*Published online 9 August 2006* *Nucleic Acids Research, 2006, Vol. 34, No. 14* ***e100***
*doi:10.1093/nar/gkl475*

# Genotyping and annotation of Affymetrix SNP arrays

**Philippe Lamy[1], Claus L. Andersen[2], Friedrik P. Wikman[2] and Carsten Wiuf[1,2,*]**

[1]Bioinformatics Research Center, University of Aarhus, Høegh-Guldbergsgade 10, Bldg 1090,
8000 Aarhus C, Denmark and [2]Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby,
Brendstrupgaardsvej 100, 8200 Aarhus N, Denmark

Received February 13, 2006; Revised June 20, 2006; Accepted June 21, 2006

## ABSTRACT

**In this paper we develop a new method for genotyping Affymetrix single nucleotide polymorphism (SNP) array. The method is based on (i) using multiple arrays at the same time to determine the genotypes and (ii) a model that relates intensities of individual SNPs to each other. The latter point allows us to annotate SNPs that have poor performance, either because of poor experimental conditions or because for one of the alleles the probes do not behave in a dose–response manner. Generally, our method agrees well with a method developed by Affymetrix. When both methods make a call they agree in 99.25% (using standard settings) of the cases, using a sample of 113 Affymetrix 10k SNP arrays. In the majority of cases where the two methods disagree, our method makes a genotype call, whereas the method by Affymetrix makes a no call, i.e. the genotype of the SNP is not determined. By visualization it is indicated that our method is likely to be correct in majority of these cases. In addition, we demonstrate that our method produces more SNPs that are in concordance with Hardy–Weinberg equilibrium than the method by Affymetrix. Finally, we have validated our method on HapMap data and shown that the performance of our method is comparable to other methods.**

## INTRODUCTION

To date, whole-genome scans of polymorphic genetic markers [e.g. single nucleotide polymorphisms (SNPs)] are routinely performed with high-throughput technologies such as Affymetrix SNP array technology. Genome scans provide comprehensive information about the genetic background of individuals and have been used among other things to (i) study linkage disequilibrium in human populations and populations of other species, (ii) perform association mapping and linkage studies of common complex diseases, and (iii) conduct analysis of the genetic content in tumor cells, where the assumption of diploidy, as found in normal cells often is violated.

Affymetrix SNP arrays have become popular and are widely used. Originally, an array with 1500 SNPs was released, later the 10k SNP array followed and quite recently arrays with up to 500k SNPs have been made available. The array technique is based on genomic hybridization to synthetic high-density oligonucleotide microarrays [see (1) and references therein]. Each of the two alleles of an SNP is represented by 10 or 14 oligonucleotides (together called a probe set) and hybridization (probe) intensities are measured for all probes in a probe set. Affymetrix has developed a software (GDAS) for genotyping SNPs based on the intensities and, subsequently, the derived genotypes can be used for further analysis of the data [such as (i–iii)].

GDAS genotypes SNPs arraywise, one SNP at a time. For the larger arrays Affymetrix has developed a new dynamic model-based algorithm (DM) that is also based on arraywise genotyping (2). Here we present an alternative algorithm (PBG— pool-based genotyping) for genotyping Affymetrix SNP arrays. If allele intensities (probe intensities combined into one value for each allele) (Materials and Methods) are plotted for a typical SNP, three distinct clusters are generated that correspond well to the three possible genotypes (Figure 1). Naturally, this suggests that the genotype of a SNP could be derived from the distribution of allele intensities by choosing the genotype of the cloud that statistically (in some sense) is closest to the observed allele intensities. PBG builds on this observation.

In addition, we base PBG on a model that allows identification and annotation of SNPs that either are difficult to genotype correctly for experimental reasons or have probes that are not suited for copy number analysis. Identification of chromosomal regions with abnormal copy numbers (i.e. deviations from two copies) is an important undertaking in cancer research (3). Potentially, these regions harbor oncogenes or other genomic elements that are involved in the progression of tumors. Exclusion of SNPs that are not suited for copy number analysis is thus likely to increase the power to infer copy numbers correctly. We return to these issues in Results and Discussion.

At the time of writing this paper we became aware of another method (RLMM) that takes an approach similar to

[*]To whom correspondence should be addressed. Tel: +45 8942 3155; Fax: +45 8942 3077; Email: wiuf@birc.au.dk.

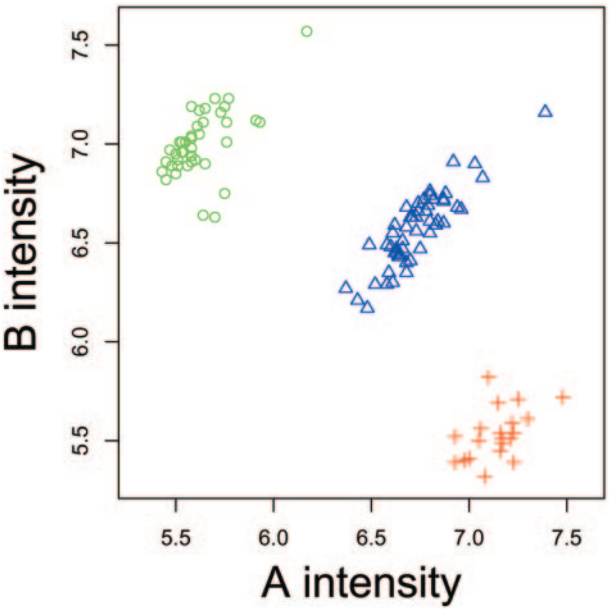

**Figure 1.** Genotype clusters. Plotted is the allele intensities for a typical SNP. Blue triangles denote the heterozygous genotype, red crosses and green circles the two homozygous genotypes. In this case PBG and GDAS agree.

genotyping as we do, although it is not based on a model that relates the intensities of different SNPs to each other (4). Also this approach will be taken up in Results and Discussion.

