Published in final edited form as:

J Appl Stat. 2013 ; 40(6): 1372–1381. doi:10.1080/02664763.2013.785499.

A note on statistical method for genotype calling of high-throughput SNP arrays

Jiaqi Yang^{1,★}, Wei Zhang², and Baolin Wu^{1,†}

¹Division of Biostatistics, School of Public Health, University of Minnesota

²Department of Computer Science and Engineering, University of Minnesota

Abstract

We study the genotype calling algorithms for the high-throughput single-nucleotide polymorphism (SNP) arrays. Building upon the novel SNP-RMA preprocessing approach and the state-of-the-art CRLMM approach for genotype calling, we propose a simple modification to better model and combine the information across multiple SNPs with empirical Bayes modeling, which could often significantly improve the genotype calling of CRLMM. Through applications to the HapMap Trio data set and a non-HapMap test set of high quality SNP chips, we illustrate the competitive performance of the proposed method.

Keywords

SNP arrays; empirical Bayes; genotype calling algorithm; mixture model

1. Introduction

The Affymetrix GeneChip System provides a series of microarray platforms that give the feasibility of exploring biological mechanisms on a whole genomic scale, for example, genotyping, copy number estimation, and differential expression identification, etc. Many modern statistical methods have been proposed to bear on various data analysis procedures involved in processing and analyzing the microarray raw measurement data and have shown a promising perspective.

In this paper, we focus on the single nucleotide polymorphism (SNP) microarrays that can be used for high throughput genotyping in genome-wide association studies (e.g., [9, 16, 17, 23]). In SNP array genotyping, we need to determine sample genotypes based on the measured relative intensities of two alleles. A naive approach based on thresholding the two allele intensity ratios does not work well due to various background noise and systematic variations. Many sophisticated statistical methods have been proposed for genotype calling. For example, for their 10K arrays Affymetrix used a modified partitioning around the medoids clustering algorithm to associate clusters with 3 different genotypes [13]. The algorithm worked well when there were enough data for each genotype, but not as well in other cases. For 100K arrays, Affymetrix changed their default procedure to a dynamic model based algorithm that uses Gaussian models for the probe intensities of each SNP, and calls genotypes for each sample based on the likelihoods for each genotype [4]. Rabbee and Speed [15] proposed a robust linear model with Mahalanobis distance (RLMM) that is a

supervised modeling approach building upon the known genotype of HapMap samples to predict other samples and has been shown to greatly outperform the Affymetrix method. For the 100K and 500K SNP chip arrays, Affymetrix introduced the Bayesian Robust Linear Model with Mahalanobis Distance Classifier (BRLMM) algorithm [1]. It adopted the dynamic model estimates as a prior for cluster characteristics of all SNPs and re-calibrated each SNP estimate in an ad hoc Bayesian manner. For the 6.0 array, Affymetrix used the Birdseed algorithm [10], which is a modification of the the RLMM genotyping algorithm. Carvalho *et al.* [2] proposed a very novel corrected RLMM procedure (CRLMM) that used empirical Bayes modeling to combine the information across multiple SNPs and improve the genotype prediction of individual SNPs. Similar approaches based on the empirical Bayes modeling has also been successfully applied to improve the genome-wide copy number prediction [22]. CRLMM has been shown to greatly outperform all the other methods [2, 12]. In this paper we propose a simple modification to the state of the art CRLMM approach and show that it could significantly improve the genotype calling of CRLMM through applications to the HapMap Trio data set and a non-HapMap test set of high quality SNP chips.

The rest of the paper is organized as follows. Section 2 gives a review of Affymetrix SNP arrays preprocessing and the CRLMM methodology for genotype calling, and describes a modified empirical Bayes modeling approach for better modeling and combining information across multiple SNPs. In Section 3, we compare the proposed method to CRLMM using the HapMap Trio data set, which consists of 30 trios and is part of the International HapMap Project with precise genotype calls that can be used as the “gold standard”, and a non-HapMap test set of 32 high quality SNP chips. We end the paper with a discussion in section 4.

