Algorithms and analysis of human disease genomics

Inouye, M.

2

A GENOTYPE CALLING ALGORITHM FOR THE ILLUMINA BEADARRAY PLATFORM

Yik Y. Teo[1,2,3], Michael Inouye[2,3], Kerrin S. Small[1,2], Rhian Gwilliam[2], Panagiotis Deloukas[2], Dominic P. Kwiatkowski[1,2], Taane G. Clark[1,2]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom

[2] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom

[3] These authors contributed equally

Bioinformatics. 2007 Oct 15;23(20):2741-6

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 20 2007, pages 2741–2746
doi:10.1093/bioinformatics/btm443

Genetics and population analysis

A genotype calling algorithm for the Illumina BeadArray platform

Yik Y. Teo[1,2,*,†], Michael Inouye[2,†], Kerrin S. Small[1,2], Rhian Gwilliam[2], Panagiotis Deloukas[2], Dominic P. Kwiatkowski[1,2] and Taane G. Clark[1,2]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN and [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Received on July 31, 2007; revised and accepted on August 20, 2007

Advance Access publication September 10, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Large-scale genotyping relies on the use of unsupervised automated calling algorithms to assign genotypes to hybridization data. A number of such calling algorithms have been recently established for the Affymetrix GeneChip genotyping technology. Here, we present a fast and accurate genotype calling algorithm for the Illumina BeadArray genotyping platforms. As the technology moves towards assaying millions of genetic polymorphisms simultaneously, there is a need for an integrated and easy-to-use software for calling genotypes.

Results: We have introduced a model-based genotype calling algorithm which does not rely on having prior training data or require computationally intensive procedures. The algorithm can assign genotypes to hybridization data from thousands of individuals simultaneously and pools information across multiple individuals to improve the calling. The method can accommodate variations in hybridization intensities which result in dramatic shifts of the position of the genotype clouds by identifying the optimal coordinates to initialize the algorithm. By incorporating the process of perturbation analysis, we can obtain a quality metric measuring the stability of the assigned genotype calls. We show that this quality metric can be used to identify SNPs with low call rates and accuracy.

Availability: The C++ executable for the algorithm described here is available by request from the authors.

Contact: teo@well.ox.ac.uk or tgc@well.ox.ac.uk

1 INTRODUCTION

The possibility of genome-wide studies hinges on advances in genotyping technology to perform large-scale genotyping quickly and cheaply, assaying up to a million single nucleotide polymorphisms (SNPs) simultaneously. In this setting, it is difficult and time consuming to determine genotypes manually from the examination of fluorescent dye intensities representing the presence or absence of alleles and automated genotype calling procedures are necessary for such large-scale genotyping.

A number of genotype calling algorithms have been established recently for the Affymetrix GeneChip and ParAllele Molecular Inversion Probe genotyping technology (Affymetrix Inc., 2006; Di *et al*., 2005; Moorhead *et al*., 2006; Plagnol *et al*., 2007; Rabbee and Speed, 2006; The Wellcome Trust Case Control Consortium, 2007; Xiao *et al*., 2007). There has been a confluence in the style and input of the calling algorithms towards the use of multi-component mixture models on hybridization intensities. These intensities are typically normalized and transformed to coordinates which yield distinct genotype clouds that are easier to call.

Generally speaking, the end-products of the genotyping process yield hybridization intensities for the alleles and genotypes are assigned by comparing the relative strength of these intensities. While the genotype of a SNP for each sample can be assigned independently, it has been shown that pooling information across multiple samples at each SNP can enhance the calling process and result in higher quality calls (Affymetrix Inc., 2006). However, this requires non-biological differences of intensities to be minimized. Various normalization schemes have been explored and implemented although these schemes are generally unique to each genotyping technology (Bolstad *et al*., 2003; Carvalho *et al*., 2007; Rabbee and Speed, 2006; Xiao *et al*., 2007). These normalized signal intensities are often transformed to different coordinate scales for ease of calling, and the contrast-strength transformation has been the preferred scale for a number of algorithms (Affymetrix Inc., 2006; Moorhead *et al*., 2006; Plagnol *et al*., 2007). The strength and contrast can be respectively interpreted as the equivalent of r and θ in polar coordinates for the allelic signals (see Fig. 1).