## MATERIALS AND METHODS

### 10k Early Access Array

For this study we used 113 samples collected at Aarhus University Hospital, Skejby. The GeneChip® Mapping 10k Early Access Array was applied to all 113 samples. The Single Primer Assay Protocol (labeling, hybridization, washing, staining and scanning) was performed according to the manufacturer's instructions (Affymetrix, Santa Clara, CA, USA) (1). Of the 113 samples, 32 are from unrelated Danish individuals, 40 from Cuban unrelated individuals and 41 from Cuban families. The Early Access Array has 10 126 SNPs. Of these 9600 (9430 autosomal and 170 X chromosomal) mapped to on a unique position in the genome [using the April 2003 genome assembly (hg15), http://www.genome.ucsc.edu]. The remaining SNPs (526) were excluded from further analysis. Genotypes were derived using (unnormalized) probe set intensities with Genechip DNA Analysis Software (GDAS) from Affymetrix. Subsequently, the probe set intensities were normalized using the dChipSNP software (6) and the allele intensities as defined in equation (1) were calculated.

### HapMap data

To evaluate PBG on an externally validated dataset, we used a dataset where both HapMap calls and Affymetrix calls are available. We downloaded HapMap genotype data from 30 CEPH trios (90 samples in total) from http://www.hapmap.org/downloads/index.html.en and Affymetrix genotype data from the same samples from http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx. There are 15 589 SNPs for which both calls exist in the 90 samples. All SNPs are on Affymetrix Xba array and genotyped with DM.

### Notation and definitions

Let $\alpha$ denote an arbitrary allele, $\alpha = \mathrm{A}$ or B, and let $\bar{\alpha}$ denote the complementary allele of $\alpha$, i.e. if $\alpha = \mathrm{A}$ then $\bar{\alpha} = \mathrm{B}$, and if $\alpha = \mathrm{B}$ then $\bar{\alpha} = \mathrm{A}$. Further, let $\gamma$ denote an arbitrary genotype, $\gamma = \mathrm{AA}$, AB, BB, AY or BY. Genotypes AY and BY denote male genotypes for X chromosome SNPs. Also let $\alpha\alpha$ denote the homozygote genotype for the $\alpha$ allele.

The probe intensities are combined into two values by taking the logarithm of the average over all probes for the $\alpha$ allele, $\alpha = \mathrm{A}$ or B, i.e.

$$I_{ij}(\alpha) = \log\left(\frac{1}{p}\sum_{k=1}^{p}\mathrm{PM}_{ijk}(\alpha)\right), \qquad 1$$

where $\mathrm{PM}_{ijk}(\alpha)$ is the intensity of the $k$-th probe of allele $\alpha$ for SNP $j$ in array $i$. Here $k$ runs over $k = 1, \ldots, p$, where $p = 10$ or 14, $i = 1, \ldots, 113$, $j = 1, \ldots, 9600$ and PM is short for perfect match (1). We do not use the mismatch probes in this approach. We use dChipSNP normalized probe intensities because they appear to have better statistical properties than the unnormalized probe intensities [cf. (5,6)]. The values $I_{ij}(\mathrm{A})$ and $I_{ij}(\mathrm{B})$ are referred to as allele intensities, or the A- and B-intensity, respectively.

Further, for SNP $j$, let $M_j(\alpha \mid \gamma)$ be the empirical average of $\alpha$-intensities for samples with genotype $\gamma$. For example, for a SNP on the X chromosome, $M_j(\mathrm{A} \mid \mathrm{AY})$ is the average of the A-intensity in male samples with genotype AY and for an autosomal SNP, $M_j(\mathrm{B} \mid \mathrm{AA})$ is the average of the B-intensity in samples with genotype AA.

### Allelic cross hybridization

The genotypes derived by GDAS were used for this part of the study. To investigate cross hybridization we focused on SNPs on the X chromosome (170 SNPs) and compared the allele intensities found in the male samples to those found in the female samples. For each SNP on the X chromosome, the samples were divided into groups according to genotype (excluding no calls, i.e. SNPs where a genotype call has not been assigned). Only groups comprising at least five samples were included in the analysis. The average of the $\alpha$-intensity was calculated in each group and a straight line was fitted to the points $(M_j(\alpha \mid \alpha\mathrm{Y}), M_j(\alpha \mid \mathrm{AB}))$, $\alpha = \mathrm{A}, \mathrm{B}, j = 1, \ldots, 170$ (Notation and Definition). Note that each SNP contributes at most two points. One for the A allele ($\alpha = \mathrm{A}$), if there are more than five male samples with genotype AY and five female samples with genotype AB. Similarly, one for the B allele ($\alpha = \mathrm{B}$), if there are more than five male samples with genotype BY and five female samples with genotype AB.

### The model

We assume the following model for the 9430 autosomal SNPs. The theoretical expectation of the intensity $I_{ij}$ $(\alpha \mid \gamma)$

