BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 1 2007, pages 57–63
doi:10.1093/bioinformatics/btl536

Genetics and population analysis

SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays

Jianping Hua[1], David W. Craig[2], Marcel Brun[1], Jennifer Webster[2], Victoria Zismann[2], Waibhav Tembe[1], Keta Joshipura[2], Matthew J. Huentelman[2], Edward R. Dougherty[1,3] and Dietrich A. Stephan[2,*]

[1]Computational Biology Division, [2]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, 445 N 5th Street, Phoenix, AZ, USA and [3]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

Received on July 28, 2006; revised on September 25, 2006; accepted on October 12, 2006
Advance Access publication October 24, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: The technology to genotype single nucleotide polymorphisms (SNPs) at extremely high densities provides for hypothesis-free genome-wide scans for common polymorphisms associated with complex disease. However, we find that some errors introduced by commonly employed genotyping algorithms may lead to inflation of false associations between markers and phenotype.

Results: We have developed a novel SNP genotype calling program, SNiPer-High Density (SNiPer-HD), for highly accurate genotype calling across hundreds of thousands of SNPs. The program employs an expectation-maximization (EM) algorithm with parameters based on a training sample set. The algorithm choice allows for highly accurate genotyping for most SNPs. Also, we introduce a quality control metric for each assayed SNP, such that poor-behaving SNPs can be filtered using a metric correlating to genotype class separation in the calling algorithm. SNiPer-HD is superior to the standard dynamic modeling algorithm and is complementary and non-redundant to other algorithms, such as BRLMM. Implementing multiple algorithms together may provide highly accurate genotyping calls, without inflation of false positives due to systematically miss-called SNPs. A reliable and accurate set of SNP genotypes for increasingly dense panels will eliminate some false association signals and false negative signals, allowing for rapid identification of disease susceptibility loci for complex traits.

Availability: SNiPer-HD is available at TGen's website: http://www.tgen.org/neurogenomics/data.

Contact: dstephan@tgen.org

1 INTRODUCTION

While high density genotyping of hundreds of thousands of single nucleotide polymorphisms (SNPs) has rapidly become available, the development of sophisticated genotype calling algorithms and analysis paradigms for high-density association analysis has lagged (Craig and Stephan, 2005). Within the past three years, the number of SNPs that can be genotyped within a single SNP microarray panel has grown from a thousand SNPs to hundreds of thousands of SNPs, and will soon exceed one million. The main driving force for the increasing SNP genotyping density is the long-held desire of geneticists to complete genome-wide association (GWA) studies using hundreds of thousands of SNPs to holistically scour the genome and identify associations between cases and controls, which would allow for localization of functional DNA variants predisposing to disease (Craig and Stephan, 2005). The exact number of SNPs to cover the majority of the genome is still being debated, but recent analysis of Phase II of the HapMap suggests that at least 250 000 well placed SNPs will be sufficient in a population with Asian or Caucasian descent (Altshuler *et al*., 2005; Thorisson, *et al*., 2005). Both Affymetrix and Illumina now provide platforms allowing for genotyping of SNPs greater than this amount. Thus one may be generating billions of genotypes in a straightforward case-control genome-wide SNP association study across hundreds or thousands of individuals.

One of the greatest unappreciated difficulties in a GWA study is accurately calling the genotypes for billions of SNPs. This is especially difficult given that of several hundred thousand SNPs, only a few SNPs may be associated with the underlying disorder. At the most fundamental level, a genotype must be called based on signal intensities for the two possible SNP alleles. This seemingly trivial problem grows when one considers that genotype calls must be made with extremely high accuracy (Huentelman *et al*., 2005). If genotype calling is biased or of low precision, the SNPs that are found to be most significant will be the result of genotyping error. For example, in one of the first successful genome-wide SNP association studies, 96 cases and 50 controls were genotyped on the Affymetrix 100K GeneChip Mapping array with two SNPs found surviving Bonferroni correction (Klein *et al*., 2005). While the most significant SNP was found to replicate and truly be associated with the phenotype, the second highly significant SNP was likely the result of genotyping error indicated by being significantly out of

*To whom correspondence should be addressed.

Hardy–Weinberg equilibrium (HWE). In this case, the concern about genotyping error is evident with only the >20 million genotypes performed in this study. In a second example, we randomly place 100 individuals genotyped on the Affymetrix 500K platform in one cohort and 100 individuals randomly into a second cohort. We use deviations from HWE at a $P < 0.05$ as a measure of genotyping accuracy. SNPs were called using the dynamic modeling (DM) genotype calling algorithm in GTYPE 4.008 (P-value setting $= 0.26$). Arrays with a call rate >90% were used and on average 6% of SNPs were out of HWE. However, if one calculates a Fisher's exact statistic and ranks SNPs by P-value, one finds that of the top 100 SNPs, 45% fail HWE. Permuting class memberships retains these results suggesting that SNPs that are miss-called by algorithms are more likely to be ranked as significantly associated with the disease (false-positives). It has been anecdotally suggested that enrichment of genotyping false positives may result from biased under-calling of heterozygotes, however extensive testing of this hypothesis has yet to be completed. Regardless, it is clear that highly accurate genotype calling is critical for the success and streamlined implementation of future GWA studies. In this report, we present a newly developed algorithm that significantly outperforms the DM algorithm from Affymetrix, and is complementary and non-redundant to the BRLMM algorithm.