While suitably normalized intensities generally result in genotype clusters with similar location characteristics across the SNPs, assigning genotypes for hundreds of thousands of SNPs means that a calling algorithm has to be suitably flexible to variations in the signal intensities. For most of the genotyping technologies, a fraction of the SNPs can have genotype clusters that are significantly shifted away from the expected positions (see Fig. 2). This can potentially result in erroneous genotype calls when the calling algorithm is unable to account for such extreme shifts.

In an independent pursuit of a calling algorithm for Illumina data, we formalize a calling strategy which is similar to the approach by Moorhead *et al*. (2006) and Plagnol *et al*. (2007),

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

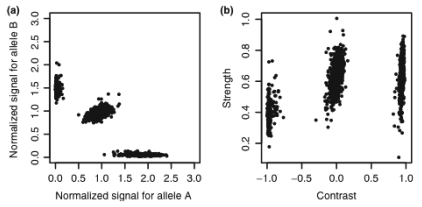

Fig. 1. A typical clusterplot of the allelic hybridization signals for a SNP: (**a**) after normalization; (**b**) after transformation of the same data to yield the contrast-scale coordinates. Each point in the figures represent the intensity data for an individual.

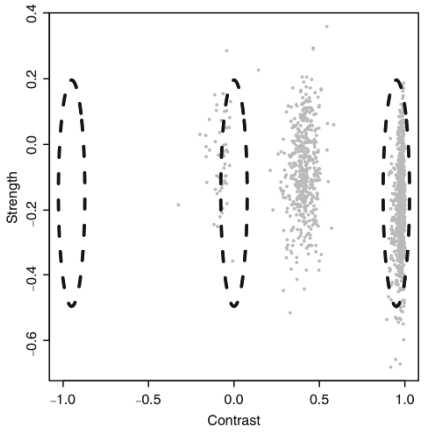

Fig. 2. Clusterplot for a SNP with shifted genotype clusters. Points in grey represent the observed signal data and the black ellipses represent the expected positions of the three genotype clouds.

and has a number of features which are specifically designed for the BeadArray genotyping technology. This strategy is set within an Expectation-Maximization framework, which is extremely fast without compromising on the quality of the genotype calls. We also explored the use of perturbation analysis to quantify the stability of the genotype calls, and provide a metric for assessing the quality of the assigned genotypes for each SNP as a whole. We emphasize the difference between the quality of the assigned genotype for each individual at a SNP, versus the quality of the assigned genotypes for each SNP across all the individuals. Our algorithm also has the ability to accommodate noisier data from whole-genome amplified DNA. It has been implemented in a C++ program Illuminus and is available by request from the authors.

2 METHODS

2.1 Chip design

Here, we provide a description of the chip design for the Illumina HumanHap 550 K SNP microarray. Other Illumina HumanHap microarrays (e.g. the 650 K and 1 M versions) which assay different number of SNPs are of similar technology and construction (Gunderson *et al.*, 2006; Steemers *et al.*, 2007). Each microarray consists of lateral stripes, each of which contains a beadpool of 55 000 different beadtypes. Every beadtype assays a single SNP and is represented by 20 beads on average. Each of these beads accommodates locus-specific 50-mer probes which correspond to the nucleotide sequence directly adjacent to the SNP. The single base extension (SBE) biochemistry format allows each beadtype to assay both SNP alleles, thus potentially providing 20 allele measurements per SNP on average for each DNA sample. Visualization of bead hybridization is achieved through the addition of hapten-labelled ddNTPs. If the labelled ddNTP corresponds to the complement of the assayed SNP, it will be incorporated by the extending polymerase and, following a staining step, the resultant fluorescent wavelength and intensity is measure by a scanner.