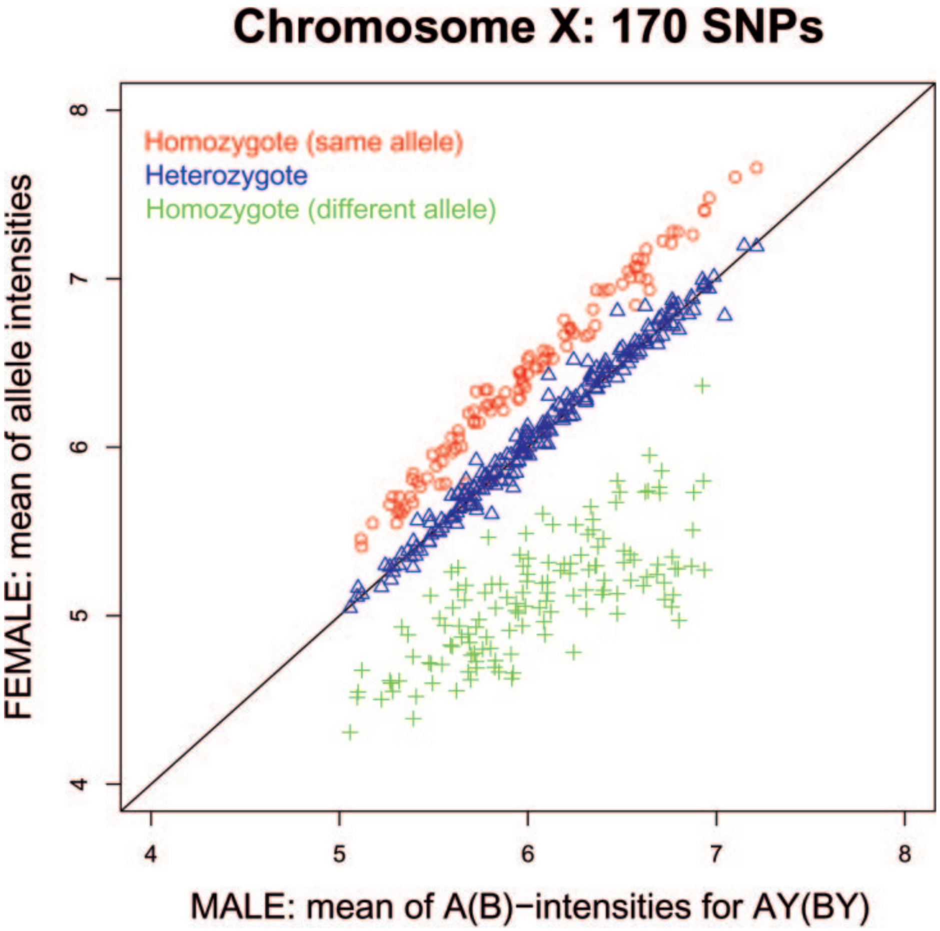

**Figure 2.** Mean intensities. Plotted is the mean allele intensities for female samples against the mean intensities for male samples. Only cases for those at least five samples have a given genotype are included. Intensity of the *x*-axis: $M_j$(A|AY) and $M_j$(B|BY). Intensity on the *y*-axis depends on the color. Red circles, $M_j$(A|AA) and $M_j$(B|BB); blue triangles, $M_j$(A|AB) and $M_j$(B|AB); Green crosses, $M_j$(B|AA) and $M_j$(A|BB).

is denoted $\mu_j(\alpha\,|\,\gamma)$, $j = 1, \ldots, 9430$, and the following relations are assumed:

$$\mu_j(\alpha \mid \alpha\alpha) = c_1 + c_2\mu_j(\alpha \mid \text{AB}), \qquad 2$$

$$\mu_j(\alpha \mid \bar{\alpha}\bar{\alpha}) = d_1 + d_2\mu_j(\alpha \mid \text{AB}) + d_3\mu_j(\alpha \mid \text{AB})^2, \qquad 3$$

where $c_1$, $c_2$, $d_1$, $d_2$ and $d_3$ are unknown parameters. The model is motivated by initial plotting of the empirical means $M_j$ $(\alpha\,|\,\gamma)$ (Figure 2). Including a quadratic term in Equation 2 does not improve the model; the term is statistically indifferent from zero ($p \approx 0$). The model postulates that the α-intensity of the heterozygous genotype is related to the α-intensity of the homozygous genotypes and further, that this relation is 'global' in the sense that the parameters $c_1$, $c_2$, $d_1$, $d_2$ and $d_3$ are not SNP specific, but the same for all SNPs.

Note that the labeling of alleles by A and B are arbitrary such that the set of A alleles is not expected to behave differently or have different chemical properties than the set of B alleles. Similarly, for the probe sets. The unknown parameters in Equations 2 and 3 should therefore not depend on α. This observation also has the consequence that the covariance matrix of ($I_{ij}$(A), $I_{ij}$(B)) for samples with genotype αα does not depend on α.

We further assume that the covariance matrices are independent of the means, i.e. are constant—by testing this does not appear to be strictly true; in particular, the variance of $I_{ij}(\bar{\alpha})$ for samples with genotype αα depends somewhat on $\mu_j(\bar{\alpha} \mid \alpha\alpha)$ (data not shown). This provides an additional five parameters. The covariance matrix

$$\Sigma_{\text{hom}} = \begin{pmatrix} \sigma^2_{\text{hom}} & \tau_{\text{hom}} \\ \tau_{\text{hom}} & \bar{\sigma}^2_{\text{hom}} \end{pmatrix} \qquad 4$$

for samples with homozygous genotypes, where $\sigma^2_{\text{hom}}$ is the variance of the α-intensity and $\bar{\sigma}^2_{\text{hom}}$ the variance of the $\bar{\alpha}$-intensity, and the covariance matrix

$$\Sigma_{\text{het}} = \begin{pmatrix} \sigma^2_{\text{het}} & \tau_{\text{het}} \\ \tau_{\text{het}} & \sigma^2_{\text{het}} \end{pmatrix} \qquad 5$$

for samples with the heterozygous genotype. In total there are 10 + number of SNPs = 9440 parameters in the model – five regression parameters, five covariance parameters and 9430 mean value parameters.

### Parameter fitting

An iterative procedure is used to estimate the model parameters. Assume that an assignment of genotypes to SNPs is given (Genotyping). For each SNP the samples were divided into three groups according to genotype (excluding no calls). The empirical mean $M_j(\alpha \mid \gamma)$ was calculated for each $\alpha$ and genotype with at least five sample points, all other cases were excluded. In the following, superscript $k$ indicates that the estimates are the estimates after the $k$-th, $k \geqslant 0$, iteration.

- *Initialization*. Define $\hat{\mu}_j^0(\alpha \mid \gamma) = M_j(\alpha \mid \gamma)$ and let $\hat{\Sigma}^0_{\text{hom}}$ and $\hat{\Sigma}^0_{\text{het}}$ be the empirical covariances with $\mu_j(\alpha \mid \gamma)$ replaced by $\hat{\mu}_j^0(\alpha \mid \gamma)$.
- *Update 1*. Use linear regression to fit a straight line to the points $(\hat{\mu}_j^k(\alpha \mid \text{AB}), \hat{\mu}_j^k(\alpha \mid \alpha\alpha))$, $\alpha = \text{A, B}$, $j = 1, \ldots, 9430$, resulting in two fitted parameters $\hat{c}_1^k$ and $\hat{c}_2^k$. Note that each SNP contributes at most two points.
- *Update 2*. Similarly, fit a second-order polynomial to the points $(\hat{\mu}_j^k(\alpha \mid \text{AB}), \hat{\mu}_j^k(\alpha \mid \bar{\alpha}\bar{\alpha}))$, $\alpha = \text{A, B}$, $j = 1, \ldots, 9430$, to obtain three estimated parameters $\hat{d}_1^k$, $\hat{d}_2^k$ and $\hat{d}_3^k$. Again each SNP contributes at most two points.
- *Update 3*. Use weighted least square to re-estimate the parameters $\mu_j(\alpha \mid \text{AB})$ and $\mu_j(\alpha \mid \alpha\alpha)$ assuming the relationship (2) and weights $\hat{\Sigma}^k_{\text{hom}}$ and $\hat{\Sigma}^k_{\text{het}}$. (Only the relevant entries in the covariance matrices are used.) The re-estimated parameters are denoted as $\hat{\mu}_j^{k+1}(\alpha \mid \text{AB})$ and $\hat{\mu}_j^{k+1}(\alpha \mid \alpha\alpha)$.
- *Update 4*. Use least square to re-estimate the parameter $\mu_j(\alpha \mid \bar{\alpha}\bar{\alpha})$ assuming the relationship (3). The re-estimated parameter is denoted $\hat{\mu}_j^{k+1}(\alpha \mid \bar{\alpha}\bar{\alpha})$.
- *Update 5*. Re-estimate the covariance matrices with $\mu_j(\alpha \mid \gamma)$ replaced by $\hat{\mu}_j^{k+1}(\alpha \mid \gamma)$. Only intensities for which $\mu_j(\alpha \mid \gamma)$ is estimated are included.
- *Iteration step*. Repeat Updates 1–5 a number of times. Here for three times.

