} \sim \mathcal{N}(\theta^{k/l}_{[1,3]}, \Sigma^{k/l}_{[1,3]}) \quad k, l = \{a, b\}$ | $\mu^{k/l}_{[2,4]} = \mu^{k/l}_{[1,3]}$ |
| 6 | $\mu^{k/k}_{[1,3]} \sim \mathcal{N}(\theta^{k/k}_{[1,3]}, \Sigma^{k/k}_{[1,3]}) \;\; k = \{a, b\}$ | $\mu^{k/k}_{[2,4]} = \mu^{k/k}_{[1,3]} \quad \mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$ |
| 7 | $\mu^{k/l}_{[1]} \sim \mathcal{N}(\theta^{k/l}_{[1]}, \Sigma^{k/l}_{[1]}) \quad k, l = \{a, b\}$ | $\mu^{k/k}_{[3]} = \mu^{l/l}_{[1]} \;\; \mu^{a/b}_{[3]} = \mu^{a/b}_{[1]} \;\; \mu^{k/l}_{[2,4]} = \mu^{k/l}_{[1,3]}$ |
| 8 | $\mu^{k/k}_{[1]} \sim \mathcal{N}(\theta^{k/k}_{[1]}, \Sigma^{k/k}_{[1]}) \;\; k = \{a, b\}$ | $\mu^{k/k}_{[3]} = \mu^{l/l}_{[1]} \;\; \mu^{k/l}_{[2,4]} = \mu^{k/l}_{[1,3]} \quad \mu^{a/b} = (\mu^{a/a} + \mu^{b/b})/2$ |

Table 2: **Priors for the Mean Parameters**. The entire prior density on the mean parameters is specified by taking the product of the prior densities in the second column. Column one identifies the model, and column three recalls how the remaining parameters are determined by the free parameters.

| M | Inverse Wishart Priors for Variance Parameters | | Dependent Matrices |
|---|---|---|---|
| | **Unconditional Distributions** | **Conditioning Events** | |
| 0 | $\Omega^{a/a} \sim IW(\Lambda^{a/a}, \nu^{a/a})$ | None | $\Omega^{b/b} = \Omega^{a/b} = \Omega^{a/a}$ |
| 1 | $\Omega^{k/l} \sim IW(\Lambda^{k/l}, \nu^{k/l}) \quad k, l = \{a, b\}$ | None | None |
| 2 | $\Omega^{k/k} \sim IW(\Lambda^{k/k}, \nu^{k/k}) \quad k = \{a, b\}$ | None | $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$ |
| 3 | $\Omega^{a/a} \sim IW(\Lambda^{a/a}, \nu^{a/a})$ | | $\Omega^{b/b} = P\Omega^{a/a}P$ |
| | $\Omega^{a/b} \sim IW(\Lambda^{a/b}, \nu^{a/b})$ | $\Omega^{a/b} = P\Omega^{a/b}P$ | |
| 4 | $\Omega^{a/a} \sim IW(\Lambda^{a/a}, \nu^{a/a})$ | None | $\Omega^{b/b} = P\Omega^{a/a}P$ |
| | | | $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$ |
| 5 | $\Omega^{k/l} \sim IW(\Lambda^{k/l}, \nu^{k/l}) \quad k, l = \{a, b\}$ | $\Omega^{a/a} = Q\Omega^{a/a}Q$ | None |
| | | $\Omega^{a/b} = Q\Omega^{a/b}Q, \;\; \Omega^{b/b} = Q\Omega^{b/b}Q$ | |
| 6 | $\Omega^{k/k} \sim IW(\Lambda^{k/k}, \nu^{k/k}) \quad k = \{a, b\}$ | $\Omega^{a/a} = Q\Omega^{a/a}Q, \; \Omega^{b/b} = Q\Omega^{b/b}Q$ | $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$ |
| 7 | $\Omega^{a/a} \sim IW(\Lambda^{a/a}, \nu^{a/a})$ | $\Omega^{a/a} = Q\Omega^{a/a}Q,$ | $\Omega^{b/b} = P\Omega^{a/a}P$ |
| | $\Omega^{a/b} \sim IW(\Lambda^{a/b}, \nu^{a/b})$ | $\Omega^{a/b} = P\Omega^{a/b}P, \; \Omega^{a/b} = Q\Omega^{a/b}Q$ | |
| 8 | $\Omega^{a/a} \sim IW(\Lambda^{a/a}, \nu^{a/a})$ | $\Omega^{a/a} = Q\Omega^{a/a}Q$ | $\Omega^{b/b} = P\Omega^{a/a}P$ |
| | | | $\Omega^{a/b} = (\Omega^{a/a} + \Omega^{b/b})/2$ |

Table 3: **Constrained IW Priors for the Variance Parameters.** Column one identifies the model. Column two lists the free variance matrices and the unconditional IW priors assumed for them; different free variance matrices $\Omega^{k/l}$ and $\Omega^{j/i}$ are independent. Column three lists the constraints on which the prior distribution has to be conditioned. Column four recalls how the remaining parameters are determined by the free parameters.

