BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 1958–1963
doi:10.1093/bioinformatics/bti275

Gene expression

Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays

Xiaojun Di*, Hajime Matsuzaki, Teresa A. Webster, Earl Hubbell, Guoying Liu, Shoulian Dong, Dan Bartell, Jing Huang, Richard Chiles, Geoffrey Yang, Mei-mei Shen, David Kulp, Giulia C. Kennedy, Rui Mei, Keith W. Jones and Simon Cawley

Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

Received on October 25, 2004; revised on December 27, 2004; accepted on January 11, 2005
Advance Access publication January 18, 2005

ABSTRACT

Motivation: A high density of single nucleotide polymorphism (SNP) coverage on the genome is desirable and often an essential requirement for population genetics studies. Region-specific or chromosome-specific linkage studies also benefit from the availability of as many high quality SNPs as possible. The availability of millions of SNPs from both Perlegen and the public domain and the development of an efficient microarray-based assay for genotyping SNPs has brought up some interesting analytical challenges. Effective methods for the selection of optimal subsets of SNPs spanning the genome and methods for accurately calling genotypes from probe hybridization patterns have enabled the development of a new microarray-based system for robustly genotyping over 100 000 SNPs per sample.

Results: We introduce a new dynamic model-based algorithm (DM) for screening over 3 million SNPs and genotyping over 100 000 SNPs. The model is based on four possible underlying states: Null, A, AB and B for each probe quartet. We calculate a probe-level log likelihood for each model and then select between the four competing models with an SNP-level statistical aggregation across multiple probe quartets to provide a high-quality genotype call along with a quality measure of the call. We assess performance with HapMap reference genotypes, informative Mendelian inheritance relationship in families, and consistency between DM and another genotype classification method. At a call rate of 95.91% the concordance with reference genotypes from the HapMap Project is 99.81% based on over 1.5 million genotypes, the Mendelian error rate is 0.018% based on 10 trios, and the consistency between DM and MPAM is 99.90% at a comparable rate of 97.18%. We also develop methods for SNP selection and optimal probe selection.

Availability: The DM algorithm is available in Affymetrix's Genotyping Tools software package and in Affymetrix's GDAS software package. See http://www.affymetrix.com for further information. 10K and 100K mapping array data are available on the Affymetrix website.

Contact: xiaojun_di@affymetrix.com

INTRODUCTION

With the release of the Mapping 10K array, oligonucleotide microarrays are currently being widely used for genotyping single nucleotide polymorphisms (SNPs) (Sellick *et al*., 2003; Kennedy *et al*., 2003; Liu *et al*., 2003; Matsuzaki *et al*., 2004b; Middleton *et al*., 2004; Puffenberger *et al*., 2004; Ranch *et al*., 2004; Koed *et al*., 2005; Shrimpton *et al*., 2004; Woods *et al*., 2004). However, an even higher density of SNP coverage on the genome is desirable and often essential for population studies, such as association. Assembling many thousands of SNPs has been the current priority (Brooks, 1999), the advances of oligonucleotide microarray technology make it possible (Fodor *et al*., 1991, 1993; Pease *et al*., 1993; Fan *et al*., 2000; Dong *et al*., 2001; Kennedy *et al*., 2003). Three technical advances have enabled an increase in SNP content from 10 000 to over 100 000: a several-fold increase in the complexity of the genomic fraction amplified in the assay to access more SNPs, a reduction in feature size to integrate more SNPs on fewer chips and the availability of a large database of SNPs formed by combining SNP data from Perlegen and the public domain.

The starting point is a combined database of around 3 000 000 SNPs to which certain bioinformatics and biochemical constraints are applied to derive a large set of about 500 000 SNPs for empirical validation. It is essential to have an algorithm to empirically screen this large SNP set and select the best SNPs effectively and genotype them accurately. The method developed for analysis of Mapping 10K data, MPAM (Liu *et al*., 2003) is very effective but faces some challenges in scaling up to the development of a 100K chip set. MPAM requires the screening of a large number of samples to observe all three genotypes and build empirical models accurately, which is both cost ineffective and time consuming. Even in a large sample set, it is still difficult to handle SNPs with low minor allele frequency; it is difficult to build models for missing or low-frequency genotypes. For the Mapping 10K chip intensity patterns of all selected SNPs were visually inspected to reject low quality SNPs undetected by systematic screening; however, this approach clearly does not scale well. The MPAM model is robust and accurate but lacks the flexibility to accommodate further improvements and optimization after product release. Changes including further optimization of experimental conditions, scanner settings, etc. may require retraining models to reflect the effects of such changes on MPAM's SNP-specific clusters.