Updates 3 and 4 are two separate steps rather than just one step. If all parameters are estimated in one step using least square, $I_j(\alpha \mid \bar{\alpha}\bar{\alpha})$ tends to dominate the least square equation with the consequence that the estimated values become less accurate (data not shown).

### Genotyping

The procedure for genotyping is iterative starting with an initial clustering for each SNP of the points $(I_{ij}(\text{A}), I_{ij}(\text{B}))$ for all samples into at most four clusters corresponding to the genotypes AA, AB, BB and NC (no call—both GDAS and our method genotype an SNP as NC if the confidence in all proper genotypes are low). For this study GDAS genotyping was used as the initial clustering; $\gamma_{ij}^0$ denotes the GDAS genotype of array $i$, SNP $j$. The procedure is continued until no more (few) changes in genotypes occur. For the $k$-th ($k > 0$) iteration the following is performed.

- *Parameter estimation*. Estimate the parameters $\mu_j(\alpha \mid \gamma)$ and the two covariance matrices (as described in Parameter Fitting) using all observations for which the estimated genotype $\gamma_{ij}^{k-1}$ has confidence higher than $C > 0$ in iteration step $k - 1$. If $k = 1$ all proper GDAS genotypes (only excluding no calls) are used.
- *Calculation of confidence*. Denote the estimated densities for genotype $\gamma$ by $f_k(x, y \mid \gamma)$ in iteration $k$. Calculate the weight for genotype $\gamma$ using the following equation:

  $$W_k(x, y \mid \gamma) = \frac{f_k(x, y \mid \gamma)}{f_k(x, y \mid \text{AA}) + f_k(x, y \mid \text{AB}) + f_k(x, y \mid \text{BB}) + \epsilon},$$

  where $(x,y) = (I_{ij}(\text{A}), I_{ij}(\text{B}))$ and $\epsilon > 0$ is a constant. If none of the genotypes provides good support for $(x,y)$, i.e. if $f(x,y \mid \gamma) \ll \varepsilon$, then all genotypes get low confidence. Here $\epsilon = 10^{-10}$ was used.
- *Genotyping*. If $\max_\gamma W(x,y \mid \gamma) > C$, then genotype $\gamma_{ij}^{(k)} = \text{argmax}_\gamma W(x, y \mid \gamma)$ is assigned to the SNP, otherwise NC is assigned.
- *Iteration step*. Repeat the three steps a number of times. Here for six times.

### SNP performance measures

Affymetrix provides a list of SNPs that were excluded/replaced in the commercial version of the 10k SNP array. It comprises 998 SNPs out of the 9430 autosomal SNPs that are used in this study. The reasons for excluding the SNPs include low call rate, low confidence, poor reproducability and visual criteria. The SNPs in the list were compared to the SNPs found by the SNP performance measures described below. The measures are designed to identify SNPs that are difficult to genotype correctly or to flag SNPs or alleles, where the probes of one or both alleles do not show a dose–response behavior, as postulated by the model. The flagged SNPs and alleles might not be suitable for copy number analysis.

*Hardy–weinberg equilibrium*. For all SNPs it was tested whether the genotype assignments complied with Hardy–Weinberg equilibrium. To avoid issues of imbreeding and admixture, two groups of arrays were defined. (i) A group of unrelated Danish individuals (32 arrays) and (ii) a group of unrelated Cuban individuals (40 arrays). For each SNP ($j$), the total number of genotypes ($n_j$; not including no calls) and the numbers of A ($a_j$) and B ($b_j$) alleles were calculated. SNPs where all arrays in a group are homozygous for the same allele are excluded (they trivially comply with Hardy–Weinberg equilibrium). Subsequently, a permutation test was conducted where the $a_j$ and $b_j$ alleles randomly were re-distributed among the $n_j$ individuals. It was counted how often a value higher than the observed value of the chi-squared statistic was obtained in the permuted samples. A total of 1000 permutations were performed.

*Distance measure*. A weighted Euclidian distance was calculated between the observed means $M_j(\alpha \mid \gamma)$ and the estimated means $\hat{\mu}_j(\alpha \mid \gamma)$. This was performed for the A- and the B-intensities alone and jointly for both. A probe set that does not perform according to expectations is likely to have a higher distance value than a probe set that does perform according to expectations. Thus, a large distance indicates that the observed intensities do not fit well to the model.

For the $\alpha$ probe, the following distance was calculated:

$$\frac{1}{\hat{\mu}_j(\alpha \mid \text{AB})^2} \sum_\gamma D(\alpha \mid \gamma)^2 \hat{\sigma}(\alpha \mid \gamma)^{-2},$$

where $D(\alpha \mid \gamma) = M_j(\alpha \mid \gamma) - \hat{\mu}_j(\alpha \mid \gamma)$, if $\hat{\mu}_j(\alpha \mid \gamma)$ lies between $M_j(\mathrm{A} \mid \gamma)$ and $M_j(\mathrm{B} \mid \gamma)$; and otherwise $D(\alpha \mid \gamma) = 0$. Generally, SNPs can be genotyped correctly if $\hat{\mu}_j(\alpha \mid \gamma)$ lies outside the interval spanned by $M_j(\mathrm{A} \mid \gamma)$ and $M_j(\mathrm{B} \mid \gamma)$. Here $\hat{\sigma}(\alpha \mid \gamma)$ is the average of the squared residuals $M_j(\alpha \mid \gamma) - \hat{\mu}_j(\alpha \mid \gamma)$. The factor $\hat{\mu}_j(\alpha \mid \mathrm{AB})$ is motivated by the model; it roughly scales intensities for different SNPs to the same range. Thus, distances become comparable between SNPs.