| M | 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 | $V^{k/l} \sim CW(\Lambda_V^{k/l}, \nu_V^{k/l}) \quad W^{k/l} \sim CW(\Lambda_W^{k/l}, \nu_W^{k/l}); \; k, l = \{a, b\}$ |
| 6 | $V^{k/k} \sim CW(\Lambda_V^{k/k}, \nu_V^{k/k}) \quad W^{k/k} \sim CW(\Lambda_W^{k/k}, \nu_W^{k/k}); \; k = \{a, b\}$ |
| 7 | $\Omega^{a/a} \sim W(\Lambda^{a/a}, \nu^{a/a}) \quad e_i^{\lambda} \sim \text{Gamma}(a_i, b_i) \;\; ind.$ |
| 8 | $V^{a/a} \sim CW(\Lambda_V^{a/a}, \nu_V^{a/a}) \quad W^{a/a} \sim CW(\Lambda_W^{a/a}, \nu_W^{a/a})$ |

Table 4: **Wishart Prior Distributions on the Free Variance Parameters**. Column one identifies the model. Column two 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.

**Genotypes Called with Posterior Probabilities of Miscalling Less than 0.01**

| | | Pedigree | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Individual** | $a/a$ | 0.330 | 0.000 | 0.000 |
| | $a/b$ | 0.000 | 0.345 | 0.000 |
| | $b/b$ | 0.000 | 0.000 | 0.323 |

| | | Individual | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Affymetrix** | $a/a$ | 0.340 | 0.005 | 0.000 |
| | $a/b$ | 0.003 | 0.308 | 0.003 |
| | $b/b$ | 0.000 | 0.006 | 0.334 |

| | | Pedigree | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Affymetrix** | $a/a$ | 0.340 | 0.005 | 0.000 |
| | $a/b$ | 0.004 | 0.304 | 0.004 |
| | $b/b$ | 0.000 | 0.006 | 0.335 |

**All Genoytpes Called**

| | | Pedigree | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Individual** | $a/a$ | 0.318 | 0.003 | 0.001 |
| | $a/b$ | 0.004 | 0.355 | 0.004 |
| | $b/b$ | 0.001 | 0.003 | 0.312 |

| | | Individual | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Affymetrix** | $a/a$ | 0.330 | 0.013 | 0.000 |
| | $a/b$ | 0.007 | 0.303 | 0.006 |
| | $b/b$ | 0.001 | 0.013 | 0.326 |

| | | Pedigree | | |
|---|---|---|---|---|
| | | $a/a$ | $a/b$ | $b/b$ |
| **Affymetrix** | $a/a$ | 0.333 | 0.011 | 0.000 |
| | $a/b$ | 0.006 | 0.304 | 0.006 |
| | $b/b$ | 0.001 | 0.011 | 0.328 |

Table 5: **Comparison of Genotype Calls**. The genotypes called by the Affymetrix software are contrasted with the individual and pedigree-based calls obtained by our hierarchical Bayes procedure under a Wishart prior and model 1-7. The top half of the table considers only calls with high posterior probabilities.

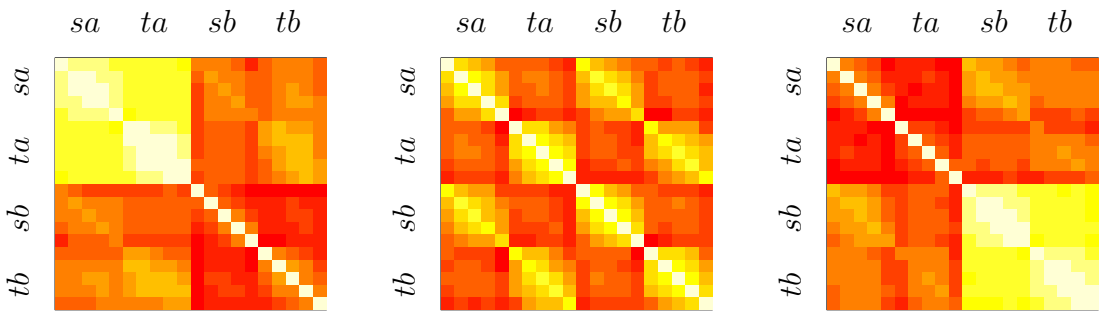


Figure 1: **Heat-Map Image of the Average Correlations between SNP Signals**. The map involves 20 signals per SNP and averages over 500 SNPs. From left to right the genotypes are $a/a$, $a/b$, and $b/b$. Lighter colors correspond to higher correlations. The signals occur in the following order: five probes for $a$ allele, sense; five probes for $a$ allele, antisense; five probes for $b$ allele, sense; and five probes for $b$ allele, antisense.