Cutler *et al*. (2001) introduced a model based genotyping approach for high-throughput variation detection microarrays, which demonstrated that it is possible to genotype SNPs without using any prior empirical information. We introduce a new dynamic model-based algorithm (DM) for the Mapping 100K array. The algorithm starts with a probe-level dynamic-model-based likelihood and performs an SNP-level statistical aggregation to provide a high-quality genotype call. By stratifying all possible states into four models Null, A, AB

*To whom correspondence should be addressed.

and B, for a given probe, DM first operates on quartets of probes calculating a log likelihood for each model, then computes a score by subtracting from the log likelihood associated with the model the largest log likelihood of the other three models. The crucial step of DM is the SNP level statistical aggregation. We use a statistical test on multiple quartets of an SNP to aggregate the quartet level scores to an SNP level genotype call along with a confidence metric. We use the one-sided Wilcoxon signed rank test producing four p-values, one for each model; the model associated with the smallest p-value will be the predicted genotype and the smallest p-value itself is the confidence metric of the genotype call.

The DM algorithm also includes methods for SNP screening and probe reduction and optimization which are directly derived from the genotyping algorithm. Being able to screen SNPs on a relatively small sample set DM makes the SNP selection process time and cost efficient.

EXPERIMENT AND CHIP DESIGN

The strategy for selecting an optimal set of 100 000 high-quality SNPs started with an *in silico* screening phase where 500 000 SNPs of the complete set of 3 000 000 were selected due to being located in genomic locations expected to be among the regions amplified by the assay. This led to the design of a screening array set consisting of six chips for each of two enzymes, XbaI and HindIII (a total of 12 types of chip). The screening array set was then used in an empirical screening process in which an optimal set of 100 000 SNPs was identified. The tiling strategy is the same as 10K screening (Liu *et al*., 2003); each SNP consists of 56 probes, 7 probe quartets for each strand with the polymorphic nucleotide having different shifts from the center of the 25-mer probe sequence. For the seven probe quartets, the shifts are set to be

$$-4, -2, -1, 0, +1, +3, +4.$$

A probe quartet includes probe pairs for both alleles A and B. A probe pair includes a perfect match cell and a mismatch cell. A total of 54 ethnically diverse samples were used to generate the screening data set, the goal being to select an optimal set of 50 000 SNPs for each enzyme to be titled on two arrays.

ALGORITHMS

This new approach for genotyping microarray data is based on a probabilistic model for intensities for each probe quartet, followed by aggregation to SNP level with a statistical test. The same underlying model is used as a basis for genotype classification, SNP selection and identification of optimal probe quartets for each SNP. The call quality metric is based on a single numeric value calculated by a Wilcoxon signed rank test.

Probe level likelihood

For generality we assume that there are n probe quartets for each SNP. Each quartet consists of two probe pairs, one pair for each allele. From the image perspective, there are four underlying states for each quartet, that is,

Null. All probes behave similarly; none is expected to be brighter than the rest. In this case, all four cells are treated as background.

A or B. The perfect match cell corresponding to the A (or B) allele is the only significantly bright cell. In this case, the bright perfect match cell is treated as the foreground, with all the other three cells being treated as the background.

AB. Both perfect match cells are bright. In this case, the two perfect match cells are treated as foreground, and the two mismatch cells are treated as background.