2.2 Normalization

Each microarray is divided into a number of sub-bead pools (25 sub-bead pools for the 550 K chip) and normalization of the bead intensities occurs at the sub-beadpool level. The normalization algorithm uses a six-degree of freedom affine transformation which occurs in five steps (Kermani, 2005):

(1) outlier removal
(2) background estimation
(3) rotational estimation
(4) shear estimation
(5) scaling estimation.

A brief description of the algorithm is as follows: within each sub-beadpool, outlier SNPs are removed if their allelic intensities are smaller than either the 5th smallest or 1st percentile as compared to all SNPs, or if their intensities are larger than the 5th largest or 99th percentile as compared to all SNPs. Background estimation occurs by uniform sampling of 400 points along each intensity axis to create a linear fitting to candidate homozygotes. The intercept of the linear fittings from both homozygotes then defines the origin. Rotation and shear of the data points by the same uniform sampling then occurs with respect to this defined origin. The final normalized intensities are then determined by mean scaling via virtual control points. This procedure at present occurs automatically within the Illumina BeadStudio software and outputs the normalized intensities, which provide a pair of coordinates corresponding to the signals for the two alleles at each SNP.

2.3 Manual and automated Illumina calling

The automated GenCall proprietary algorithm which Illumina provides with BeadStudio was initiated within the BeadStudio analysis on all Illumina 550 K SNPs. These genotypes are herein referred to as 'GenCall' genotypes. Manual curation of the Illumina SNPs was initiated on the GenCall clustered genotypes and are herein called 'GenCall-C' genotypes. Manual inspection and adjustment of the genotype classifications was performed on all SNPs with: (a) call rates $<95.0\%$ at a GC score cutoff of 0.20; (b) call rates $>95.0\%$ with cluster separation scores <0.25 or average GC score <0.60; (c) Hardy–Weinberg equilibrium χ^2 P-values <0.0001. In addition, all mitochondrial and SNPs located on the X or Y chromosomes were inspected.

2.4 Mixture model

Let (x_{jl}, y_{jl}) denote the normalized signal intensities for the two alleles which we generically define as A and B, respectively, for sample j at SNP l. We define the contrast and strength respectively as

$$c_{jl} = \frac{x_{jl} - y_{jl}}{x_{jl} + y_{jl}} \quad (1)$$

$$s_{jl} = \log(x_{jl} + y_{jl}) \quad (2)$$

We fit a three-component bivariate mixture model for $\boldsymbol{X}_{jl} = (c_{jl}, s_{jl})$ using multivariate truncated t distributions, where the three components correspond to the genotype classes of AA, AB and BB, respectively. Let $f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denote the density function for data $\boldsymbol{x}$ at a t distribution with location parameter $\boldsymbol{\mu}$, variance–covariance matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom. The density for $\boldsymbol{X}_{jl}$ can be written as

$$F(\mathbf{X}_{jl}) = \sum_{k=1}^{3} \lambda_k \phi_k(\mathbf{X}_{jl}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) \quad (3)$$

where $(\lambda_1, \lambda_2, \lambda_3)$ corresponds to the mixture proportions obtained by assuming Hardy–Weinberg equilibrium, and

$$\phi_1(\mathbf{X}_{jl}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \nu_1) = \frac{f(\mathbf{X}_{jl}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \nu_1)}{1 - \int_{-\infty}^{-1} f(\mathbf{X}_{jl}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \nu_1) dc} \quad (4)$$

$$\phi_2(\mathbf{X}_{jl}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu_2) = \frac{f(\mathbf{X}_{jl}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu_2)}{\int_{-1}^{1} f(\mathbf{X}_{jl}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu_2) dc} \quad (5)$$

$$\phi_3(\mathbf{X}_{jl}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \nu_3) = \frac{f(\mathbf{X}_{jl}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \nu_3)}{1 - \int_{1}^{\infty} f(\mathbf{X}_{jl}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \nu_3) dc}. \quad (6)$$

We also introduce a fourth bivariate Gaussian component with zero covariance and significantly large variances such that the density is flat across the possible range of values. This serves as an outlier class for samples with intensity profiles which do not clearly belong to any of the three valid genotypes.