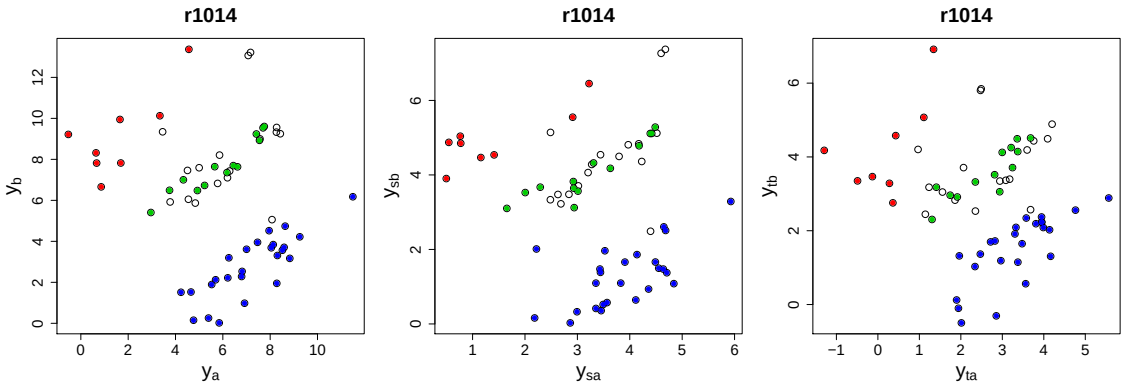


Figure 2: **Scatter Plot of the Four Statistics for SNP r1014**. From left to right we plot the average of $y_{sa}$ and $y_{ta}$ versus the average of $y_{sb}$ and $y_{tb}$, $y_{sa}$ versus $y_{sb}$, and $y_{ta}$ versus $y_{tb}$. The clusters corresponding to the three genotypes can be easily identified. Samples that are called by the Affymetrix software are indicated with filled circles, blue for $a/a$, green for $a/b$, and red for $b/b$. The substantial positive correlation between $a$ allele and $b$ allele probes is obvious.

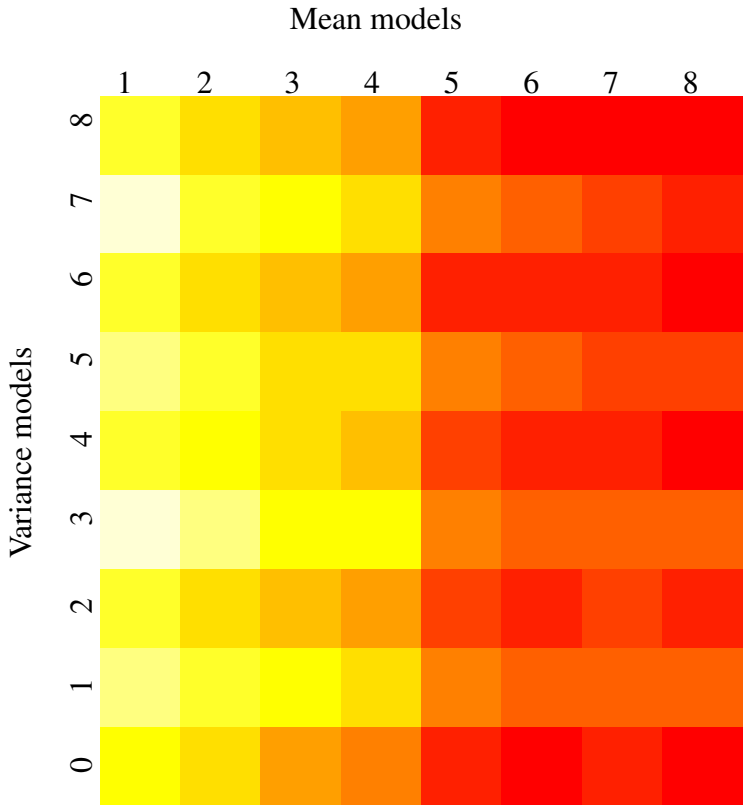


Figure 3: **Heat-map Image of the BIC Values for the 72 Models**. Each column corresponds to a different mean model (from left to right, models 1 through 8) and each row to a different variance model (from top to bottom, models 0 through 8).