To determine the genotype for an SNP from a given probe quartet, the only thing we need to know is which state the probe quartet supports. Since all four states are possible, a more natural question to ask is which state is most likely. We quantitatively calculate the likelihood for each state in order to determine the most likely genotype; the state with the maximum likelihood will be the model best fitting the state. We select between the four models: Null, A, B and AB providing No Call (NC), A, B and AB in terms of genotype call. Let us assume that $L(\text{state})$ is a well-defined likelihood function; we compute four likelihoods for each probe quartet

$$L(1) = L(\text{Null}), \quad L(2) = L(\text{A}), \quad L(3) = L(\text{B}),$$
$$L(4) = L(\text{AB}).$$

SNP level aggregation

Even though probes for each SNP are only a few bases apart from each other, the hybridization efficiencies and level of non-specific cross hybridization can vary significantly. This probe-specific hybridization inconsistency may lead to different probe quartets supporting different states. To handle unexpected hybridization behavior we robustly aggregate over all probe quartets. To this end, define the score for a given model m as

$$S(m) = L(m) - \max\{L(k),\ k = 1, 2, 3, 4,\ k \neq m\}. \qquad (1)$$

Then for each model m we can formulate a vector of the scores with n observations for each SNP,

$$\mathrm{V}_m = \{S_1(m), S_2(m), \ldots, S_n(m)\}, \quad m = 1, 2, 3, 4, \qquad (2)$$

where m are the model indices and n is the number of probe quartets for a given SNP. Notice that for a given model m and quartet i, if $S_i(m) > 0$, then model m would be the best fitting model for that quartet and quartet i supports model m. If $S_i(m) < 0$, then model m would not be the best fitting model for that quartet and quartet i does not support model m. Since we have n quartets, a very natural question to ask is how strong is the support for a given model across all probe quartets. We use a Wilcoxon signed rank test (Hollander and Wolfe, 1999) for a robust non-parametric answer to this question. By applying the one-sided Wilcoxon signed rank test on vector (2) for all four models with hypotheses

$$\begin{aligned} H_0 :&\quad \text{median}\{S_i(m)\} = 0 \\ &\qquad \text{versus} \\ H_1 :&\quad \text{median}\{S_i(m)\} > 0\,, \end{aligned} \qquad (3)$$

we obtain four p-values, $\{p_1, p_2, p_3, p_4\}$. It is straightforward to see that the model with the least p-value will be the best fitting model across all probe quartets. Now sort the above four p-values; the corresponding model m_0 with the least p-value will provide the genotype call, and

$$p \equiv p_{m_0} = \min\{p_1, p_2, p_3, p_4\} \qquad (4)$$

will be the metric for the confidence of the genotype call. A threshold α is set to control the confidence of the genotype calls; if p is greater

than this threshold, the call is defined as a no-call (NC), and therefore, no-call genotype is supported either when Null model is picked or when $p > \alpha$. Most of the analysis described in this manuscript used $\alpha = 0.05$. If there are ties for the least p-value, the order of tie breaker is Null, A, B and AB. When ties happen among A, B and AB, none of the p-values is significant, and hence the final genotype will be no-call after the threshold α is applied.

Identification of optimal probe quartets

Consider a training dataset with l samples. Let

$$\{S_{1j}, S_{2j}, \ldots, S_{nj}\}, \quad j = 1, 2, \ldots, l \tag{5}$$

be the vector of quartet scores associated with the smallest p-value (p_{m_0}) for sample j. For each probe quartet formulate a new vector

$$\{S_{i1}, S_{i2}, \ldots, S_{il}\}, \quad i = 1, 2, \ldots, n. \tag{6}$$

Again by applying one-sided Wilcoxon signed rank test on the above n vectors, we obtain n p-values, one for each probe quartet

$$\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n\}. \tag{7}$$

Notice that the smaller $\hat{p}_i$ is, the more the i-th probe quartet contributes to the aggregated SNP call in each of the l samples. Therefore, by sorting Equation (7), we are able to rank probe quartets according to how strongly they support the call and select optimal subsets of any given size.

We initially tiled 14 probe quartets (7 for each strand) in the screening arrays; then to achieve greater SNP density in the final array we reduced to 10 probe quartets per SNP. Rather than uniformly removing 4 probe quartets across all SNPs we selected 10 optimal probe positions out of 14 for each SNP. Using this probe optimization method we are able to select ~5% more SNPs with the same high quality compared to uniform selection of 10 probe quartets.

Likelihood function

The algorithm design of DM is not specific to any particular likelihood function, as long as it is well-defined corresponding to the stratification of four models: Null, A, B and AB. Here, we use a dynamic model-based likelihood function. For this dynamic model-based likelihood function, to simplify the above models we assume that, for any given cell, the pixel signal intensities are independent, identically distributed (i.i.d.) normal random variables. We also assume that all four cells of a quartet are independent. In addition, we assume that all probe quartets are independent. For a given model, we assume that both foreground and background are evenly distributed, namely, all foreground cell(s) have the same distribution (Gaussian), and all background cells have the same distribution (Gaussian).