2 METHODS

2.1 500K genotyping

Array-based SNP genotyping: samples were obtained from brain banks under approved IRB from Caucasian donors, and DNA extracted using Gentra DNA extraction kits according to manufacturer's instructions, and processed as described in the Mapping 500K Protocol (Affymetrix).

2.2 SNiPer-HD

SNiPer-HD utilizes a multi-chip-based genotyping algorithm. Unlike the current dynamic modeling (DM) approach implemented in GTYPE 4.008 (Di *et al*., 2005), SNiPer-HD treats the genotyping as a classification procedure, and designs a parametric classifier for each SNP by applying expectation-maximization (EM) clustering to estimate the distribution parameters. Besides giving calls, SNiPer-HD also provides two quality control indicators to help filter out bad calls: a confidence score (per-sample per SNP quality measure) and a quality index (a SNP quality measure across a cohort). The confidence score works on individual calls like a P-value: the lower the score, the more reliable the call. The quality index works on each SNP as a whole: the higher the quality, the more reliable the SNP is over all samples. Consequentially, one may assess genotype accuracy for a SNP, as compared to all other SNPs on the platform, while assessing significance of an association signal.

SNiPer-HD works on one SNP at a time. Consider for the current SNP, D probe quartets (for the 500K chipset $D = 6$ or 10) are used to detect the genotypes on the sense, anti-sense or both-strands. Each probe quartet contains two probe pairs, one for SNP allele A, another for allele B. Each probe pair consists of perfect match (PM) and mismatch (MM) sequences for the target allele. A relative allele signal (RAS) for the dth probe quartet is defined as

$$x_d = A_d/(A_d + B_d), \quad (1)$$

where A_d is the PM signal of allele A, and B_d of allele B. Note that this definition is slightly different from what originally defined in (Liu *et al*., 2003), but identical to the one suggested in the discussion section of the same paper. The major difference is that the mismatch signals are no longer used here. Based on our observations, including background subtraction in a ratio-based measure can increase the signal variance significantly and induce too many outliers.

The SNP at sample point i is represented by RAS-value vector $X_i = (x_1, x_2, \ldots, x_D)$. Clearly if X is close to $(0, 0, \ldots, 0)$, then allele B dominates all probes and the genotype should be BB. If X_i is close to $(1, 1, \ldots, 1)$, then the genotype should be AA. And if X_i is close to $(0.5, 0.5, \ldots, 0.5)$, then the genotype is AB. Thus if there is a whole set of RAS vectors, $\{X_1, X_2, \ldots\}$ on one certain SNP, one should expect them to form three mass concentrations in the D-dimensional space, which could be identified through an appropriate clustering approach. In the ideal case, the RAS vectors of three genotypes AA, AB and BB should be $(1, 1, \ldots, 1)$, $(0.5, 0.5, \ldots, 0.5)$ and $(0, 0, \ldots, 0)$, respectively. However due to background noise, sample points of AA and BB are seldom close to $(1, 1, \ldots, 1)$ and $(0, 0, \ldots, 0)$. We can assume that for each SNP, the RAS vectors are generated from a mixture of three Gaussian distributions, with each Gaussian distribution representing one genotype. If one has the distribution of the RAS vector, then for any sample point, its genotype should be assigned to the one with the highest posterior probability according to the Bayesian rule. Note that although the Gaussian distribution may not perfectly fit the points close to the border, i.e. one or more probes have RAS values close to 0 or 1, there is minimal impact on the genotyping accuracy since they are usually too far away from decision boundaries that lie between genotype centers to cause any mistake in their labels when genotyping.

The SNiPer-HD algorithm has two components: parameter estimation followed by classification. In the parameter estimation component, SNiPer-HD will estimate the distribution parameters of each SNP through an EM clustering algorithm, and give calls to the training sample. The size of the training sample should be large enough to represent the true distribution of each SNP to ensure accurate estimation. In the classification component, SNiPer-HD will give calls to any sample based on the parameters obtained. The outline of SNiPer-HD is summarized as follows:

2.2.1 *Parameter estimation*

(1) Load the RAS values of one SNP from all training sample points;

(2) Based on original calls provided by DM, estimate the number of clusters/genotypes G in the samples, and assign the initial seeds;

(3) Apply EM clustering algorithm;

(4) If the number of clusters/genotypes after clustering is $<G$, remove the empty clusters, use the clustering results as the initial seeds, repeat step 3;

(5) Set the genotypes for all sample points, calculate the confidence scores;

(6) Calculate the reliability of the call results of the current SNP, and save the distribution parameters for future genotyping;

(7) Repeat step 1–6 until all SNPs are processed.

2.2.2 *Classification*

(1) Load the RAS value of one SNP from one sample point;

(2) Based on the distribution parameters estimated, set the call of current SNP with corresponding confidence score;

(3) Repeat steps 1 and 2 until all SNPs from all sample points are processed.