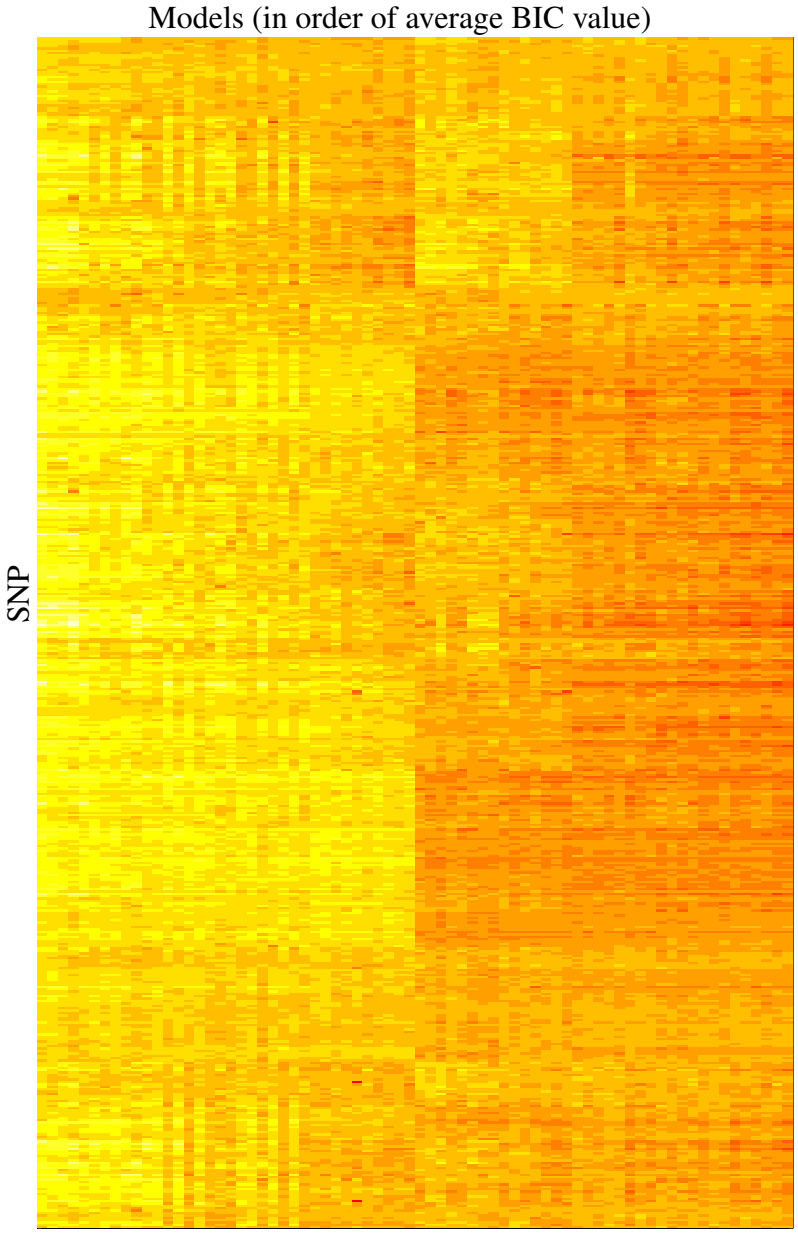


Figure 4: **Heat-Map Image of the Maximum Likelihood for Each Combination of SNP (vertical axis) and Model (horizontal axis)**. The maximum likelihood values for a given SNP are standardized by subtracting the average maximum likelihood across models for that snp.

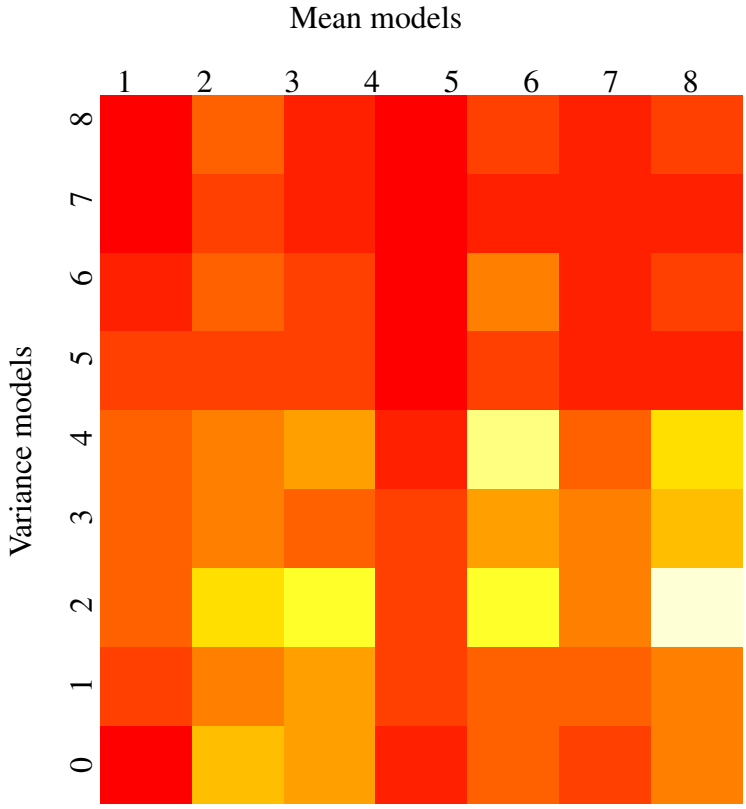


Figure 5: **Heat-Map Image of the Fraction of Discordant Calls**. Each column corresponds to a different mean model (from left to right, models 1 through 8) and each row to a different variance model (from top to bottom, models 0 through 8). Red corresponds to low values and yellow to high values.