Alleles were flagged if the distance was >0.15. SNPs were flagged if the A and B distances both were >0.15.

## RESULTS

### Allelic cross hybridization

The genotypes derived by GDAS were used for this part of the study. Figure 2 shows the relationship between the mean of the allele intensities for the male and the female samples for all 170 SNPs on the X chromosome.

To investigate cross hybridization we focused on SNPs on the X chromosome (170 SNPs) and compared the allele intensities found in the male samples to those found in the female samples. Specifically, we compared the α-intensity for samples with genotype AB (females) to the α-intensity for samples with genotype αY (males). The curve of $M_j(\alpha \mid \mathrm{AB}) - M_j(\alpha \mid \alpha\mathrm{Y})$ plotted against $M_j(\alpha \mid \mathrm{AB})$ is not statistically different from the constant line with intercept 0 ($P = 0.71$). For definition of $M_j(\alpha \mid \gamma)$ see Notation and definitions. It is concluded that the presence of the B allele for genotype AB (in females) does not affect hybridization of the A allele and vice versa. Also, the curve $M_j(\alpha \mid \bar{\alpha}\bar{\alpha}) - M_j(\alpha \mid \bar{\alpha}\mathrm{Y})$ plotted against $M_j(\alpha \mid \bar{\alpha}\bar{\alpha})$ is not statistically different from the constant line with intercept 0 ($P = 0.32$). In consequence, hybridization of the A allele is not affected by the copy number of the B allele (1 or 2) and vice versa.

Thus, it appears that the effect of allelic cross hybridization generally is insignificant. This observation does not have immediate consequences for our genotyping method but will have consequences for copy number analysis. It will be taken up in Discussion.

### Genotyping

Initially, we selected all arrays for which the arraywise residuals after iteration 1 of parameter fitting (Parameter fitting) were <0.27 (Supplementary Figure S1). A total of 10 arrays were excluded in this way, leaving 103 arrays. The 10 arrays were used for testing. We performed the genotyping method on the 103 arrays as described in Parameter fitting and Genotyping. Subsequently, the 10 arrays were genotyped with the parameters estimated from the 103 arrays, and thus served as an independent test of the method.

In majority of cases PBG agrees with GDAS and only one round of iteration is necessary for PBG to stabilize. In some cases PBG improves with the number of iterations because the starting point (GDAS genotypes) is inaccurate and/or the variation in the data requires extra iterations before stabilization occurs. Figure 3 shows examples of SNPs genotyped with GDAS and PBG after five rounds of iteration and confidence level $C = 0.90$. This level of confidence gives fewer no calls than GDAS (PBG 1%, GDAS 6.6%) (Table 1) but we have found from studying plots of the estimated genotypes that this level of confidence appears reasonable. With confidence $C = 0.998$ the same number of NCs are made (∼6.5%) but the two methods only agree on ≈26% (=1.3/5.0) of these.

Table 2 shows how often the two methods agree on genotype when a call has been made for different levels of confidence. A mere 392 SNPs (not listed in the table) in 113 arrays (total of 1 064 686 SNPs) were homozygous AA (BB) with PBG and BB (AA) with GDAS when using confidence 0.90. This number was reduced to 263 SNPs when using confidence 0.998.

The examples in Figure 3 illustrate some of the differences between PBG and GDAS. In many cases where PBG outperforms GDAS, GDAS does not provide a call to one cluster of points or to a group of points located within a cluster. This is not just a question of the level of confidence. Even with a higher level of confidence, e.g. $C = 0.95$, PBG is able to genotype these SNPs (data not shown).

There are also few cases where GDAS outperforms PBG—one case is also shown in Figure 3. In other cases, there is one cluster and PBG might also fail here. The presence of just one cluster is either (i) because one allele has low population frequency and by chance is not represented in the sample, or (ii) because of poor experimental conditions that make it statistically impossible to distinguish the genotypes from one another. In these cases PBG often provides an overrepresentation of heterozygous genotypes. For 31 SNPs, 60% or more (excluding NC) are heterozygous. In a panmixing population 50% is the maximum theoretical heterozygosity level, and an overrepresentation of heterozygous genotypes is thus in conflict with expectations and results in strong violation of Hardy–Weinberg equilibrium. In contrast, for the 31 SNPs, GDAS shows a scatter of different genotypes—whether GDAS genotyping is correct in these cases requires experimental verification.

For 440 SNPs we found 4 or more differences between PBG and GDAS, only counting cases where both methods make a proper call (i.e. excluding case where one or both methods assign NC). The group of 440 SNPs is referred to as Group B, the remaining SNPs as Group A.

It is difficult to get exact numbers for when one method performs better on a particular SNP than the other, because we do not know the true genotypes. A manual expection indicates that for ∼150 SNPs PBG does a better job than GDAS, and for ∼40 SNPs the opposite is true. These all appear in Group B and ∼90% of the cases, where we believe GDAS is superior, are spotted by the perfomance distance measure introduced in the following section.

Despite PBG is based on a model which assumes that the intensity raises with the number of alleles present, PBG successfully genotypes SNPs where one probe does not perform according to the model. To illustrate this point we genotyped the 170 SNPs on chromosome X using only female samples. Subsequently, the male samples were genotyped (see Table 3).