2.3 Parameter estimation

We use three-class labels $\{0, 1, 2\}$ to represent three genotypes $\{AA, AB, BB\}$, respectively. We will use class and genotype interchangeably. For each SNP, we assume the RAS vector is drawn from a certain three-class Gaussian mixture distribution. To be exact, the RAS vector is drawn from the three genotypes with prior probabilities τ_0, τ_1 and τ_2. A SNP of genotype k has its RAS vector X generated according to the corresponding class conditional

distribution, which is a Gaussian:

$$f_k(X \mid \mu_k, \Sigma_k) = \frac{\exp[-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)]}{(2\pi)^{D/2} \mid \Sigma_k \mid^{1/2}}, \quad (2)$$

where μ_k is the mean vector and Σ_k the covariance matrix. For each SNP, τ_k, μ_k and Σ_k, $k = 0, 1, 2$, are the distribution parameters to be estimated. Here we assume the covariance matrices to be spherical and of equal volumes: $\Sigma_0 = \Sigma_1 = \Sigma_2 = \lambda \boldsymbol{I}$. For each RAS vector X_i, there is an indicator vector $Z_i = (z_{i0}, z_{i1}, z_{i2})$, where z_{ik} indicate the posterior probability that X_i is generated from genotype k. Assume there are altogether N sample points, then the likelihood of all data is

$$L = \sum_{i=1}^{N} \sum_{k=0}^{2} z_{ik}[\log \tau_k f_k(X \mid \mu_k, \Sigma_k)]. \quad (3)$$

The EM clustering algorithm (Fraley and Raftery, 1998) is then used to estimate the parameters of the above Gaussian mixture model:

(1) Initialization: $z_{ik} = 1$, if according DM the genotype of sample i is k, with its confidence score <0.26; otherwise, $z_{ik} = 0$;

(2) M-step: compute τ_k, μ_k and Σ_k;

$$\tau_k = n_k/N \quad (4)$$

$$\mu_k = \sum_{i=1}^{N} z_{ik} X_i / n_k \quad (5)$$

$$\lambda = tr(W)/ND, \quad (6)$$

where, $n_k = \sum_{i=1}^{N} z_{ik}$, $W = \sum_{k=0}^{2} \sum_{i=1}^{N} z_{ik}(X_i - \mu_k)(X_i - \mu_k)^T$ (Celeux and Govaert, 1995);

(3) E-step: compute $f_k(X_i \mid \mu_k, \Sigma_k)$ and z_{ik};
$f_k(X_i \mid \mu_k, \Sigma_k)$ is computed as shown in Equation (2)

$$z_{ik} = \tau_k f_k(X_i \mid \mu_k, \Sigma_k) / \sum_{j=0}^{2} \tau_j f_j(X_i \mid \mu_j, \Sigma_j) \quad (7)$$

(4) Repeat steps 2 and 3 until either the relative change in the overall likelihood, computed according to Equation (3), is smaller than a predefined threshold, or the maximum iteration time is reached.

In SNiPer-HD, the threshold for convergence is set to 0.001, and the maximum number of iteration is set to 30. In most cases, the clustering will converge after 3–5 iterations.

Note that during the E-step, the posterior probability z_{ik} has already been computed for each sample point. Hence for the training sample, one does not need to apply the classification component again; rather the calls can be obtained directly. Once converged, the genotype will be assigned to each individual according to its indicator vector Z_i: the genotype of sample i is k, if $z_{ik} > z_{ij}$, $j = 0, 1, 2$ and $j \neq k$. Several special cases are considered before all parameters and call results are saved:

(1) Ties: if a tie is encountered when comparing z_{ik}, then a value randomly chosen from the genotypes with maximal probability will be assigned. Note that in real situations, a tie rarely happens. Furthermore, even if a tie occurs for $z_{ik} = z_{ij}$, then both z_{ik} and z_{ij} should be <0.5, indicating low confidence in the calling of either call;

(2) Class center order: the centers of the three genotypes *AA*, *AB* and *BB* should be aligned in a descending order when projected onto the diagonal line lined up $(0, 0, \ldots, 0)$ and $(1, 1, \ldots, 1)$ at the RAS space. If that is not the case, swapping of the calls must be made to correct the error. The corresponding model parameters should also be changed;

(3) Missing genotypes: when reordering the class center, it is possible that only one or two genotypes are presented. If only one class is presented, the genotype should be assigned to *AA* or *BB*, depending on whether the genotype center is close to $(1, 1, \ldots, 1)$ or $(0, 0, \ldots, 0)$. If two genotypes are presented, then the two classes will be assigned to either *AA* and *AB*, or *AB* and *BB*, depending on whether the overall mass center of all points is close to $(1, 1, \ldots, 1)$ or $(0, 0, \ldots, 0)$, with the corresponding model parameters being changed.

The accuracy of parameter estimation is highly dependent on the quality of the training samples. For SNiPer-HD, we suggest using only the samples with overall call rate larger than 85% by GTYPE, though it is feasible to use other algorithms, such as BRLMM. Unreliable calls will be presented and are largely a function of poor probe quality and/or poor DNA sample quality. Hence, SNiPer-HD provides two quality control indicators to help filter out bad calls, one on the individual call level, and one on each SNP as a whole.