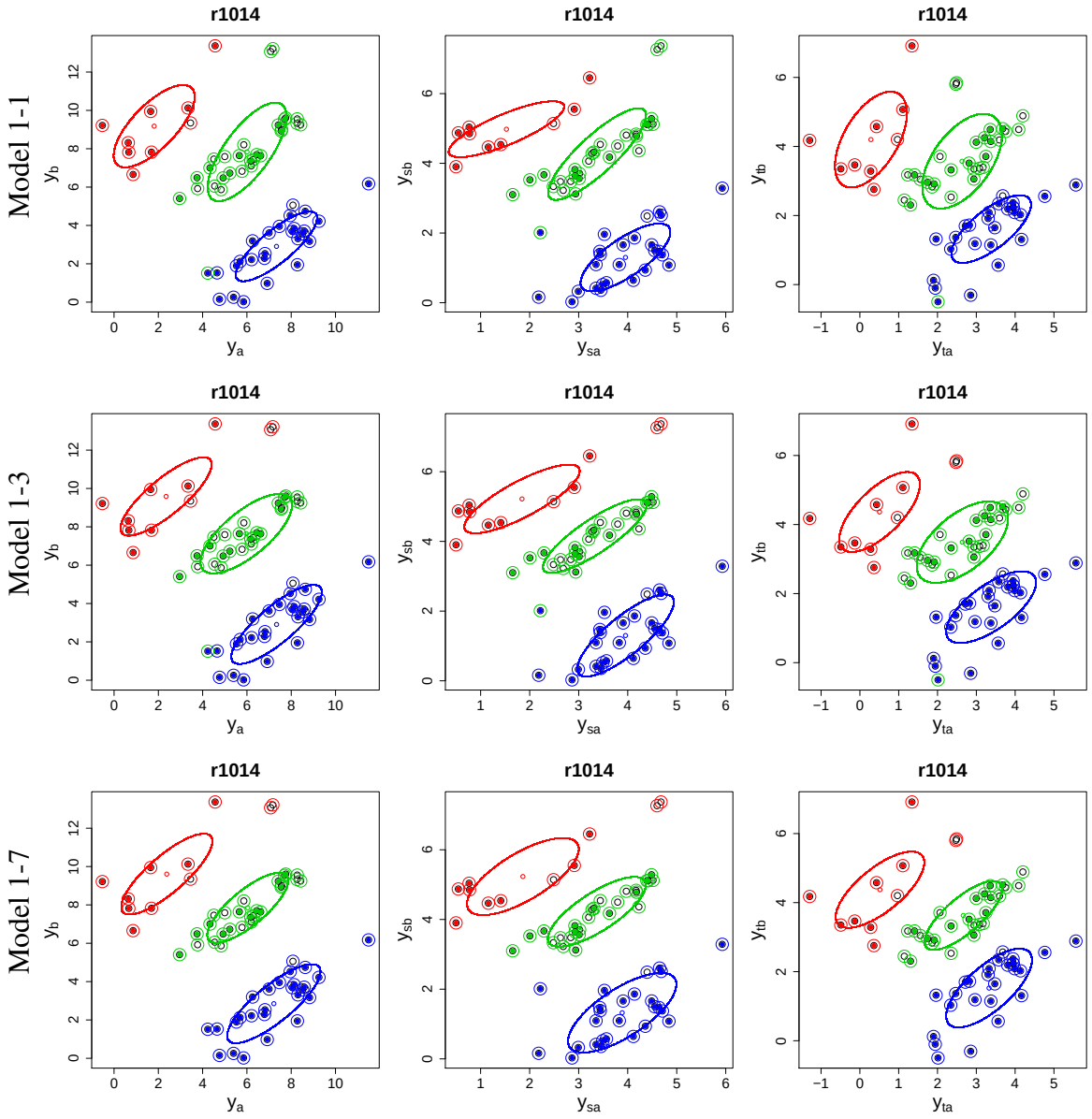


Figure 6: **Maximum Likelihood Estimates and Genotype Calls for SNP r1014**. The top row depicts model 1-1, corresponding to unrestricted means and variance parameters; the center row depicts model 1-3, corresponding to allelic symmetry in the variance parameters; and the bottom row depicts model 1-7, corresponding to allelic and strand symmetry in the variance parameters. From left to right occur plots of 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 circles, blue for $a/a$, green for $a/b$, and red for $b/b$. Samples called by our procedure are indicated by unfilled larger circles and the same color code. A colored ellipse surrounds each estimated mean with scale determined by the estimated variance matrix.