The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the data, while ν_k are pre-determined. As the Illumina BeadArray technology yields extremely sharp and uniform signals for the homozygotes as compared to the heterozygotes at bulk of the SNPs, the variance profiles for the distributions of the contrast for the AA and BB genotype clusters are significantly peaked (see Fig. 3). This can potentially bias the genotype calling since any homozygote samples with contrast intensities which are marginally different from the uniform signals may be assigned heterozygous genotypes. The use of t distributions means we can model this feature with heavier-tails for the homozygous clusters as compared to the heterozygous cluster. In practice, this means we fit $\nu_1 = \nu_3 < \nu_2$ for most SNPs, and where the variance profiles for the contrast of the three genotype clusters are similar, we fit a constant degree of freedom for all three components.

An Expectation-Maximization procedure is implemented which alternates between recalibrating the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and λ_k using maximum-likelihood estimation conditional on the assigned genotypes (the M-step), and reassigning the genotypes to the intensity data conditional on the recalibrated cluster characteristics (the E-step). We iterate between the process of recalibration and reassignment until the reassignment yields exactly the same genotype configuration as that of the previous iteration. At every reassignment step, we introduce an additional step which reflects the natural decision-making process of assigning genotypes. Samples with contrast values that are larger than the mean contrast of the heterozygous cluster will never be assigned an AA genotype, while samples with contrast values that are smaller than the mean contrast of the heterozygous cluster will never be assigned a BB genotype. The genotype for the j sample can be assigned to the genotype class with the maximum posterior probability, subject to the constraint that the maximum posterior probability is above some threshold. Samples where the maximum posterior probability is below the threshold will be assigned a NULL genotype.

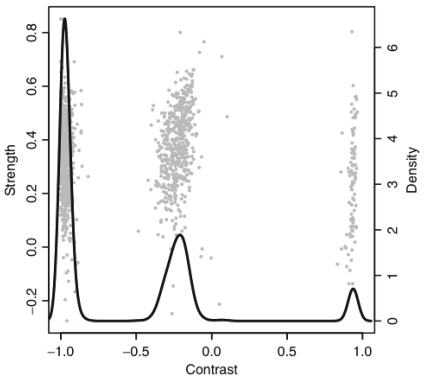

Fig. 3. The clusterplot of a typical SNP on the Illumina array which yields highly homogeneous signals for the homozygous clusters, resulting in significantly peaked variance profiles for the homozygous clusters. Lines in black represent the kernal densities of the observed data (in grey).

Initialization of the EM procedure for each SNP is performed using a one-dimension three-component Gaussian mixture model with equal mixing for the contrast axis, and searches across five guided starts to identify the optimal set of parameters which best fits the data. Of the five starting sets of parameters, two are guided by the data and this modification allows genotype clusters which are shifted to be accommodated accordingly. The spread of the Gaussian distributions are similarly pre-determined for the initialization and are assumed to be identical across all three genotype classes for each of the five guided starts. As with the location parameters, the SDs for the first three starts are pre-determined, and are calculated from the data for the last two starts. A NULL component with a flat density is introduced to accommodate samples with contrast scores which are ambivalent to cluster membership. The EM algorithm is initialized with the set of starting parameters which yields the largest likelihood.

2.5 Chromosome X

For SNPs on chromosome X which are not located in the pseudo-autosomal regions, the genotype calling procedure is modified to incorporate gender information as males contain only one copy of chromosome X and thus will never be heterozygous at these SNPs. As we expect only the strength coordinates to differ between males and females with similar distribution for the contrast coordinates, there is no change to the calling procedure for females, whereas

$$\phi_2(\mathbf{X}_{jl}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu_2, \text{male}) = 0$$

in Equation (5). The mixture proportions utilizing Hardy–Weinberg equilibrium is calculated such that males only contribute one allele copy when evaluating the allele frequencies and females contribute two allele copies as usual.