For individual genotype calls of a sample, the posterior probability automatically provides a confidence in the call. SNiPer-HD uses $1 - z_{ik}$ as the confidence score. The smaller the confidence score, the more reliable the call result.

Since the outcome of the parameter estimation procedure depends on the quality of the available sample and probe design, it is possible to have poor parameter estimation on some SNPs, which not only induce errors in the training sample, but also propagate into future predictions. To check the quality of SNP parameter estimation as a whole, SNiPer-HD introduces a quality index based on a silhouette (Rousseeuw, 1987). If X_i's genotype is k, then silhouette width of X_i is defined as

$$S(X_i) = \frac{b(X_i) - a(X_i)}{\max[b(X_i), a(X_i)]}, \quad (8)$$

where $a(X_i)$ is the average distance between X_i and all other sample points of genotype k, while $b(X_i)$ is the minimum of the two average distances between X_i and the points of another genotype. A Euclidean distance metric is used in SNiPer-HD. The silhouette index of genotype k is then defined as the median value of silhouette widths over all sample points of genotype k. The silhouette value ranges from -1 to 1. The higher the value, the more compact and separated is the genotype, thus more reliable the calls on that genotype. SNiPer-HD picks the smallest silhouette index among all three genotypes as the quality index of the SNP.

2.4 Classification

With the distribution information obtained from the parameter estimation, the classification component of SNiPer-HD gives calls on any sample by calculating the posterior probability directly. For the current SNP, we load its corresponding parameters τ_k, μ_k and Σ_k. If the RAS vector of the sample point i is X_i, then SNiPer-HD simply applies the E-step in the EM clustering algorithm to calculate $f_k(X_i \mid \mu_k, \Sigma_k)$ and z_{ik}, and assigns the genotype and confidence score according to z_{ik}. Note that if one applies the classification to the samples that were used in the parameter estimation component, then one will obtain exactly the same call. To ensure the classification performance, the testing samples should be from the same ethnic and geographic population as the training samples.

3 RESULTS

For this study, we compiled a database of ~900 individuals, which have been genotyped in our lab. To construct our database we removed arrays that have call rates <85% when using the DM algorithm with the default setting. Of critical importance, the remaining 500K arrays are not of perfect quality, but they do represent what researchers are currently encountering in practice using DNA samples of mixed quality. Since the selection is conducted separately on NSP and STY chips, the numbers of chips are not equal between the two types of chips. SNP calling was made and

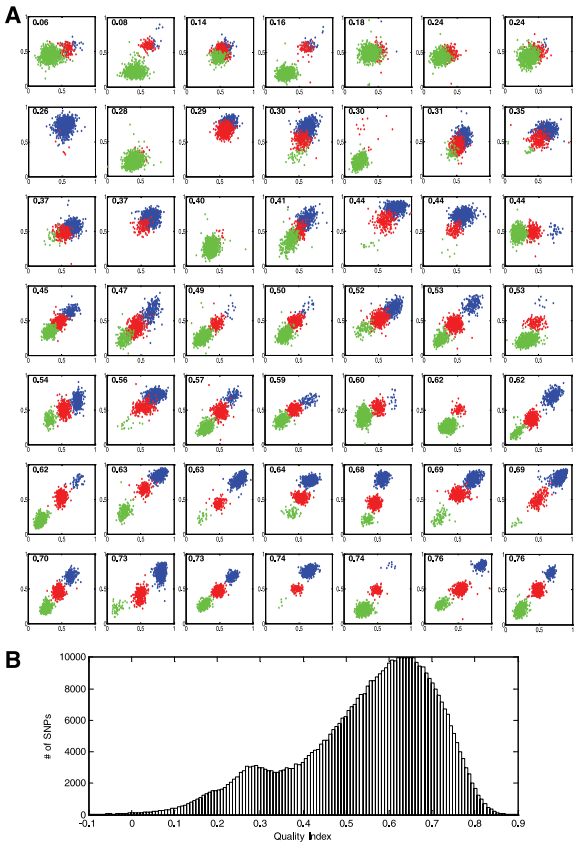

Fig. 1. Genotype calls on 49 SNPs of various quality indices and the distribution of quality index. In (**A**) each scatter-plot corresponds to one SNP. The SNPs are ordered according to the quality index, which is given at the upper left corner of each plot. The SNPs with lowest quality index values are in the top rows, and highest quality index values in the bottom. The *x*- and *y*-axis are two selected RAS values. Genotype *AA* is denoted as blue point, *AB* as red point and *BB* as green point. The histogram of quality index on all SNPs is shown in (**B**). Only the major portion of the whole range of quality index is shown for better viewing.