Dynamic model-based likelihood function For each probe in a quartet let

$$\mu_x, \ \sigma_x^2, \ n_x, \quad x = 1, 2, 3, 4$$

be the mean, variance and number of observations (number of pixels), respectively. Let

$$\hat{\mu}_x, \ \hat{\sigma}_x^2, \quad x = 1, 2, 3, 4$$

be the estimated mean and variance assuming model m. Because of the assumptions above, an explicit log likelihood function for a quartet can be given as (see Cutler *et al*., 2001)

$$L(m) = -\frac{1}{2}\sum_{x=1}^{4} n_x \left[\ln(2\pi\hat{\sigma}_x^2) + \frac{\sigma_x^2 + (\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2}\right]. \tag{8}$$

To minimize the estimation error, for each of the four models we find maximum-likelihood estimators for all parameters by using the assumptions and by differentiating Equation (8) with respect to all parameters and solving the corresponding system of equations.

For model Null, all four features are assumed as background and evenly distributed; hence we have

$$\begin{aligned} \hat{\mu} \equiv \hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \hat{\mu}_4 &= \frac{\sum_{x=1}^{4} n_x \mu_x}{\sum_{x=1}^{4} n_x}, \\ \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2 = \hat{\sigma}_4^2 &= \frac{\sum_{x=1}^{4} n_x[\sigma_x^2 + \mu_x^2]}{\sum_{x=1}^{4} n_x} - \hat{\mu}^2. \end{aligned} \tag{9}$$

For model A, the perfect match for A allele is assumed as the foreground, and all the other three are assumed as the background and evenly distributed; hence we have

$$\hat{\mu}_1 = \mu_1, \quad \hat{\sigma}_1^2 = \sigma_1^2 \tag{10}$$

and

$$\begin{aligned} \hat{\mu}_2 = \hat{\mu}_3 = \hat{\mu}_4 &= \frac{\sum_{x=2}^{4} n_x \mu_x}{\sum_{x=2}^{4} n_x}, \\ \hat{\sigma}_2^2 = \hat{\sigma}_3^2 = \hat{\sigma}_4^2 &= \frac{\sum_{x=2}^{4} n_x[\sigma_x^2 + (\hat{\mu}_x - \mu_x)^2]}{\sum_{x=2}^{4} n_x}. \end{aligned} \tag{11}$$

For model B, the perfect match for B allele is assumed as the foreground, and all the other three are assumed as the background and evenly distributed; hence we have

$$\hat{\mu}_3 = \mu_3, \quad \hat{\sigma}_3^2 = \sigma_3^2 \tag{12}$$

and

$$\begin{aligned} \hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_4 &= \frac{\sum_{x\neq 3} n_x \mu_x}{\sum_{x\neq 3} n_x}, \\ \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_4^2 &= \frac{\sum_{x\neq 3}^{4} n_x[\sigma_x^2 + (\hat{\mu}_x - \mu_x)^2]}{\sum_{x\neq 3} n_x}. \end{aligned} \tag{13}$$

For model AB, the perfect matches for both A and B allele are assumed as the foreground, and both mismatches are assumed as the background and evenly distributed; hence we have

$$\begin{aligned} \hat{\mu}_1 = \hat{\mu}_3 &= \frac{n_1\mu_1 + n_3\mu_3}{n_1 + n_3}, \\ \hat{\sigma}_1^2 = \hat{\sigma}_3^2 &= \frac{n_1[\sigma_1^2 + (\hat{\mu}_1 - \mu_1)^2] + n_3[\sigma_3^2 + (\hat{\mu}_3 - \mu_3)^2]}{n_1 + n_3} \end{aligned} \tag{14}$$

and

$$\begin{aligned} \hat{\mu}_2 = \hat{\mu}_4 &= \frac{n_2\mu_2 + n_4\mu_4}{n_2 + n_4}, \\ \hat{\sigma}_2^2 = \hat{\sigma}_4^2 &= \frac{n_2[\sigma_2^2 + (\hat{\mu}_2 - \mu_2)^2] + n_4[\sigma_4^2 + (\hat{\mu}_4 - \mu_4)^2]}{n_2 + n_4}. \end{aligned} \tag{15}$$

Observe that if we add or multiply a constant to all the raw pixel intensities, by definition, Equations (1) and (2) will not change, and

hence the p-values will remain the same. Therefore, with the above likelihood function, DM is invariant to linear transformation of the pixel intensities.