### SNP performance measures

Individual SNPs that are not in Hardy–Weinberg equilibrium after genotyping are likely to be falsely genotyped. Thus,

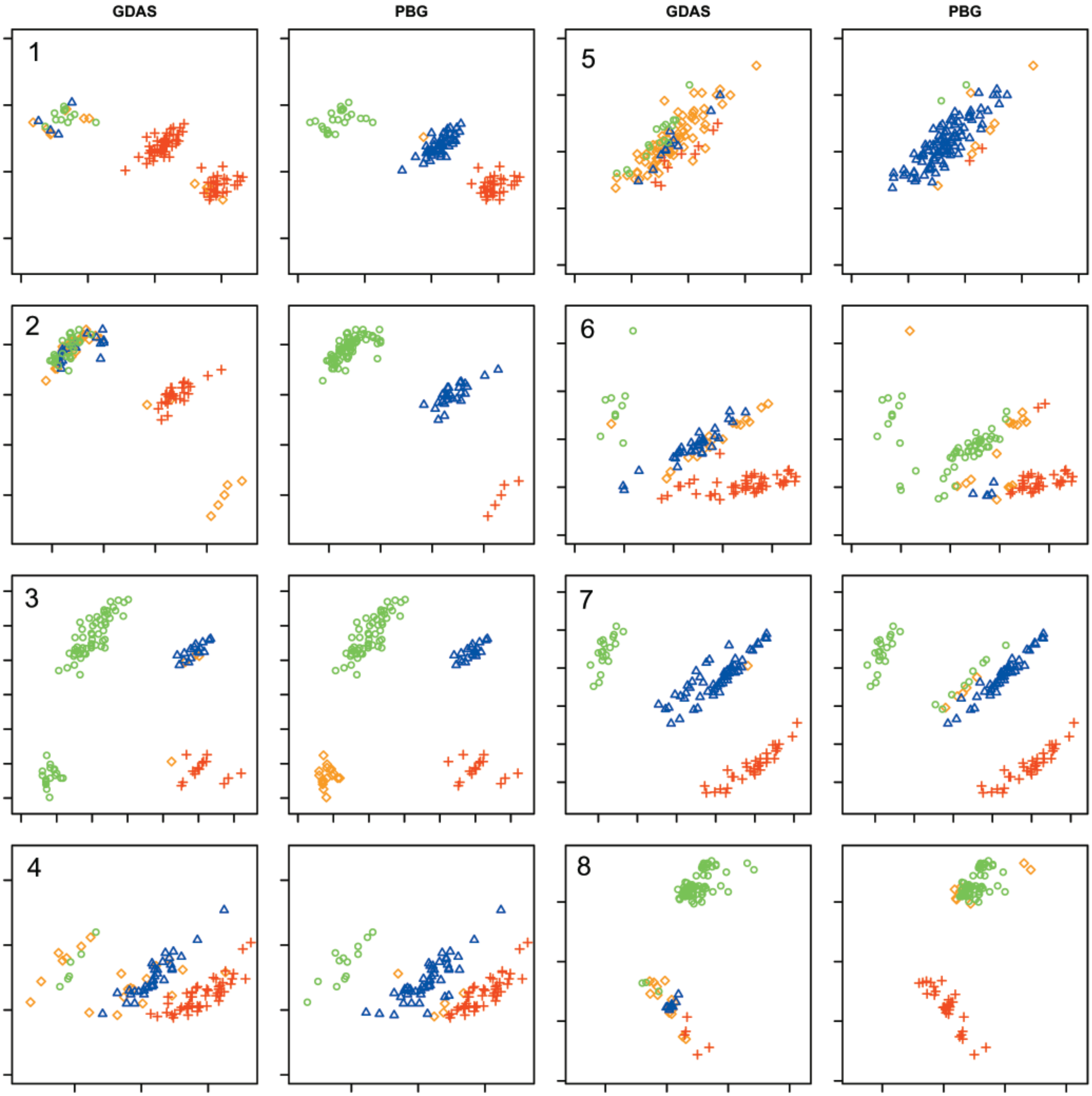

**Figure 3.** Genotyping examples—PBG versus GDAS. Shown are eight SNPs genotyped with GDAS and PBG, respectively. Red crosses: Homozygous for AA; blue triangles, heterozygous; green circles, homozygous for BB; yellow rhombuses, no call. SNPs 1–3 show cases where PBG outperforms GDAS; SNP 4 shows an example where the A allele does not reflect the number of A alleles in the genotype, but still GDAS and PBG genotype correctly; SNPs 5 and 8 show cases where none of the probes apparently functions correctly; and SNPs 6–7 show cases where GDAS outperforms PBG.

failure to pass a test for Hardy–Weinberg equilibrium is an indication of poor SNP quality. If Hardy–Weinberg equilibrium generally is not fulfilled it indicates poor performance of the method.

We performed a permutation test for Hardy–Weinberg equilibrium for all SNPs in two populations: the samples of unrelated Danish individuals (32 arrays) and the sample of unrelated Cuban individuals (40 arrays). Table 4 summarizes the findings. SNPs were excluded if all samples were homozygous for the same genotype or were no calls. These SNPs trivially comply with Hardy–Weinberg equilibrium. In general, PBG provides more genotypes that are in concordance with Hardy–Weinberg equilibrium and with statistical expectations.

An overview of the results of the performance measures are collected in Table 5. Lists of SNPs being selected by the performance measures are provided in Supplementary Table S1.

**Table 1.** PBG versus GDAS

| Conf. | PBG | GDAS | | | | | |
|---|---|---|---|---|---|---|---|
| | | 103 Arrays | | 10 Arrays | | 113 Arrays | |
| | | Call | NC | Call | NC | Call | NC |
| 0.9 | Call | 93.1 | 6.0 | 88.6 | 8.8 | 92.7 | 6.3 |
| | NC | 0.6 | 0.3 | 1.7 | 0.9 | 0.7 | 0.3 |
| 0.95 | Call | 92.8 | 5.9 | 87.9 | 8.5 | 92.4 | 6.2 |
| | NC | 0.8 | 0.4 | 2.4 | 1.3 | 1.0 | 0.5 |
| 0.99 | Call | 91.6 | 5.6 | 85.6 | 7.6 | 91.1 | 5.7 |
| | NC | 2.1 | 0.8 | 4.6 | 2.2 | 2.3 | 0.9 |
| 0.998 | Call | 89.0 | 5.0 | 82.5 | 6.5 | 88.4 | 5.1 |
| | NC | 4.7 | 1.3 | 7.8 | 3.3 | 5.0 | 1.5 |

Shown is how often PBG and GDAS make a genotype call and an NC using different confidence levels for PBG. Standard settings were applied for GDAS. The additional 10 arrays were genotyped using the parameters fitted when genotyping the 103 arrays. In total 970 466 SNPs in 103 samples were available for genotyping, and 94 220 in the remaining 10 arrays.