2. Preprocessing and genotype calling of Affymetrix SNP arrays

In a typical Affymetrix high-throughput SNP array, hundreds of thousands of SNPs are interrogated simultaneously. For each SNP, four sets of perfect match probes (PM) are used to measure the intensities of two alleles, denoted as A and B, from sense and antisense strands, denoted as $+/ -$. In addition, mismatch probes (MM) are also included in the array for the purpose of measuring background noise, e.g., non-specific binding etc. And they have been commonly used to correct the bias of PM probe intensities [4, 13]. However as demonstrated in, e.g., Irizarry *et al.* [8] and Carvalho *et al.* [2], we can often improve the preprocessing and downstream statistical analysis by just using the PM probe intensities in studying high-dimensional Affymetrix gene chip data. Therefore in our following discussions we just focus on the PM probe intensities.

The purpose of preprocessing is to remove systematic and non-biological variation and make the intensities comparable across different samples. In preprocessing high-throughput SNP arrays, regression modeling and quantile normalization are first used to remove probe sequence and DNA fragment length effects and unwanted array-to-array variation on the log intensities as in the widely used robust multi-chip average (RMA) approach for preprocessing high-throughput microarray gene expression data [8].

After preprocessing, denote the log intensity measurement as θ_{ijas} for SNP $i = 1, \dots, m$ and sample $j = 1, \dots, n$, where allele $a = A, B$ and strand $s = +, -$. In genotype calling, we need to decide which allele has relatively larger intensities (corresponding to homozygous genotypes, AA or BB) or whether they have the same intensities (corresponding to heterozygous genotype, AB). Intuitively we just need to consider the log ratios, $M_{ijs} = \theta_{ijAs} - \theta_{ijBs}$, for calling genotypes. Denote the three genotypes (AA, AB, BB) by (1, 2, 3). There

are mainly two modeling steps involved, within each sample and across samples, in the CRLMM approach for genotype calling [2].

1. For sample j , further remove the unwanted variation due to probe sequence, DNA fragment length, and total intensities by explicitly modeling both strands of all SNPs with a finite normal mixture model

$$\pi_{1j}N(f_{j1}(X_{ijs}), \sigma_{1j}^2) + \pi_{2j}N(f_{j2}(X_{ijs}), \sigma_{2j}^2) + \pi_{3j}N(f_{j3}(X_{ijs}), \sigma_{1j}^2), \quad \sum_{k=1}^3 \pi_{kj} = 1.$$

Here covariate X_{ijs} includes fragment length L_i , average intensity $S_{ijs} = \theta_{ijAs} + \theta_{ijBs}$, and a factor coding the base pair bp_i for SNP i , and $f_{j2} = 0$, $f_{j3} = -f_{j1}$ with

$$f_{j1}(X_{ijs}) = \mu_{bp_i} + f_L(L_i) + f_S(S_{ijs}),$$

where μ_{bp_i} represents the mean level for each SNP base pair $bp_i \in \{AC, AG, AT, CG, CT, GT\}$, f_L is modeled as a cubic spline with 3 degrees of freedom, and f_S a cubic spline with 5 degrees of freedom. With hundreds of thousands of SNPs, we can obtain very precise estimates of f_{j1} using the EM algorithm [3].

2. Given a set of training samples with known genotypes Z_{ij} , characterize genotype specific intensities of each SNP with a mixed-effects model

$$[M_{ijs}|Z_{ij}=k, m_{iks}] = f_{jk}(X_{ijs}) + m_{iks} + \varepsilon_{ijks},$$

where m_{iks} is the SNP specific shift from the typical genotype region centers not accounted for by the covariate X_{ijs} , and ε_{ijks} represent the random error assumed to follow normal distribution $N(0, \sigma_{iks}^2)$. When not enough data are available, an inverse χ^2 prior is used to improve estimate, $1/\sigma_{iks}^2 \sim \chi_{d_{0k}}^2 / \{d_{0k} s_{0k}^2\}$. The vector of SNP-specific region centers, $m_i = (m_{i1+}, m_{i2+}, m_{i3+}, m_{i1-}, m_{i2-}, m_{i3-})$, are treated as random effects and modeled with a multivariate normal distribution $N(0, V)$.