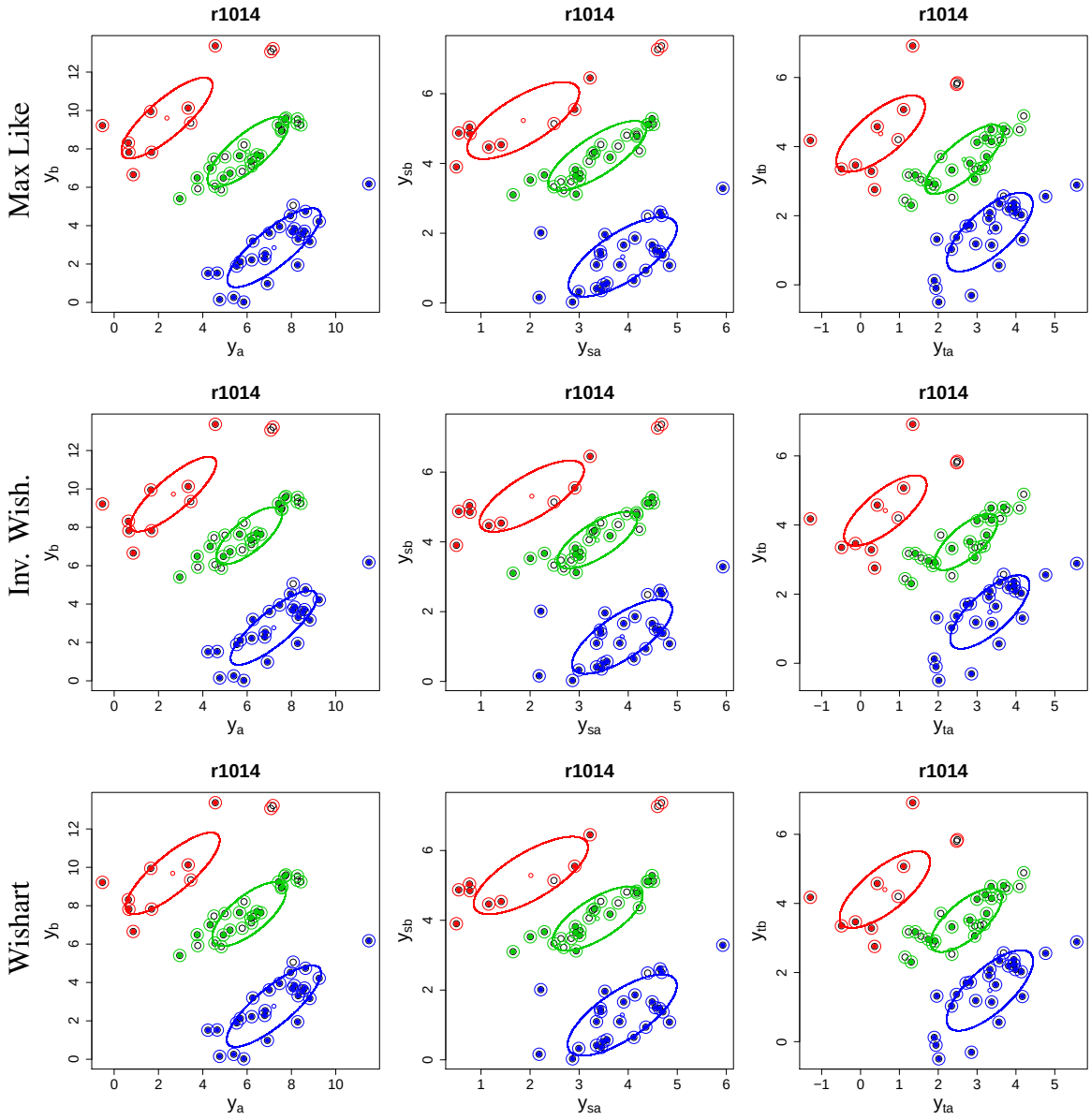


Figure 7: **Comparison of Maximum Likelihood and Maximum a Posteriori Estimates for SNP r1014**. The top row shows the maximum likelihood estimates, the center row the maximum a posteriori estimates with inverse Wishart priors, and the bottom row the maximum a posteriori estimates with Wishart priors. Model 1-7 is assumed throughout. From left to right occur plots of 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 circles, blue for $a/a$, green for $a/b$, and red for $b/b$. Samples called by our procedure are indicated by unfilled larger circles and the same color code. A colored ellipse surrounds each estimated mean with scale determined by the estimated variance matrix.