Algorithm implementation

The implementation of the DM algorithm is straightforward. Since DM is a single sample-based algorithm and uses only Affymetrix standard cel file, DM always starts from a cel file (or multiple cel files for probe quartet optimization). For any given SNP, the steps of the DM algorithm are summarized as follows:

(1) Read in from a cel file all intensities, standard deviations and number of pixels for a probe quartet.

(2) Calculate the log likelihood $L(\text{Null})$ for the Null model by treating all four features of a probe quartet as background using formulae (8) and (9).

(3) Calculate estimates for model A using formulae (10) and (11).

(4) Check whether background estimate $\hat{\mu}_2$ for model A is greater than foreground estimate $\hat{\mu}_1$. If true set log likelihood $L(\text{A}) = L(\text{Null})$; otherwise calculate the log likelihood $L(\text{A})$ for model A using formula (8) and estimates in step 3.

(5) Repeat steps 3 and 4 for models B and AB.

(6) Repeat steps 1–5 for all available probe quartets for the given SNP.

(7) Formulate score vectors $\text{V}_m, m = 1, 2, 3, 4$ as defined in Equations (1) and (2).

(8) Standardize vector $\text{V}_m, m = 1, 2, 3, 4$ by removing all zeros from them. Zero has no signed rank.

(9) Compute the Wilcoxon signed rank statistics using $\text{V}_m, m = 1, 2, 3, 4$.

(10) Get the corresponding p-values through a pre-built distribution look-up table. If after step 8 the size of V_m is 0, then set $p_m = 0.5$.

(11) Construct a vector of pairs of model with associated p-value in the order of Null, A, B and AB.

(12) Sort the above vector ascendingly on p-values. The model in the first pair is the genotype, and the corresponding p-value is the confidence metric. Note that the preferences for tie breaker between models are Null > A > B > AB. When there are ties between A, B or AB, p-values are very insignificant.

For the implementation of identification of optimal probe quartets, we first loop over all samples for the above 12 steps and identify the genotype for each sample and its corresponding score vector (5); then we formulate the vectors as defined in Equation (6) and apply Wilcoxon signed-rank test to get the p-values for each available probe quartets; then we sort these p-values to get the given number of desired probe quartets. The distribution table was pre-built for the number of observations from 1 to 14 with exact probabilities; for the number of observations >14 we use large-sample approximation. To be able to screen many samples on arrays with a large number of SNPs, we create a sequential access binary file for each cel file so that we can fetch information from multiple samples efficiently for any given SNP.

SNP selection

The goal of the SNP screening process was to select about 100K high quality SNPs for subsequent development efforts, including an entropy-based SNP selection process to choose SNPs that provide the most uniform information across the genome (Hubbell *et al.*, 2004, http://recomb04.sdsc.edu/ posters/hubbell_earlat affymetrix.com_245.pdf). That means selection of ∼20% of the 500K SNPs is necessary. We applied the following rules in our SNP screening process: (1) Call consistency: SNP called in >85% of the samples when SNPs with $p > \alpha$ were defined as no-calls. (2) Overall minor allele frequency: at least two heterozygous calls must be observed. Then α was varied to yield a subset of ∼20%. With a total of 54 samples in the training set, each of the selected SNPs has to call in at least 46 samples with similar high confidence and with at least 2% minor allele frequency.

RESULTS

For the SNP screening phase, we ran hybridization experiments for all 54 ethnically diverse samples: 24 individuals from the Polymorphism Discovery panel (PD), 6 individuals from the Centre du Etude Polymorphisme Humain (CEPH), 12 Caucasians, 6 African-Americans and 6 Asians, for each of our 12 arrays. DM's optimal probe quartet selection method was used on all 54 samples to choose a set of 10 optimal probe quartets for each SNP, after which DM's SNP selection rules were applied to select high quality SNPs, leading to the selection of 144 189 very high quality SNPs. The top 126 757 SNPs (63 379 XbaI and 63 378 HindIII) were selected by using entropy-based SNP selection process. A final selection was done after tiling this set of SNPs across one XbaI and one HindIII array and genotyping 330 individuals including 30 CEPH family trios genotyped by the HapMap Project, giving a set of 116 204 SNPs (58 960 for XbaI, 57 244 for HindIII; see Matsuzaki *et al.*, 2004a). Genotype results along with other information of these SNPs are available on the Affymetrix website.