2.6 Modelling parameters

The parameters used for our implementation of the calling algorithm are listed in detail here:

2.6.1 Initialization The location parameters for the five guided starts are:

- $(-0.9, 0.0, 0.9)$
- $(-0.9, -0.5, 0.9)$
- $(-0.9, 0.5, 0.9)$
- $(-0.9, 0.5\ (\max(\boldsymbol{c}) + \min(\boldsymbol{c})), 0.9\ \max(\boldsymbol{c}))$
- $(0.9\ \min(\boldsymbol{c}), 0.5\ (\max(\boldsymbol{c}) + \min(\boldsymbol{c})), 0.9)$.

The SDs for the three genotype classes are identical for each guided start, and they are:

- 0.1
- 0.1
- 0.1
- 0.05 (max $(\boldsymbol{c})$ + min $(\boldsymbol{c})$)
- 0.05 (max $(\boldsymbol{c})$ + min $(\boldsymbol{c})$).

The location and SD of the NULL component are 0 and 100 000, respectively for genomic DNA, and 0 and 1000, respectively for whole-genome amplified DNA.

2.6.2 EM mixture model The location parameter for genotype class k is updated by:

$$\boldsymbol{\mu}_k = (\bar{c}^k, \bar{s}^k) = \left(\frac{1}{n_k}\sum_j^{n_k} c_{jl}^k, \frac{1}{n_k}\sum_j^{n_k} s_{jl}^k\right),$$

where n_k denote the number of samples assigned to genotype class k, and the superscript refers to the points which have been assigned to genotype class k.

The variance–covariance matrix for genotype class k, $\boldsymbol{\Sigma}_k$, is updated by:

$$\frac{1}{n_k - 1}\begin{pmatrix} \sum_j^{n_k}(c_{jl}^k - \bar{c}^k)^2 & \sum_j^{n_k}(c_{jl}^k - \bar{c}^k)(s_{jl}^k - \bar{s}^k) \\ \sum_j^{n_k}(c_{jl}^k - \bar{c}^k)(s_{jl}^k - \bar{s}^k) & \sum_j^{n_k}(s_{jl}^k - \bar{s}^k)^2 \end{pmatrix}$$

For most of the SNPs, we fit multivariate t distributions with heavier tails for the two homozygous classes. This is represented by $\nu_1 = \nu_3 = 6$, $\nu_2 = 20$. For a fraction of the SNPs where the variance profiles for the contrast of the three genotype clusters are similar, we fit $\nu_1 = \nu_2 = \nu_3 = 20$.

The location and variance–covariance matrix for the NULL component are

$$(0, 0) \quad \text{and} \quad \begin{pmatrix} 100000 & 0 \\ 0 & 100000 \end{pmatrix}$$

respectively for genomic DNA, and

$$(0, 0) \quad \text{and} \quad \begin{pmatrix} 1000 & 0 \\ 0 & 1000 \end{pmatrix}$$

respectively for whole-genome amplified DNA.

2.7 Perturbation analysis

In the process of calling the genotypes for hundreds of thousands of SNPs, it is inevitable that there will be SNPs with high-confidence and yet erroneous calls which will filter through despite stringent thresholds on the maximum posterior probabilities. As it is impossible to manually curate the assigned genotypes for all the SNPs, we implemented a perturbation analysis step, which provides a metric for quantifying the stability of the assigned genotypes to minor perturbations in the normalized intensities (submitted). This works on the remit of comparing the genotype calls obtained from two independent runs of the algorithm using the original and perturbed intensities respectively, where a high rate of discordance between the two runs implies that the genotype assignments are less reliable for the particular SNP. While this means that every SNP needs to be called twice, once on the original intensities (x_{jl}, y_{jl}) and once on the perturbed intensities $(x_{jl} + \epsilon_1, y_{jl} + \epsilon_2)$ where ϵ_1 and ϵ_2 are independent and identically $N(0, 0.05^2)$, the algorithm within the EM framework is extremely fast and the time taken to run the analysis with perturbation analysis is realistic.