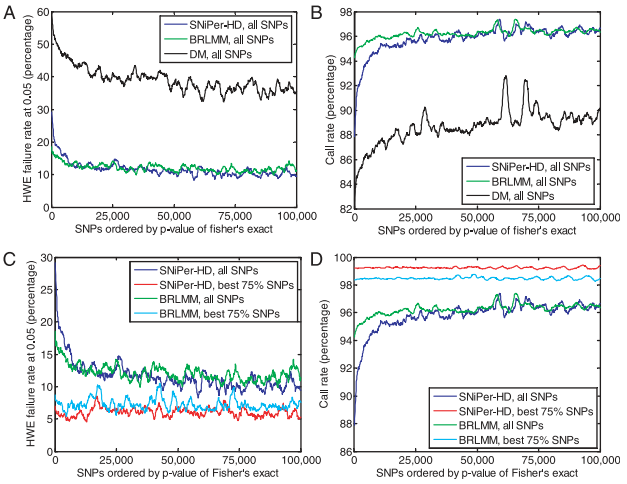

Fig. 2. HWE failure rate/call rate versus top SNPs ordered by *P*-value of Fisher's exact on DM, SNiPer-HD and BRLMM calls of all samples. Exact test of HWE is used. The *x*-axis is the SNPs ordered by the *P*-value of Fisher's exact on a case-control study, and the *y*-axis is the percentage of SNPs that fail HWE at 0.05 for control samples in (**A** and **C**), and call rate on all individuals in (**B** and **D**). Default settings are used for DM and BRLMM to set 'NoCall'. For SNiPer-HD, 'NoCall' is set to any call with confidence score >0.05.

compared on 939 NSP chips and 834 STY chips, including a set of case-control study samples: 578 cases and 300 controls from NSP chip, and 494 cases and 284 controls from STY chip. Additionally, there are also six technical replicates of the Affymetrix control DNA on the NSP chips and five on the STY chips.

Figure 1A shows the RAS signal intensity values of 49 SNPs that were randomly selected based on the quality index (after SNiPer-HD was run) to illustrate the quality spectrum of SNPs that must be called by any algorithm. The calls of each SNP are shown as a scatter-plot in a two-dimension subspace of the original data space represented by the RAS values of probes. SNPs are ordered according to the quality index, which is marked at the upper left corner of each plot. One can clearly see that if the samples of different genotypes are well separated, that SNP has a high-quality index. For SNPs with quality indices <0.4, samples of two or all three genotypes are mixed up in the plot, and result in inaccurate calls. Figure 1B shows the distribution of quality indices of all SNPs. The peak close to quality index value 0.3 in Figure 1B indicates that a considerable amount of SNP calls are of low reliability. To verify this, we genotyped 10 000 SNPs again by randomly setting seeds in EM clustering to artificially generate a huge amount of bad calls. The resulting quality index shows a strong peak at 0.25–0.3, which confirms our suspicion. While the exact choice of the quality index is arbitrary, it is clear by inspection of Figure 1 that selection of SNPs with a quality index at ∼0.45 will remove those SNPs that are most likely to be poorly called, while still keeping those SNPs with highest separation between genotype classes. As with all calling algorithms, the exact choice of filtering those SNPs most likely to yield poor calls is dependent on an individual's requirements for accuracy while balancing SNP coverage.

The importance of quality control can be seen in the results of the case-control study on typical experimental data generated on the Affymetrix 500K platform using multiple calling algorithms. A Fisher's exact statistic is first applied to the autosomal SNPs between the cases and controls. Next, HWE testing is conducted for each autosomal SNP based on the control samples only. To show the relationship between the two statistics, in the *x*-axis, from left to right, the autosomal SNPs are then ordered in increasing order according to the *P*-value obtained through the Fisher's exact test. In the *y*-axis, we compute the percentage of SNPs that fail HWE at $P < 0.05$. To ensure a smooth curve, the percentage is calculated based on a window size of 2000 SNPs along the ordered SNPs. The results for the most significant 100 000 SNPs are shown in Figure 2A. Along with SNiPer-HD, the results of two other genotype calling algorithms (Affymetrix: DM and BRLMM v1.0) are shown. Default settings are used for DM and BRLMM to set 'NoCall'. For SNiPer-HD, any call with a confidence score >0.05 is set to 'NoCall'. The exact test of HWE was used (Wigginton *et al.*, 2005). Unless specifically mentioned, this set-up is used for all comparisons shown in this study.

Although the performance of the three algorithms differs dramatically, with BRLMM and SNiPer-HD clearly superior to DM, a

common problem exists in all three: the SNPs with the most significant *P*-values have the highest HWE failure rate, which is reflected by up-biased tails in the extreme left of Figure 2A. This problem is confirmed by checking the average call rates of the top SNPs in a similar manner, as shown in Figure 2B: the top SNPs also have the lowest call rates and the lowest quality indices. There are a considerable number of SNPs that are poorly called, and they represent false-positive association findings. These results indicate that although the new genotyping algorithms (SNiPer-HD and BRLMM) are superior to the DM algorithm, it is still necessary to filter out unreliable SNPs so that false-positive associations do not overwhelm the replication phase of a WGA study.