We assessed our algorithm performance by (1) comparing genotypes with data from the HapMap Project (release 8, June 2004; see HapMap, 2003), (2) measuring Mendelian inheritance consistency and consistency between DM and our 10K genotyping algorithm MPAM and (3) checking genotype call consistency over multiple technical replicates.

Of the 116 204 SNPs we selected, 18 558 SNPs overlap with HapMap release 8. Based on 30 CEPH family trios (90 individuals) at 95.91% call rate, our concordance rate with HapMap genotypes is 99.81% (1 584 607/1 587 624). Since HapMap defines its genotypes as CC,CT,CG, etc. and it is highly unlikely for the two homozygote genotypes to be switched, we ignored possible discordant comparisons between the two homozygote genotypes.

Ten trios (including five CEPH family trios genotyped by HapMap project) not being used in any of the selection processes were used to check Mendelian inheritance errors by counting calls with inconsistent Mendelian relationship. Our Mendelian inheritance inconsistency is 0.018% (501/2 833 137).

The reproducibility of the DM algorithm was measured by checking the inconsistency errors over technical replicates. We ran five technical replicates on nine different samples (not included in the samples used for SNP selection), then genotyped them independently. Inconsistentency errors were counted by comparing genotype call on each of the five replicates to the consensus call, where

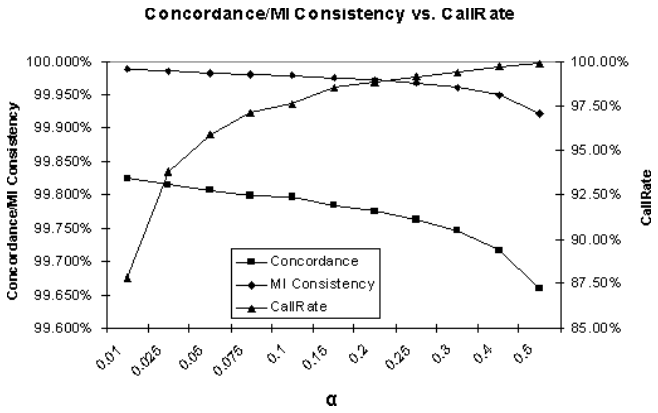

Fig. 1. The vertical axis on the right is the average call rate over all 90 CEPH family samples; the vertical axis on the left is either HapMap reference genotype concordance based on all 90 CEPH family samples, or Mendelian inheritance consistency based on 10 CEPH trios with 5 trios among the 90 CEPH family samples; the horizontal axis is the variation of α. Notice that the more stringent the threshold α, the higher the concordance/MI consistency and the lower the call rate. Notice that as the threshold α decreases, concordance/accuracy increases.

consensus call was defined as the majority call in five replicates, while no-calls were omitted from the consensus building and comparisons. At $\alpha = 0.05$, there were 47 inconsistency errors out of 4 976 938 comparisons with an average call rate of 94.13%, where the average was taken over 90 experiments (5 replicates per sample, 9 samples on 2 arrays).

All 11 555 SNPs from the Mapping 10K array were tiled onto the screening array(s), 8014 of them ended among the final set of 116 204 SNPs, of which we have genotypes for both 10K and 100K mapping arrays for 19 CEPH family samples. The consensus rate between DM and MPAM on this subset of SNPs is 99.90% (147 814/147 964) for all comparable genotypes (both algorithms making genotype calls) and the comparable rate is 97.18%. This shows great consistency between the two algorithms on the common subset of SNPs.

The DM algorithm is very flexible for balancing call rate and accuracy. By adjusting the cutoff of the confidence metric α, one can easily get higher call rates or even higher accuracy, as shown in Figure 1 and Table 1.