3 RESULTS

We compared the performance of Illuminus on genotype data for 1409 samples with genomic DNA which have been genotyped on both the Illumina 550 K and the Affymetrix 500 K GeneChip genotyping arrays. These samples are from the 1958 British Birth Cohort which have been genotyped as part of the Wellcome Trust Case Control Consortium (The Wellcome Trust Case Control Consortium, 2007). Of these data, we identified 82 981 SNPs which overlap on the two chips, and the genotypes for the Affymetrix data are available. Genotypes for the Affymetrix data have been called using the automated program Chiamo (The Wellcome Trust Case Control Consortium, 2007). Four metrics are used in assessing the performance of Illuminus against the genotype calls obtained using the GenCall algorithm from the BeadStudio software (Version 3): (i) call rate, defined as the percentage of valid genotype calls (excluding the NULL calls); (ii) concordance, defined as the percentage of valid calls made by each method which are identical to the genotypes from the Affymetrix GeneChip technology (see Discussion Section); (iii) overall concordance, defined as the percentage of concordant genotype calls out of all possible calls—essentially a product of the concordance and the call rate (iv) filtered SNPs, defined as the number of SNPs which have been dropped from further analysis due to low SNP-wise call rates, defined at 0.95 for samples with genomic DNA and 0.90 for whole-genome amplified samples. GenCall genotypes have been thresholded at a confidence score of 0.20, and any genotype with a score less than 0.2 is assigned as a NULL genotype. The BeadStudio software comes with the flexibility of manually curating the calls made by GenCall. We also assess the performance of Illuminus against GenCall genotypes which have undergone manual curation. We calculated the call and concordance rates for genotypes by Illuminus at thresholds of 0.95. We also run Illuminus with perturbation analysis, flagging and excluding SNPs with more than 5% discordant genotypes between two runs of Illuminus with the original and perturbed intensities. The results are summarized in Table 1.

GenCall yielded high quality genotypes, successfully assigning a valid genotype to 99.597% of the data, with high accuracy as reflected by a concordance of 99.785% with genotypes from a different platform. This yielded an overall concordance of 99.383%. The process of manual curation appeared to be conservative and chose to classify more potentially ambiguous calls as NULL genotypes. This reduced the call rates to 99.419%, but increases the concordance to 99.801%. While

Table 1. Comparison of Illuminus against GenCall on 82 981 SNPs for 1409 samples genotyped using genomic DNA

Method	Call rates (%)	Concordance (%)	Overall concordance (%)	Number of SNPs filtered
GenCall	99.597	99.785	99.383	1644
GenCall-C	99.419	99.801	99.221	636
Illuminus	99.873	99.729	99.603	215
Illuminus-PA	99.896	99.762	99.659	11(+2805)
GenCall-PA	99.675	99.791	99.467	521(+2805)

Numbers in the last column refer to the number of SNPs with per-SNP call rates <95% which would have been excluded from further analyses. GenCall-C refers to genotypes which have undergone manual curation. Illuminus-PA refers to the use of perturbation analysis to exclude SNPs with unstable genotypes. GenCall-PA refers to the genotype calls made by GenCall on the same SNPs which remained after exclusion based on perturbation analysis on Illuminus.

Table 2. Comparison of Illuminus against GenCall on 95 578 SNPs for 252 samples genotyped using whole-genome amplified DNA

Method	Call rates (%)	Concordance (%)	Overall concordance (%)	Number of SNPs filtered
GenCall	82.207	95.867	78.809	41 332
Illuminus	95.374	94.561	90.187	1607
Illuminus-PA	95.598	95.031	90.848	379(+7530)
GenCall-PA	82.958	96.051	79.682	35 697(+7530)

Numbers in the column with the Filtered header refer to the number of SNPs with per-SNP call rates <90% which would have been excluded from further analyses. A lower threshold is used because of whole-genome amplification. Illuminus-PA refers to the use of perturbation analysis to exclude SNPs with unstable genotypes. GenCall-PA refers to the genotype calls made by GenCall on the same SNPs which remained after exclusion based on perturbation analysis on Illuminus.