**Table 2.** Percentage agreement (%-agrm) between PBG and GDAS

| Conf. | 103 Arrays | | 10 Arrays | | 113 Arrays | |
|---|---|---|---|---|---|---|
| | %-agrm | NC | %-agrm | NC | %-agrm | NC |
| 0.9 | 99.25 | 0.8 | 98.98 | 2.6 | 99.23 | 1.0 |
| 0.95 | 99.30 | 1.2 | 99.09 | 3.6 | 99.28 | 1.4 |
| 0.99 | 99.41 | 2.8 | 99.31 | 6.8 | 99.40 | 3.2 |
| 0.998 | 99.52 | 6.1 | 99.46 | 11.0 | 99.51 | 6.5 |

Shown is how often the two methods agree when both methods make a call, and the percentage of no calls obtained with PBG.

**Table 3.** Genotyping of SNPs on chromosome X

| Conf. | Female | | Male | | All | |
|---|---|---|---|---|---|---|
| | %-agrm | NC | %-agrm | NC | %-agrm | NC |
| 0.9 | 99.39 | 0.70 | 98.32 | 1.80 | 98.92 | 1.20 |
| 0.95 | 99.41 | 1.00 | 98.54 | 2.51 | 99.02 | 1.68 |
| 0.99 | 99.59 | 2.41 | 99.07 | 3.82 | 99.36 | 3.05 |
| 0.998 | 99.65 | 4.53 | 99.25 | 4.84 | 99.47 | 4.67 |

Out of 113 samples, 62 are females and 51 males. For comparison, GDAS produces 5.15% NC in females, 7.29% in males and 6.11% in total.

**Table 4.** Test for Hardy–Weinberg equilibrium

| | Group A | | Group B | |
|---|---|---|---|---|
| | Mean | Var | Mean | Var |
| PBG | 0.50 | 0.081 | 0.37 | 0.096 |
| GDAS | 0.48 | 0.084 | 0.34 | 0.094 |

Group A is defined as SNPs where PBG and GDAS disagree on the genotype (excluding NC) in <4 cases, and Group B (440 SNPs) is defined as those where there are ⩾4 disagreements. The mean and variance are expected to follow a uniform distribution, i.e. the mean should be 0.50 and the variance 0.083. Both methods have problems with Group B. The GDAS mean 0.48 is significantly different from 0.50 ($P < 0.001$).

**Table 5.** Comparison of different performance measures

| | Distance | | NC | | HW | | |
|---|---|---|---|---|---|---|---|
| | One | Both | 5% | 10% | 0.1% | 1% | Group B |
| Total | 485 | 34 | 408 | 125 | 57 | 384 | 440 |
| Rejected | 158 | 14 | 185 | 70 | 24 | 107 | 191 |

The list of rejected SNPs comprises 998 SNPs. NC 10% (5%) is the group of SNPs with a no call rate of at least 10% (5%). If an SNP obtains a $P$-value <1% (0.1%) in the test for Hardy–Weinberg equilibrium in either of the two populations it counts in HW 1% (0.1%). Generally, we find an overrepresentation of SNPs in the list of rejected SNPs compared with the list of non-rejected SNPs.

**Table 6.** Comparison with other methods on HapMap data

| No. of SNPs | PBG | DM | No.of SNPS | RLMM |
|---|---|---|---|---|
| 15 589 | 99.50% | 99.60% | 15 910 | ? |
| 14 509 | 99.57% | 99.65% | 11 446 | 99.86% |

Shown is the percentage agreement with HapMap calls for different methods. PBG and DM are run on the same data set, RLMM on a different, although similar, dataset (see text). The SNPs in the second row form a subset of the SNPs in the first row. SNPs are excluded if they fulfill the criteria that there is at most one member in two genotype groups (based on HapMap calls). Results are not available for RLMM on the full dataset.

Generally, we find an overrepresentation of SNPs in the list of rejected SNPs of the Affymetrix compared to the list of non-rejected SNPs. The distance measure was calculated for each allele intensity and jointly for both, as described in Materials and Methods. Plots of all SNPs with a distance >0.15 is shown in Supplementary Figure S2. The distance measure is adequate to identify SNPs that are difficult to genotype, i.e. SNPs with experimentally poor performance, or where the probes for one or both alleles are not reacting in a dose–reponse manner, i.e. probes that are not suitable for copy number analysis.

Even if a combination of measures are used, the frequency of flagged SNPs in the list of rejected SNPs of the Affymetrix does not exceed 32% (out of 998). To achieve this, SNPs are flagged if the number of NC exceeds 5% (NC 5% in Table 5) or the $P$-value for the test for Hardy–Weinberg equilibrium is <1% (HW 1% in Table 5). Oppositely, the frequency of rejected SNPs out of all flagged SNPs does not exceed 36% (combining distance with NC 5% and HW 0.1%). A table is provided in Supplementary Table S2.

## Comparison with other methods on HapMap data

In the previous sections we have evaluated PBG on a large dataset and compared PBG to GDAS. To evaluate PBG on an externally validated dataset, we followed the procedure in (4) closely. This procedure also allows us to compare PBG with DM (2) and RLMM (4).

We selected 15 589 SNPs from Affymetrix Xba array where both HapMap and DM calls were available (Materials and Methods). We ran PBG and DM on this dataset. Unfortunately, we could not get RLMM to run on our computers and we therefore used results from (4) to compare with RLMM. These results are based on 15 910 SNPs selected in the same way as our dataset. The discrepancy between the sizes of the two datasets is unknown to us. For each SNP (in both datasets) calls are made for 90 individuals. Table 6 summarizes the results.

Interestingly, PBG genotypes are identical to HapMap genotypes in all 90 samples for 81.4% of the SNPs in the full dataset and for 87.7% of the 1080 SNPs excluded by the criteria used in (4) (see also Table 6). The criteria excludes SNPs if there is at most one member in two genotype groups (based on HapMap calls). This shows one strength of PBG, because it is able to genotype accurately, even though some genotypes are sparsely represented in the data. It is not shown in (4) how RLMM performs on the excluded set of SNPs.

It is also worth pointing out that RLMM used HapMap calls to train the algorithm, whereas PBG used Affymetrix calls (inferred by DM). Affymetrix calls are always available for Affymetrix arrays, whereas HapMap calls naturally are not. Thus, training with HapMap calls is not generally possible. Generally, this might lead to lower performance of RLMM than reported in Table 6, because HapMap calls are believed to be highly accurate.