We can obtain estimates of prior parameters, (d_{0k}, s_{0k}^2, V) , by empirical Bayes modeling of all SNPs [18]. Then to combine information of all SNPs, we use the posterior mean of m_i and σ_{iks}^2 (conditional on all SNPs) in the Gaussian likelihood to predict genotypes for future samples.

In the previous two steps, we have tried to simultaneously model all SNPs in order to borrow information to improve individual SNP inference. Our intuitive idea is that if we can make all SNPs more similar to each other, we could fit the empirical Bayes model more accurately and hence better model and combine information across SNPs. We propose to consistently make the first homozygous genotype formed by the minor allele (estimated based on the set of training samples with known genotypes) for all SNPs to achieve more similarity among different SNPs (see Appendix for some justification). Specifically in calculation of the log ratios for SNP i , $M_{ijs} = \theta_{ijAs} - \theta_{ijBs}$, we consistently make allele A the major allele among training samples across all SNPs.

In the next section we consider two numerical examples. In the first example, we use cross validation on the 90 HapMap samples (with known genotypes as the gold standard) to compare the genotype calling performance of the proposed method (denoted as 'hCRLMM') relative to CRLMM. In addition, we also illustrate the impact of genotype calling algorithms

on the genetic association study. In the second example, we train both genotype calling algorithms on the HapMap samples to call the genotypes of a non-HapMap test set of high quality SNP chips. These two examples will illustrate the relative performance of genotype calling within and across different sample pools.

3. Numerical studies

3.1 Genotype calling of HapMap trio data set

The HapMap (CEPH) Trio data set consists of 30 trios and is part of the International HapMap Project with precise genotype calls that can be used as “gold standard” [19]. We analyzed their Affymetrix GeneChip Human Mapping 100K array data publicly available at http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx. Each array simultaneously interrogates 116204 SNPs. We train genotype calling algorithms on the training set with randomly selected 60 samples, which are then used to predict the testing set with the other 30 samples.

Table 1 shows the genotype calling results for a testing set of randomly selected 30 samples. We can see that the proposed method miss-called total 87566 SNPs, while CRLMM miss-called 97111 SNPs. In total 65982 SNPs are miss-called by both methods. Overall the proposed method has relatively smaller genotype calling error rate, 2.51% compared to 2.79% for CRLMM, and performs significantly better with $p\text{-value} < 10^{-15}$ based on the McNemar test [14].

We have repeated the random cross validation fifty times and consistently observed the typical results of Table 1: the proposed method always has significantly less genotype calling errors than CRLMM (all McNemar test $p\text{-values}$ are less than 10^{-15}). Figure 1 summarizes the genotype calling errors over 50 cross validations for both methods using boxplot. Based on 50 cross validations, we can see that the proposed method performed significantly better than CRLMM ($p\text{-value} = 0.5^{50} = 10^{-15}$).

The Hardy-Weinberg equilibrium (HWE) test has been commonly used as a genetic association study quality control [5, 20, 21]. Here we show that accurate genotyping is critical for appropriate HWE test and small genotyping error can often lead to largely inflated false positives and discard some potentially causal genetic markers. For a test set of 30 HapMap samples, we use an exact test based on the conditional distribution of the number of heterozygotes to test HWE [6, 7, 24]. We count the number of SNPs with $p\text{-values}$ less than $\frac{\alpha}{m}$ based on the Bonferroni correction, where m is the total number of SNPs. Table 2 summarizes the results based on the two genotype calling algorithms and the gold standard for $\alpha = 0.1, 0.05, 0.01, 0.005, 0.001$. We can see that overall the proposed method has a smaller rejection rate compared to CRLMM, and both methods have inflated false positives. We have consistently observed similar results across all test sets.