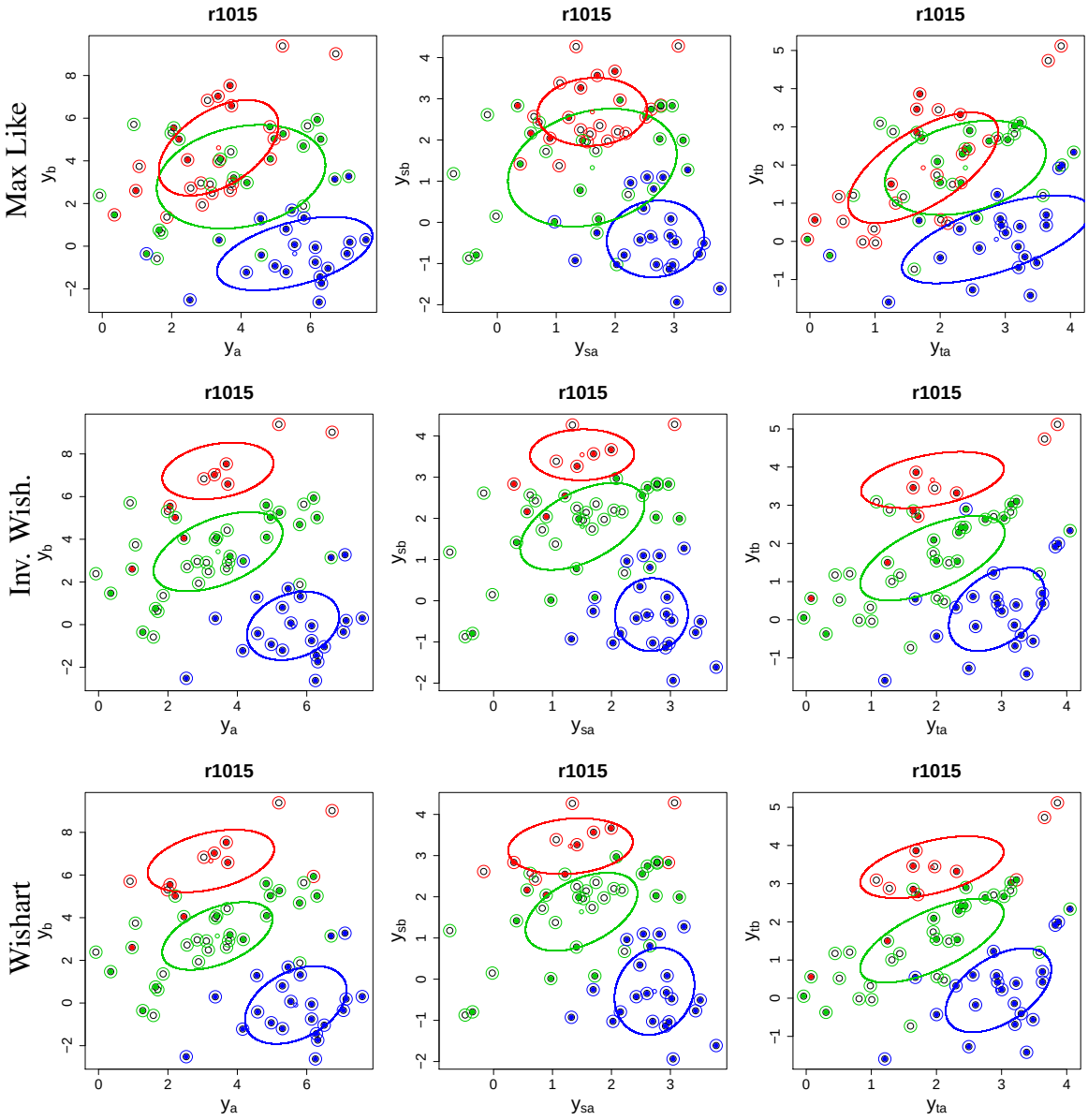


Figure 8: **Comparison of Maximum Likelihood and Maximum a Posteriori Estimates for SNP r1015**. The same conventions hold as in Figure 7. Note the poor definition of the genotype clusters for this SNP.

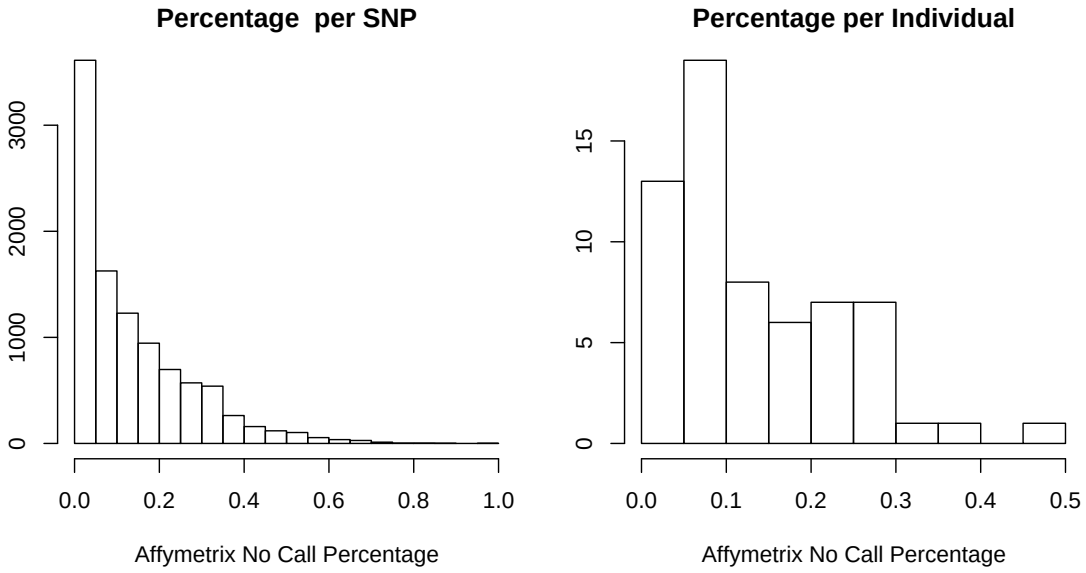


Figure 9: **Histograms of the Proportions of Affymetrix No Calls**. The left hand side shows the proportions within a SNP, and the right hand side the proportions within an individual.