## DISCUSSION

We have developed a new method for genotyping Affymetrix SNP arrays and compared the performance of our method (PBG) to that of Affymetrix (GDAS). PBG is based on analyzing multiple arrays at the same time, in contrast to GDAS that analyses SNPs arraywise, one SNP at a time. Generally, the two methods agree, but PGB appears to be able to genotype correctly with a lower no call rate and also appears to produce more genotypes than GDAS that comply with Hardy–Weinberg equilibrium. In addition, PBG is based on a model that relates allele intensities from different SNPs to each other. We use this relationship to annotate SNPs and alleles. The plots provided in Supplementary Figure S2 show that we are able to annotate poor performing SNPs and alleles. We also compared PBG to two other recently published methods, DM and RLMM. Overall the methods seems to have similar performances; some of the differences are explained below.

Our method is based on dChipSNP normalized probe intensities. One array is selected as reference array and all other arrays are normalized relatively to the reference array. This has the advantage that new arrays (a test set) can be genotyped using fitted parameters obtained from a training set. If the test set is normalized relatively to the reference array of the training set the fitted parameters of the training set can be used to genotype the test set. Particularly, this should be useful when genotyping only few arrays, provided the fitted parameters of the test set and the reference array is publically available. We showed that this approach is feasible by analyzing an additional 10 arrays that was not used for fitting (Table 1).

Our model has 10+ number of SNPs = 9440 parameters. For some SNPs only one or two genotypes are observed. In these cases, we use the model to estimate the mean intensity of the non-observed genotypes. In contrast, the model RLMM proposed in (4) has 15 × number of SNPs = 139 450 parameters (if applied to the 10k array), because their model does not assume a relationship between parameters for different SNPs. If a genotype is not observed or sparsely represented, the parameters for that genotype are predicted using estimated parameters from other SNPs. Note that it is not known how RLMM performs on SNPs where only one genotype is present (or some genotypes are sparsely represented). In (4) results are not shown for these SNPs, even though they comprise ∼28.7% of the SNPs in their dataset.

Naturally, the structure of the data can be modeled more accurately with a large dimensional parameter than a small dimensional parameter (in the sense that 9440 is small compared to 139 450). PBG is thus likely to fail in genotyping some SNPs that might be correctly genotyped by RLMM. However, since these SNPs do not fit the model, PBG will flag them as ‘poor’ and they can be excluded from the analysis. Flagging or annotation of ‘poor’ performing SNPs offers a two-sided advantage. First of all, SNPs that perform ‘poor’ because of experimental reasons can be excluded from the analysis. Second, SNPs can be ‘poor’ performing, as illustrated in Figure 3 and Supplementary Figure S1, because for one or both alleles the probes do not behave in a dose-response manner and should therefore be excluded. These SNPs might still be genotyped correctly, but are not suitable for copy number analysis. Several research groups have demonstrated that a typical SNP shows a linear relationship between the log-copy number and the log-intensity and used the intensity levels in diploid samples to infer copy numbers in abnormal samples, e.g. in tumor samples for instance see (3,6,8,9). This relationship is documented both with the data normalization procedure introduced by Affymetrix and with dChipSNP’s procedure, which is used in PBG.

Analysis of SNP arrays often requires correction for multiple testing. To avoid too many false positives the significance level of a single test should be chosen low. Excluding SNPs that are poorly performing because of experimental reasons should reduce the number of false positives and thus increase the power.

It appears that GDAS genotypes tumor samples reliably at the cost of an increased no call rate are compared to normal samples. Our initial investigations show that PBG seems to make more errors while genotyping tumor samples (data not shown). This is expected because we explicitly apply a model which assumes that two copies of the DNA are present for each SNP, whereas a copy number of two is often found violated in tumor samples. Whether, the method in (4) can genotype samples with abnormal DNA content correctly is presently unknown.

In (3,6,8,9), genotyping and copy number analysis are separate issues; i.e. if genotypes are used in a copy number analysis the genotypes are obtained before the copy number analysis is conducted. It would be natural to combine the two into a single analysis. We showed in Allele Cross Hybridization that the level of the A-intensity is not affected by the copy number of the B allele, and vice versa. This leads us to speculate that cross hybridization can be ignored generally in the sense that the level of the A-intensity is only affected by the copy number of the A allele, not the copy number of the B allele. Assuming a linear relationship between log-copy number and log-intensity, the intensity levels for higher allele copy numbers could be extrapolated from the observations made in this paper.

A version of PBG implemented in Perl is available from the authors upon request.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

P.L. and C.W. is supported by the Danish Cancer Society and by the Fraenkel Foundation, Denmark. C.L.A. is supported by the Danish Research Councils and the John and Birte Meyer Foundation. Ole Mors, Center for Basic Psychiatric Research, is thanked for providing 81 of the SNP arrays. Funding to pay the open Access publication charges for this article was provided by the Fraenkel Foundation, Denmark.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Kennedy,G.C., Matsuzaki,H., Dong,S., Liu,W.M., Huang,J., Liu,G., Su,X., Cao,M., Chen,W., Zhang,J. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol*., **21**, 1233–1237.
2. Di,X., Matsuzaki,H., Webster,T.A., Hubbell,E., Liu,G., Dong,S., Bartell,D., Huang,J., Chiles,R., Yang,G. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.
3. Bignell,G.R., Huang,J., Greshock,J., Watt,S., Butler,A., West,S., Grigorova,M., Jones,K.W., Wei,W., Stratton,M.R. *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*., **14**, 287–295.
4. Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.
5. Li,C., Tseng,G.C. and Wong,W.H. (2003) Model-based analysis of oligonucleotide arrays and issues in cDNA microarray analysis. In Speed,T. (ed.), *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall, NY, pp. 1–34.
6. Zhao,X., Li,C., Paez,J.G., Chin,K., Jänne,P.A., Chen,T.-H., Girard,L., Minna,J., Christiani,D., Leo,C. *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorpism arrays. *Cancer Res*., **64**, 3060–3071.
7. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
8. Huang,J., Wei,W., Zhang,J., Liu,G., Bignell,G.R., Stratton,M.R., Futreal,P.A., Wooster,R., Jones,K.W. and Shapero,M.H. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics*, **4**, 287–299.
9. Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*., **65**, 6071–6079.
10. Ming,L., Wei,L.-J., Sellars,L.R., Lieberfarb,M., Wong,W.H. and Li,C. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.