this reduces the overall concordance to 99.221%, manual curation can increase the accuracy of the calls made, and also reduce the number of SNPs with <90% call rate. At a threshold of 0.95, Illuminus made around 0.28% more calls, and this translated to 322 699 genotypes which were assigned by Illuminus but failed to be assigned by GenCall. The rate of concordance for Illuminus was marginally lower but overall Illuminus achieved a much higher concordance rate of 99.603%. Perturbation analysis identified 2805 SNPs to be removed. These SNPs have mean call and concordance rates of 99.302 and 98.787%, and removing these SNPs from the analysis resulted in higher call and concordance rates of 99.896 and 99.762% for the remaining SNPs. This suggests that perturbation analysis can correctly identify SNPs with higher rates of erroneous calls. In order to provide a fair comparison of GenCall and Illuminus, we also investigated the performance of GenCall on the same set of remaining SNPs after exclusion based on perturbation analysis on Illuminus. Our analysis showed that removing SNPs identified by perturbation analysis also improved the performance of GenCall, suggesting that the SNPs removed were performing poorly. SNP-wise call rates have often been used as a criterion for filtering SNPs (Gudmundsson *et al*., 2007; Rioux *et al*, 2007; Saxena *et al*, 2007; Scott *et al*, 2007; The Wellcome Trust Case Control Consortium, 2007; Yeager *et al*., 2007), and a common approach is to remove SNPs with <95% call rates. The use of GenCall resulted in the loss of at least 636 SNPs, while Illuminus achieved a better performance by removing only 215 SNPs.

We also investigated the performance of Illuminus on 252 samples which have been genotyped on both the Illumina 650 K and the Affymetrix 500K GeneChip genotyping arrays with whole-genome amplified DNA (Table 2). We analysed 95 578 SNPs found on both platforms, and the genotypes for the Affymetrix data has similarly been called using Chiamo. Perturbation analysis with a discordance threshold of 5% identified 7530 SNPs to be removed. These SNPs have mean call and concordance rates from Illuminus of 92.624 and 88.819%, and removing these SNPs from the analysis resulted in higher call and concordance rates for the remaining SNPs. We observed a drop in the performance of both GenCall and Illuminus for whole-genome amplified samples. GenCall has a lower call rate of 82%, but calls that were made generally are highly concordant with Chiamo calls. Illuminus performed significantly better with a much higher call rate of 95% while achieving similar concordance rates as GenCall. Overall, this means Illuminus made at least 2.5 million more concordant calls across 252 samples, corresponding to 12% more concordant calls on top of GenCall. While it appears that the numbers are considerably low, we caution against over-interpreting these figures since the performance of both GenCall and Illuminus is necessarily bounded by the error rates of Chiamo and we believe this may be substantial for whole-genome amplified DNA (see Discussion Section). GenCall removed 41 332 SNPs with SNP-wise call rates of <90%. By comparison, Illuminus experienced a more tolerable loss of 1607 SNPs.

4 DISCUSSION

We have introduced a model-based approach to call genotypes for the Illumina BeadArray platforms and this has been implemented within an Expectation-Maximization framework in the program Illuminus. By comparing the performance of Illuminus against GenCall on data where the genotypes for the same SNPs from a different genotyping platform is available, we see that Illuminus made more concordant calls and resulted in a smaller number of SNPs which are excluded on the basis of per-SNP call rates (see below). This improvement over GenCall is significantly more substantial for DNA samples which have undergone whole-genome amplification (see below). We have also investigated the use of perturbation analysis to provide a metric for assessing the stability of the assigned genotypes to minor changes in the allelic signals. Empirical results suggest that perturbation analysis correctly identified SNPs with lower call and concordance rates.

In quantifying the performance of Illuminus, we have compared the assigned genotypes against calls made on the

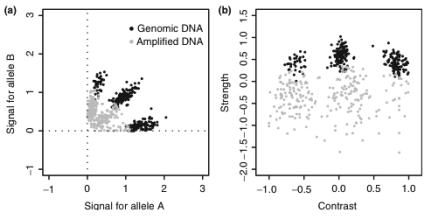

Fig. 4. Clusterplots of a SNP which has been typed on both genomic and whole-genome amplified DNA. Points in black correspond to samples with genomic DNA while points in grey correspond to samples with amplified DNA. The plots are made on the (**a**) normalized allelic signal coordinates; (**b**) strength-contrast transformed coordinates.