In the following comparison, we use a quality index threshold of 0.45 to pick out the reliable SNPs from SNiPer-HD calls. The number of SNPs that pass this threshold depends on the overall data quality. For the data in this study, ~76% SNPs (380 330 SNPs) pass the threshold. After removing SNPs on the X chromosome, which do not obey HWE, there are ~75% SNPs (375 000 SNPs) that pass this threshold. We will refer to these SNPs as the 'best 75% SNPs'. In Figure 2C and D, we compare the HWE failure rate and call rate between all SNPs and best 75% SNPs, respectively. Since the call results of DM are far inferior to BRLMM and SNiPer-HD, we only show the results of BRLMM and SNiPer-HD. For comparison, we picked the same number of SNPs from BRLMM's calls based on the call rate, a traditional way to select reliable SNPs. Among the 375 000 SNPs that are called by each algorithm, ~90% of those SNPs (335 554 SNPs) are called (versus 'NoCall') in both algorithms. For the 100 000 SNPs shown in Figure 2C and D, this number is reduced to 80% that are called in common (80 319 SNPs). For the top 100 SNPs, this number is further reduced to 70% that are called in common (70 SNPs). Thus when one compares the top 100 SNPs in the *P*-value list, there is less agreement between calling algorithms, and thus less confidence in the accuracy as measured by agreement of just calls versus 'NoCalls'.

After removing the SNPs with a quality index <0.45, both HWE failure rate and call rate are significantly improved. Furthermore, no observable bias exists in the top SNPs. SNiPer-HD has better performance than BRLMM in both HWE failure rate and call rate for the best 75% SNPs. For the top 100 000 SNPs shown, the HWE failure rates are 6.06 and 7.06%, and the call rates are 99.29 and 98.63%, for SNiPer-HD and BRLMM, respectively. We have also conducted a comparison of HWE over all individuals in our database. The results are shown in Figure 3, where SNiPer-HD is compared with BRLMM and DM for both all SNPs and the best 75% SNPs. SNiPer-HD outperforms BRLMM and DM in all cases. The HWE failure rates at several thresholds are also shown in Table 1.

The reproducibility test is conducted on the available technical replicates. If one call is different from the majority call of all replicates, then a concordance error is reported. 'NoCall' calls are omitted when considering the majority call. The results of reproducibility are shown in Table 2. When all SNPs are considered, DM has the highest concordance and lowest call rate. When best 75% SNPs are considered, all three methods have very good concordance rates. SNiPer-HD has the highest call rate, and a concordance rate of 99.84% is between DM and BRLMM.

We have also considered SNiPer-HD to the available HapMap samples that were genotyped by Affymetrix on the 500K SNP platform. We find that the HapMap sample data are of superior quality compared to the usual sample encountered in our lab, likely a result of optimization of site-specific laboratory techniques and ultra-pure high molecular weight cell line DNA as a starting material. However, due to the limited publicly-available Affymetrix

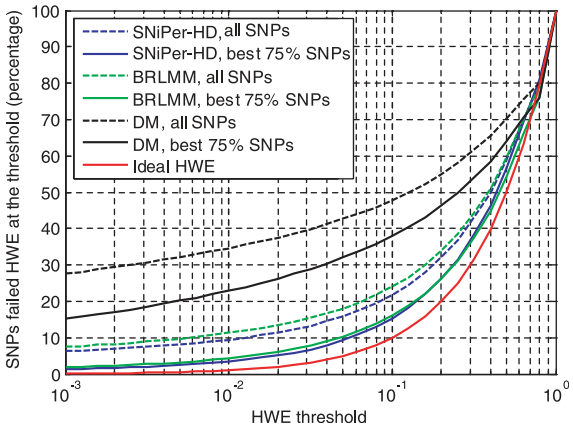

Fig. 3. HWE on DM, BRLMM and SNiPer-HD calls. The *x*-axis is the HWE test threshold, and the *y*-axis is the percentage of SNPs that fails at the corresponding threshold. Exact test of HWE is used. For DM, 'NoCall' is set to any call with confidence score >0.26. For BRLMM, 'NoCall' is set to any call with confidence score >0.5. For SNiPer-HD, 'NoCall' is set to any call with confidence score >0.05.

Table 1. HWE test failure rate of DM, BRLMM and SNiPer-HD genotype calls

	HWE threshold	0.001 (%)	0.005 (%)	0.01 (%)	0.05 (%)	0.10 (%)
All SNPS	DM	27.54	32.13	34.56	42.66	47.82
	BRLMM	7.43	9.73	11.27	18.02	24.03
	SNiPer-HD	6.23	7.67	9.46	15.57	21.83
Top 75% SNPs	DM	15.16	20.14	22.86	32.03	37.98
	BRLMM	1.81	3.24	4.33	10.20	16.06
	SNiPer-HD	1.39	2.58	3.50	9.51	15.35

The percentage of SNPs that fails the HWE test is reported. The exact test of HWE is used. For DM, 'NoCall' is set to any call with confidence score >0.26. For BRLMM, 'NoCall' is set to any call with confidence score >0.5. For SNiPer-HD, 'NoCall' is set to any call with confidence score >0.05.

Table 2. Reproducibility measures of accuracy of the genotype calling algorithms

	All SNPs		Top 75% SNPs	
	Call rate (%)	Concordance (%)	Call rate (%)	Concordance (%)
DM	91.54	99.37	96.36	99.79
BRLMM	97.23	99.15	98.96	99.89
SNiPer-HD	97.09	99.15	99.51	99.84

Comparison is done on technical replicates, six for NSP chip, and five for STY chip. For DM, 'NoCall' is set to any call with confidence score >0.26. For BRLMM, 'NoCall' is set to any call with confidence score >0.5. For SNiPer-HD, 'NoCall' is set to any call with confidence score >0.05.