In the following, we explore the difference between the two genotype calling algorithms in details using a training set of 60 HapMap samples.

3.2 Comparative exploratory analysis of genotype calling algorithms

For a selected training set of 60 HapMap samples, we observed that 57403 out of total 116204 SNPs have A/B alleles switched. Table 3 compares the genotyping errors stratified by whether the A/B alleles are switched. The proposed method has genotyping errors of 2.44% and 2.58% respectively, compared to 2.74% and 2.83% for CRLMM. The differences are significant with both McNemar test $p\text{-values}$ less than 10^{-15} . For each SNP, we also compare the number of miss-called samples (out of total 30 HapMap test samples) for both methods. Figure 2 summarizes their count differences across all SNPs and stratified by

whether the A/B alleles are switched. We define one method performs better for a SNP when it has smaller number of mis-called samples, and two methods perform equally well when they mis-called the same number of samples for a SNP. Table 4 summarizes the results. Overall the proposed method performs better than CRLMM (all binomial test p-values less than 10^{-20}).

Figure 3 compares the histogram of M_{ijs} for CRLMM and hCRLMM for the 30 test samples. For the CRLMM (based on the raw data), we can observe roughly equal proportions for the positive mode around 2 and the negative mode at -2 . While for hCRLMM, we observe the majority is around the positive mode around 2 as expected.

Figure 4 shows the genotype calling details for one SNP where CRLMM miss-called 19 samples and hCRLMM correctly called all 30 samples (A/B alleles switched). Here we have plotted the log intensity ratios (denoted as $M_{+/-}$ for sense and anti-sense strands), the estimated covariate effects $f_{j1+/-}$ for both strands, the predicted SNP-specific region centers $m_{+/-}$, and the difference between $M_{+/-}$ and the corresponding $f_{jk+/-}$. Here samples are plotted in different shape and color based on their true genotypes. We can see that for CRLMM, the predicted $m_{+/-}$ for the missing genotype (not observed in the training set, shown as black circle in the m_+ versus m_- plot) have large negative values, which caused some test samples to be mis-called. While for the proposed hCRLMM, majority of the SNPs have positive log intensity ratios (see figure 3). And the predicted $m_{+/-}$ for the missing genotype had relatively small negative values and were close to the other two genotypes. As a result, it didn't bias the calling of all samples for the hCRLMM. Figure 5 shows another SNP where CRLMM correctly called all 30 samples, while hCRLMM mis-called 18 samples due to the predicted large positive $m_{+/-}$ values of the missing genotype (shown as green plus in the m_+ versus m_- plot).

3.3 Genotype calling of non-HapMap samples based on the HapMap trios

In principle we need to have some data with “gold standard” to compare the genotype calling algorithms. Here we use a test set consisting of 32 high quality replicate SNP chips run on a single non-HapMap individual DNA sample [12]. Specifically we use the “40 Chakravarti replicate Xba chips” available at <http://rafalab.jhsph.edu/snppaffycomp> (eight replicates of low quality were removed as suggested in [12]). For each SNP, the genotype calls comprising more than 50% of the calls is designated the gold standard call following [12].

Firstly we run both genotype calling algorithms based on randomly selected 60 HapMap samples to call the 32 replicate SNP chips. And we also use all 90 HapMap samples to train both genotype calling algorithms.

Table 5 shows the genotype calling results for the 32 replicate SNP chips based on the 60 HapMap samples. The proposed method miss-called 71549 SNPs, while CRLMM miss-called 74191 SNPs. In total 52789 SNPs are miss-called by both methods. The proposed method has a smaller genotype calling error rate, 3.79% compared to 3.93% for CRLMM, and performs significantly better with a McNemar test p-value $< 10^{-15}$. When using all 90 HapMap samples to train both algorithms, we observe similar results: the proposed method has a calling error rate of 4.31% compared to 4.40% for CRLMM with the McNemar test p-value $< 10^{-15}$.