same set of SNPs on the Affymetrix 500 K genotyping array. This is clearly not ideal as it assumes that the genotype calls made by Chiamo on the Affymetrix 500 K array are correct. The concordance rates calculated through such comparisons are naturally bounded above by the error rates of Chiamo. We visually inspected the clusterplots for SNPs where the discordance between Chiamo genotypes and GenCall or Illuminus genotypes is >90% and found that all these SNPs have been correctly called using GenCall and Illuminus but the genotypes are inconsistent with the Chiamo genotypes from the Affymetrix data. As we expect such inconsistencies to be more prevalent for whole-genome amplified DNA, we visually verified the Chiamo calls on the Affymetrix data for a random collection of 4000 SNPs and discarded 1481 (37%) SNPs with suspect genotype calls. On the remaining set of 2519 SNPs with ascertained accurate Chiamo calls, the concordance rates of Illuminus and GenCall genotypes increased to 82.769 and 90.509%, respectively. This suggests that while the use of genotypes from an independent platform can provide a measure of comparative performance between GenCall and Illuminus, the concordance rates calculated here tend to underestimate the actual performance.

While whole-genome amplification allows the recovery of DNA samples with inadequate concentration by performing *in vitro* reproduction of quality template DNA, empirical evidence suggests that amplified DNA suffers a reduction in hybridization signals. In the context of genotyping large number of samples simultaneously which is common in a genome-wide association study, this can result in a proportion of SNPs experiencing shifts in the positions of the genotype clusters (see Fig. 4), resulting in indistinct clusters and increasing the uncertainty of a call. While the performance of both GenCall and Illuminus were less optimal for hybridization data from whole-genome amplified DNA, Illuminus comes with a modification for calling data from amplified DNA and this provides a substantial improvement over GenCall.

Genotyping technology is moving towards assaying millions of polymorphisms simultaneously. This requires automated genotyping algorithms to be efficient and accurate. While the use of computationally intensive algorithms may aim to maximize the potential of the hybridization data, they require the use of large computing clusters which the typical laboratory may not have access to. There is a need for an integrated, simple-to-use and yet accurate genotype calling software. The software for the algorithm comes equipped with a metric for identifying SNPs with unstable genotypes. We have presented a fast and accurate calling algorithm which is designed to work with data from the Illumina BeadArray genotyping platforms.

ACKNOWLEDGEMENTS

The work of all the authors is supported by the Wellcome Trust. In addition, Y.Y.T., K.S.S., D.P.K. and T.G.C. also acknowledge support from the Grand Challenges in Global Health initiative (Gates Foundation, Wellcome Trust and FNIH) and the UK Medical Research Council.

Conflict of Interest: none declared.

REFERENCES

Affymetrix Inc (2006) BRLMM: an improved genotype calling method for the GenChip Human Mapping 500K Array Set. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Carvalho,B. *et al.* (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.

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

Gudmundsson,J. *et al.* (2007) Genome-wide association study identifies a second prostate cancer susceptibility variantat 8q24. *Nat. Genet.*, **39**, 631–637.

Gunderson,K.L. *et al.* (2006) Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics*, **7**, 641–648.

Kermani,B.G (2005) Artificial intelligence and global normalization methods for genotyping. US Patent 20060224529.

Moorhead,M. *et al.* (2006) Optimal genotype determination in highly multiplexed SNP data. *Eur. J. Hum. Genet.*, **14**, 207–215.

Plagnol,V. *et al.* (2007) A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.*, **3**, e74.

Rabbee,N. and Speed,T.P (2006) A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, **1**, 7–12.

Rioux,J.D. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy n disease pathogenesis. *Nat. Genet.*, **39**, 596–604.

Saxena,R. *et al.* (2007) Genome-wide association analysis identifies loci for Type 2 diabetes and triglyceride levels. *Science*, **316**, 1331–1336.

Scott,L.J. *et al.* (2007) A genome-wide association study of Type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.

Steemers,F.J. *et al.* (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnol. J.*, **2**, 41–49.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Xiao,Y. *et al.* (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **27**, 1459–1467.

Yeager,M. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.