Table 3. Performance on 15 CEU samples from HapMap data at different control levels

Quality index	>−1 (All SNPs)		>0.45 (Top 75% SNPs)	
Confidence score	Call rate (%)	Concordance (%)	Call rate (%)	Concordance (%)
0.01	98.52	99.42	99.63	99.51
0.05	99.14	99.34	99.79	99.49
0.1	99.37	99.30	99.85	99.48
1	100	99.07	100	99.43

Classifiers are trained on our own 900 sample database, and called on 15 HapMap samples. The call rate and concordance rate to HapMap data are given for different confidence score and quality index thresholds. Only the common SNPs appeared in both 500K chip and HapMap data are checked.

probe intensity data available (39 individuals), comparisons are likely unreliable because both BRLMM and SNiPer-HD require a relatively large training database of individual genotypes (BRLMM requires >48 samples to train for high-quality chips). We applied SNiPer-HD classifiers, which are trained on our own 900 sample database, to the 15 available CEU individuals of HapMap, which are of similar ethnic background to the training data in our database. The results are shown at Table 3. For the top 75% SNPs which also appear in the HapMap data, SNiPer-HD reports 99.8% call rate and 99.5% concordance at confidence score threshold 0.05, which is used in previous comparison. Obviously there is significant improvement in both call rate and concordance by filtering out the low-quality index SNPs. Considering that SNiPer-HD is actually trained on the much lower quality chips, the results are quite satisfactory. If the full dataset of HapMap sample is available, one would expect the results improve in all aspects, especially the quality index. However, for array signal intensity data of this quality, all algorithms should achieve excellent performance, and the difference could be very small, which says little for more realistic scenarios. These results suggest that under ideal situations, such as is the case with the HapMap samples genotyped by Affymetrix, all algorithms perform well. However, when optimal genotyping conditions are not available, such as when limited to DNA of variable quality, training-based algorithms may be more accurate.

4 DISCUSSION

There are several important aspects one has to consider in order to design or judge a genotype calling algorithm. The accuracy and call rate are certainly the two most important measures. One important and more often ignored aspect is the bias towards certain genotypes in the calls. This problem becomes significant when sample size increases. For example, miss-calls or failures to call a genotype in only a few individuals of the homozygous genotype of the minor allele can immediately force a HWE failure, although the overall accuracy and call rate on all genotypes are still high. It is known that the DM algorithm (currently used in Affymetrix GTYPE) is relatively precise, as can be seen from our test of reproducibility in Table 2. But DM often has low call rates, and calls are often biased away from heterozygotes. This deficiency severely affects the accuracy as measured by HWE, and leads to false associations in the case-control GWA studies. As previously highlighted, the first major GWA success by Klein and colleagues also showed a false-positive as the second most significant SNP. As GWA studies move to complex diseases with common variants of lower effect size, these false positives may mask true associations.

In this study, we have developed a highly accurate genotype calling program that includes quality index for each SNP. The quality index introduced is a quality metric that does not depend on call rate, which is often the standard for filtering poor quality SNP data. From Table 3, one can see by picking SNPs of high quality index, the concordance rate with HapMap data significantly increases, and call rate also increases. With the quality index, SNiPer-HD can pick out of the good quality SNPs with the most accurate parameter estimation, and which achieve better accuracy in their calls. This numerical estimate can be visually inspected for significant association findings, and lend a level of confidence to any single SNP association finding that would be unwieldy to do by visually inspecting the signal intensities across a large cohort. However, this does not mean that SNPs of low quality index are without use. The quality index can be low due to many reasons. For example, if some individuals have copy number change at certain regions, it can lead to poor initial seeds and poor separation of genotype clusters, which finally induce poor quality index. Hence currently SNiPer-HD will output calls on all SNPs with the corresponding quality index and confidence scores, and one can do their own filtering or joint analysis according to one's own needs.

The major problem in accurately calling SNP genotypes, as opposed to other common classification problems, is the low minor allele frequency. The difficulty is that training-based algorithms require training sets that include enough sample points from all genotypes for reliable parameter estimation. Although the performance improves with the increase of training sample size, efficient sampling is unlikely for tens of thousands of SNPs with very low allelic frequency even the training sample size is considerably large, e.g. 500–1000. Without considering the effects of minor allele frequency, poor separation will lead to poor parameter estimation in many SNPs, especially due to the iterative nature of EM algorithm used. Hence in SNiPer-HD, we carefully select our data model and parameters. We put the prior probability into the mixed Gaussian model to reflect the minor allele frequency of the training sample. Since the small number of individuals of certain genotypic classes can lead to unstable estimation of the parameters of that genotype, we enforce an identical covariance matrix for all genotypes. These settings in the data model not only help the EM algorithm to successfully capture the genotypes of small size, but also provide the right balance to avoid bias toward any certain genotype. For example, in our study, we found that changing the threshold of the confidence score from 0.05 to 0.01 can cause an undetectable change of +0.1% on the HWE failure rate for the best 75% SNPs, although the call rate drops from 99.2% to 98.7%. Importantly, we use a quality metric to identify those SNPs for which SNiPer-HD provides high accuracy calls, since it is unlikely that any algorithm is universally effective at calling over 500 000 SNPs with resilience due to variable assay performance.