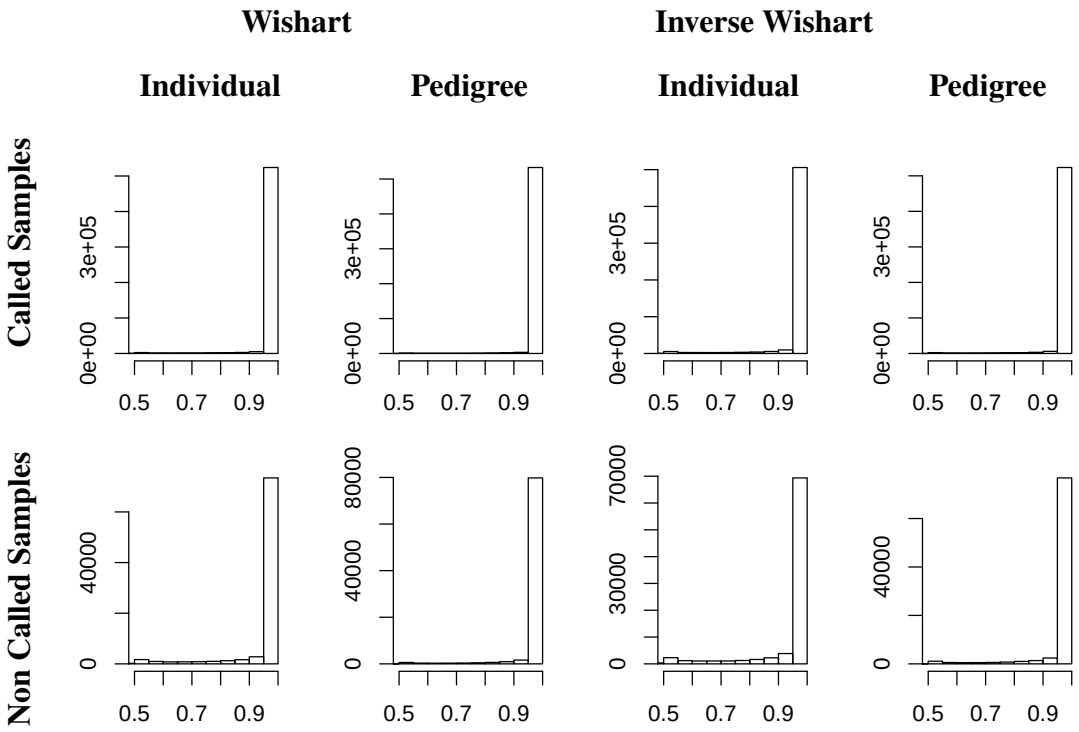


Figure 10: **Histograms of the Purity Index**. On the top our empirical Bayes procedure is applied to the entire sample; on the bottom it is applied to the subset of people classified as no calls by the Affymetrix software.

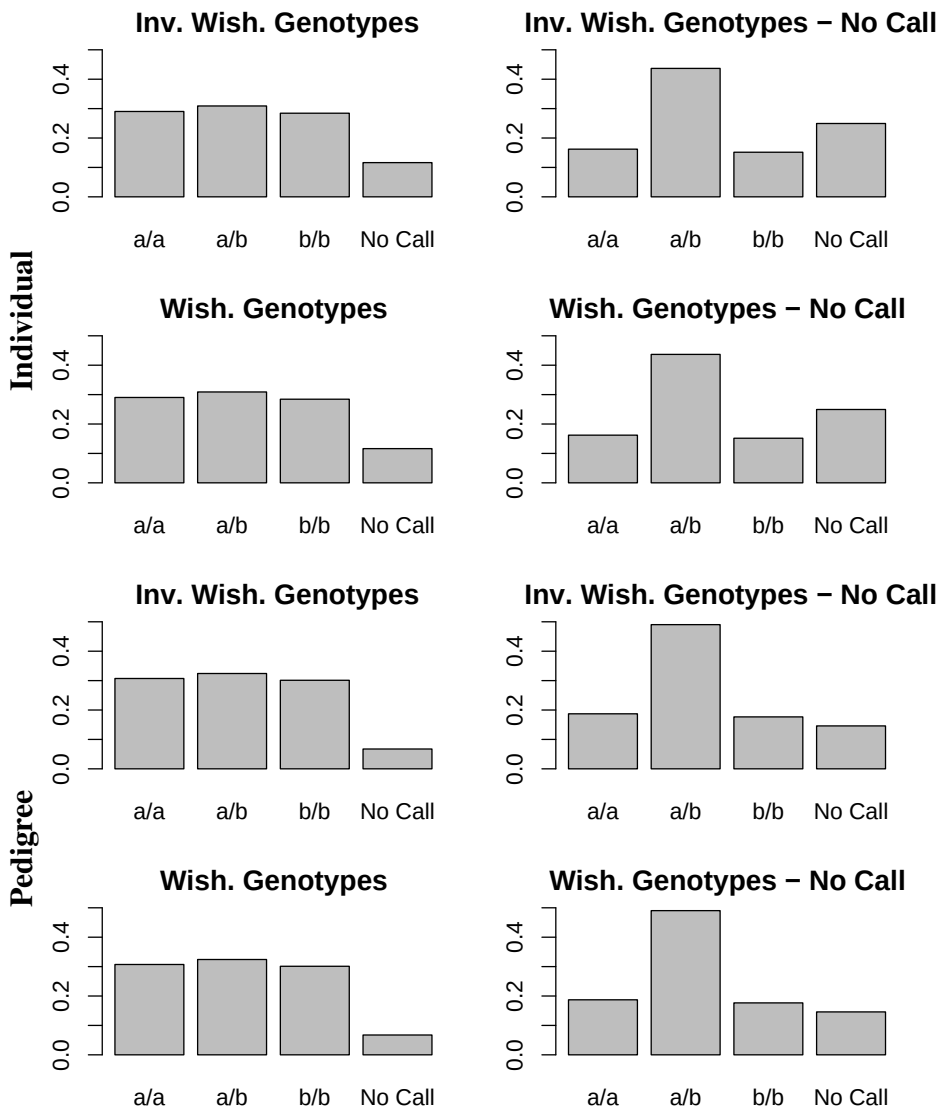


Figure 11: **Comparison of the Posterior Distributions of Assigned Genotypes**. On the left our empirical Bayes procedure is applied to the entire sample; on the right it is applied to the subset of people classified as no calls by the Affymetrix software. Model 1-7 is used with both inverse Wishart and Wishart priors on the variance matrices; calls based on individual and pedigree information are both reported.