CONCLUSION

The dynamic model-based approach presented here provides a highly accurate genotype calling method, is effective for SNP screening, is robust against changes in experimental conditions, flexible to experiment designs, and scalable to more SNPs. Future work includes extension to account for probe-specific effects and to analyze multiple chips simultaneously; similar generalizations have been very effective in the expression analysis field and are likely to be of benefit in the genotyping context too. The combination of the assay, chip and analysis method provides a cheap and extensible solution to massively parallel genotyping in a quick and simple experiment, which will be of great value in a wide range of genetic studies.

Table 1. Concordance, MI consistency and call rates

α	CallRate (%)	Concordance (%)	MI consistency (%)
0.50	99.88	99.659	99.921
0.40	99.66	99.716	99.949
0.30	99.41	99.746	99.961
0.25	99.16	99.763	99.968
0.20	98.84	99.775	99.972
0.15	98.52	99.783	99.975
0.10	97.60	99.896	99.978
0.075	97.11	99.799	99.980
0.05	95.91	99.807	99.982
0.025	93.79	99.814	99.985
0.01	87.86	99.824	99.989

For a given α, column CallRate lists average call rate over 90 CEPH family samples, column MapMap lists concordance with HapMap genotypes based on all 90 CEPH family samples, and column MI lists Mendelian inheritance consistency based on 10 trios with 5 trios among the 90 CEPH family samples.

ACKNOWLEDGEMENTS

We are grateful to Manqiu Cao, Wenwei Chen, Amy He, Matthew Ho, Saima Kassam, Jane Law, Howard Lee, Weiwei Liu, Halina Loi, Gregory Marcus, Michael Mittman, Carsten Rosenow, Mei-mei Shen, Sean Walsh and Jane Zhang for valuable discussions and/or providing data. We thank the referees for their valuable comments and suggestions.

REFERENCES

Brooks,A.J. (1999) The essence of SNPs. *Gene*, **234**, 177–186.

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

Dong,S. *et al*. (2001) Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res.*, **11**, 1418–1424.

Fan,J. *et al*. (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.*, **10**, 853–860.

Fodor,S.P. *et al*. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.

Fodor,S.P. *et al*. (1993) Multiplexed biochemical assays with biological chips. *Nature*, **364**, 555–556.

HapMap (2003) The international hapmap consortium. *Nature*, **426**, 789–796.

Hollander,M. and Wolfe,D. (1999) *Nonparametric Statistical Methods*, 2nd edn. Wiley, NY.

Hubbell,E., Webster,T. and Matsuzaki,H. (2004) Quickly choosing choice snps for chips. Poster presentation. In Proceedings of RECOMB 2004, San Diego, CA.

Kennedy,G. *et al*. (2003) Large scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233–1237.

Koed,K. *et al*. (2005) High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. *Cancer Res.*, **65**, 34–45.

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

Liu,W.-M. *et al*. (2001) Rank-based algorithms for analysis of microarrays. *Proc. SPIE*, **4266**, 56–57.

Matsuzaki,H. *et al*. (2004a) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.

Matsuzaki,H. *et al*. (2004b) Parallel genotyping of over 10,000 SNPs using a one primer assay on a high density oligonucleotide array. *Genome Res.*, **14**, 414–425.

Middleton,F.A. *et al*. (2004) Genomewide Linkage Analysis of Bipolar Disorder by Use of a High-Density Single-Nucleotide–Polymorphism (SNP) Genotyping Assay: A Comparison with Microsatellite Marker Assays and Finding of Significant Linkage to Chromosome 6q22. *Am. J. Hum. Genet.*, **74**, 886–897.

Pease,A.C. *et al*. (1993) Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proc. Natl. Acad. Sci. USA*, **91**, 555–556.

Puffenberger,E.G. *et al*. (2004) Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYLloss of function. *PNAS*, **101**, 11689–11694.

Ranch,A. *et al*. (2004) Molecular karyotyping using an SNP array for genomewide genotyping. *J. Med. Genet.*, **41**, 916–922.

Sellick,G.S. *et al*. (2003) A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13. *Diabetes*, **52**, 2636–2638.

Shrimpton,A.E. *et al*. (2004) A HOX Gene Mutation in a Family with Isolated Congenital Vertical Talus and Charcot-Marie-Tooth Disease. *Am. J. Hum. Genet.*, **75**, 92–96.

Woods,C.G. *et al*. (2004) A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J. Med. Genet.*, **41**, e101.