Concomitant to the development of SNiPer-HD, other genotyping algorithms are being developed (Cuttler *et al*., 2001; Di *et al*., 2005; Rabbee and Speed, 2006). BRLMM is another classification-based approach that is a modification of the supervised algorithm: Robust Linear Model with Mahalanobis distance classifier (RLMM)

by Speed and colleagues (Rabbee and Speed, 2006). Of major importance is that BRLMM was designed to optimize call accuracy and call rate for the Affymetrix 500K array, after those ~500 000 SNPs had been selected and the platform launched. Conversely, SNiPer-HD was designed to improve accuracy and only call those SNPs that are best called by an EM based approach. Consequentially, we introduce an algorithm that calls ~375 000 of the most informative SNPs on the Affymetrix 500K platform with extremely high accuracy. Fortunately, most SNPs left-out are of very low allelic frequency, which often have low power due to limited information content.

SNiPer-HD and BRLMM differ algorithmically, and thus it is feasible to use algorithms jointly to identify those SNPs that are most likely called correctly. Differences between SNiPer-HD and BRLMM lie in two aspects: the data space and parameter estimation.

BRLMM applies normalization and keeps the probe intensity as a feature of the data space. SNiPer-HD does not normalize the data because we do not see any obvious tailing effects on the M-A plot of probe intensities of randomly selected individuals. The simple RAS value is chosen in SNiPer-HD as our data space, where we believe the ratio of the A to the $A + B$ alleles can correct most of the shifting in the intensity. Other more elaborate transformations may be able to improve the data space concentration of each genotype and increase the separation between different genotypes. Although we achieve excellent results in our present data space, our approach does not limit other novel transforms from being explored in the future.

For parameter estimation, SNiPer-HD and BRLMM choose quite different approaches. While both SNiPer-HD and BRLMM strongly depend on the initial calls of DM, both algorithms improve over it in different ways. SNiPer-HD relies on DM calls to decide how many genotypes presented in the current SNP and set seeds. But it only use this as the start point, and utilizes the iterative nature of the EM algorithm to find the most compact clusters that represent the existing genotypes. By this way, it can correct wrong calls and 'NoCall' calls of DM. If the initial seeds are too bad, the iteration can amplify this by forming poorly separate clusters of low-quality index, which is hard to detect if only utilizing the initial seeds. On the other side, BRLMM uses the Bayesian priors collected from other SNPs to help estimate the genotypes of minor allele, even totally missing ones, which are caused by the low call rate of DM. However, BRLMM self-correction mechanism is limited and the supervised learning relies heavily on the accuracy of DM calls and estimated priors. If DM calls and priors do not match the current SNP, or do not match each other, then there is high probability of genotyping error or low call rate on certain genotypes.

The immediate question for the researchers undertaking a GWA study is which genotyping algorithm or program should be employed. Each algorithm (BRLMM and SNiPer-HD) has strengths and weaknesses, and different GWA analysis strategies have different sensitivities towards accuracy and coverage. In the case where accuracy is critical and one does not require SNPs with low minor allelic frequency, our results found that SNiPer-HD performed marginally better than BRLMM and significantly better than DM (currently implemented by GType). In the case that one requires the greatest number of SNPs called we find that BRLMM provides the greater coverage, with only slightly lower quality. To merge data, we suggest calculation of association statistics, whether they be Fisher's exact, TDT, FBAT, for each algorithm separately so as not to propagate calling inaccuracies. One can then compare rank or statistical significance between algorithms, and weighing the quality index to identify false positives. Ideally, if the same SNP is in the top few SNPs of several hundred thousand statistical tests, those genotype calls are more likely to be accurate.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Affymetrix and those members of the collaborative team developing BRLMM for sharing early alpha and beta versions of their software. The authors acknowledge NIH ENDGAME consortium grant U01-HL086528-01 to D.A.S. and D.W.C., a philanthropic donation by the Stardust Foundation to D.W.C. and NIH grant U24 NS051872 and funds from the State of Arizona to D.A.S. The authors also thank the Kronos Life Sciences division for support in generating the 500K genotype dataset. Funding to pay the Open Access publication charges was provided by TGen institutional funds.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Celeux,G. and Govaert,G. (1995) Gaussian parsimonious clustering models. *Pattern Recognit.*, **28**, 781–793.

Craig,D.W. and Stephan,D.A. (2005) Applications of whole-genome high-density SNP genotyping. *Expert Rev. Mol. Diagn.*, **5**, 159–170.

Cutler,D.J. *et al*. (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res.*, **11**, 1913–1925.

Di,X. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.

Fraley,C. and Raftery,A.E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis *Comp. J.*, **41**, 578–588.

Huentelman,M.J. *et al*. (2005) SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC Genomics*, **6**, 149.

Klein,R.J. *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Liu,W.M. *et al*. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397–2403.

Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.

Rousseeuw,P.J. (1987) A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Thorisson,G.A. *et al*. (2005) The international HapMap project web site. *Genome Res.*, **15**, 1592–1593.

Wigginton,J.E. *et al*. (2005) A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887–893.