4. Discussion

As it has been a common theme for high-throughput and high-dimensional biomedical data, we often observe small number of samples while hundreds of thousands of features (e.g.,

proteins/genes/SNPs). The empirical Bayes modeling approach has proven to be extremely useful for utilizing the parallel structure of multiple features to help improve the individual feature inference, which could perform very poorly due to small sample size when analyzed in isolation without information pooling from other features. For high-throughput SNP array genotype calling, our intuitive idea is to make multiple SNPs as homogeneous as possible so that we could combine and borrow more information by empirical Bayes modeling to improve individual SNP inference. Specifically we build upon the state of the art CRLMM genotype calling algorithm and consistently make the first homozygous genotype formed by the major allele (estimated based on the training set) across multiple SNPs. So SNPs are more similar to each other, and we can achieve a better empirical Bayes model fitting and borrow more information across SNPs. When applied to the HapMap Trio and a high quality non-HapMap data, we consistently observed that the proposed method outperformed CRLMM with significantly smaller genotype calling errors.

Acknowledgments

This research was supported in part by a Biomedical Informatics and Computational Biology grant from the University of Minnesota-Mayo-IBM Collaboration, and NIH grant GM083345 and CA134848. We would like to thank two anonymous referees for their constructive comments that have improved the presentation of the paper.

References

1. Affymetrix. BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set. 2006. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf
2. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*. 2007; 8:485–499. [PubMed: 17189563]
3. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1977; 39 (1):1–38.
4. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen Mm, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*. 2005; 21 (9):1958–1963. [PubMed: 15657097]
5. Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N. Hardy-weinberg quality control. *Annals of human genetics*. 1999; 63 (Pt 6):535–538. [PubMed: 11246455]
6. Graffelman J. The number of markers in the HapMap project: some notes on chi-square and exact tests for Hardy-Weinberg equilibrium. *American Journal of Human Genetics*. 2010; 86 (5):813–818. [PubMed: 20466092]
7. Haldane JBS. An exact test for randomness of mating. *Journal of Geneticss*. 1951; 52:631–635.
8. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4 (2):249–264. [PubMed: 12925520]
9. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor h polymorphism in age-related macular degeneration. *Science*. 2005; 308 (5720):385–389. [PubMed: 15761122]
10. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics*. 2008; 40 (10):1253–1260. [PubMed: 18776909]
11. Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951; 22:79–86.

12. Lin S, Carvalho B, Cutler DJ, Arking DE, Chakravarti A, Irizarry RA. Validation and extension of an empirical bayes method for SNP calling on affymetrix microarrays. *Genome Biology*. 2008; 9 (4):R63. [PubMed: 18387188]
13. Liu W, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G, Jones KW, Kennedy GC, Kulp D. Algorithms for large-scale genotyping microarrays. *Bioinformatics*. 2003; 19 (18):2397–2403. [PubMed: 14668223]
14. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947; 12:153–157. [PubMed: 20254758]
15. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*. 2006; 22 (1):7–12. [PubMed: 16267090]
16. Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PIW, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettrec G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rstam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjgren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007; 316 (5829):1331–1336. [PubMed: 17463246]
17. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*. 2007; 316 (5829):1341–1345. [PubMed: 17463248]
18. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3:Article 1.
19. The International HapMap Consortium. The International HapMap Project. *Nature*. 2003; 426 (6968):789–796. [PubMed: 14685227]
20. Thakkestantian A, McElduff P, D'Este C, Duffy D, Attia J. A method for meta-analysis of molecular association studies. *Statistics in medicine*. 2005; 24 (9):1291–1306. [PubMed: 15568190]
21. Xu J, Turner A, Little J, Bleeker ER, Meyers DA. Positive results in association studies are associated with departure from hardy-weinberg equilibrium: hint for genotyping error? *Human genetics*. 2002; 111 (6):573–574. [PubMed: 12516594]
22. Wang W, Carvalho B, Miller ND, Pevsner J, Chakravarti A, Irizarry RA. Estimating genome-wide copy number using allele-specific mixture models. *Journal of computational biology*. 2008; 15 (7): 857–866. [PubMed: 18707534]
23. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447 (7145):661–678. [PubMed: 17554300]
24. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of hardy-weinberg equilibrium. *American Journal of Human Genetics*. 2005; 76 (5):887–893. [PubMed: 15789306]

Appendix

SNP genotype frequency distribution difference

Consider two SNPs with their first allele frequencies denoted as π and η . Without loss of generality assume $\pi < 0.5$. Assuming Hardy-Weinberg equilibrium, we can derive the genotype frequencies as $\{\pi^2, 2\pi(1 - \pi), (1 - \pi)^2\}$ and $\{\eta^2, 2\eta(1 - \eta), (1 - \eta)^2\}$. We can characterize the SNP differences by their Kullback-Leibler (KL) distance [11]

$$L_1 = \pi^2 \log(\pi^2 / \eta^2) + 2\pi(1-\pi) \log\{\pi(1-\pi) / \eta / (1-\eta)\} + (1-\pi)^2 \log\{(1-\pi)^2 / (1-\eta)^2\}.$$

If we switch the two alleles for the second SNP, their KL distance is then

$$L_2 = \pi^2 \log\{\pi^2 / (1-\eta)^2\} + 2\pi(1-\pi) \log\{\pi(1-\pi) / \eta / (1-\eta)\} + (1-\pi)^2 \log\{(1-\pi)^2 / \eta^2\}.$$

We can easily check that

$$L_2 - L_1 = 2(2\pi - 1) \log \frac{\eta}{1-\eta},$$

which is positive when $\eta < 0.5$. Therefore the two SNPs have smaller genotype frequency distribution difference if both their first alleles have minor frequencies. It is easily verified that the previous results hold when we switch the two SNPs for calculating their KL distance. This partially justifies our proposed approach of consistently making the first homozygous genotype formed by the minor allele (estimated based on the training set) for all SNPs.

Figure 1. Genotype calling errors (%) over 50 cross validations for the HapMap samples.

Figure 2.
Genotype calling comparison for 30 HapMap samples.

Figure 3.
Histogram of log ratios M_{irs} for 30 HapMap samples.

Figure 4.
Genotype calling of 30 test samples for one SNP (A/B allele switched).

Figure 5.
Genotype calling of 30 test samples for one SNP (A/B allele not switched).

Table 1

Genotype calling results for 30 HapMap test samples.

		hCRLMM		
		Correct	Wrong	Sum
CRLMM	Correct	3367425	21584	3389009
	Wrong	31129	65982	97111
Sum		3398554	87566	3486120

Table 2

Number of rejected SNPs for HWE test based on 30 HapMap samples.

α	0.1	0.05	0.01	0.005	0.001
CRLMM	1211	988	627	471	289
hCRLMM	1007	822	546	408	245
Gold standard	56	56	36	36	1

Table 3

Detailed genotype calling results for 30 HapMap test samples.

A/B alleles switched (57403 SNPs)				
hCRLMM				
		Correct	Wrong	Sum
CRLMM	Correct	1664380	10466	1674846
	Wrong	15624	31620	47244
Sum		1680004	42086	1722090
A/B alleles not switched (58801 SNPs)				
hCRLMM				
		Correct	Wrong	Sum
CRLMM	Correct	1703045	11118	1714163
	Wrong	15505	34362	49867
Sum		1718550	45480	1764030

Table 4

SNP calling comparison for 30 HapMap test samples.

hCRLMM versus CRLMM	worse	equally well	better
116204 SNPs	6238	100103	9863
57403 SNPs (A/B alleles switched)	3004	49454	4945
58801 SNPs (A/B alleles not switched)	3234	50649	4918

Table 5

Genotype calling results for 32 high quality Chakravarti replicate Xba chips (trained on 60 HapMap samples).

		hCRLMM		
		Correct	Wrong	Sum
CRLMM	Correct	1792649	18760	1811409
	Wrong	21402	52789	74191
Sum		1814051	71549	1885600