GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population

Eleni Giannoulatou, Christopher Yau, Stefano Colella, Jiannis Ragoussis, Chrisopher C. Holmes

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 19 2008, pages 2209–2214
doi:10.1093/bioinformatics/btn386

Genetics and population analysis

GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population

Eleni Giannoulatou[1,2,†], Christopher Yau[1,2,†], Stefano Colella[3,‡], Jiannis Ragoussis[3] and Christopher C. Holmes[1,4,*]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, [2]Life Sciences Interface Doctoral Training Centre, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, [3]Genomics Group, Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN and [4]MRC Mammalian Genetics Unit, MRC Harwell, Harwell, OX11 0RD, UK

Received on April 28, 2008; revised on July 22, 2008; accepted on July 23, 2008

Advance Access publication July 24, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Current genotyping algorithms typically call genotypes by clustering allele-specific intensity data on a single nucleotide polymorphism (SNP) by SNP basis. This approach assumes the availability of a large number of control samples that have been sampled on the same array and platform. We have developed a SNP genotyping algorithm for the Illumina Infinium SNP genotyping assay that is entirely *within-sample* and does not require the need for a population of control samples nor parameters derived from such a population. Our algorithm exhibits high concordance with current methods and >99% call accuracy on HapMap samples. The ability to call genotypes using only within-sample information makes the method computationally light and practical for studies involving small sample sizes and provides a valuable independent quality control metric for other population-based approaches.

Availability: http://www.stats.ox.ac.uk/~giannoul/GenoSNP/

Contact: cholmes@stats.ox.ac.uk

1 INTRODUCTION

The success of projects, such as the International HapMap Project (The International HapMap Consortium, 2005) in mapping single nucleotide polymorphisms (SNPs) and the Wellcome Trust Case Control Consortium (2007) in finding associations with common diseases has made SNP genotyping arrays an indispensable tool in genetic epidemiology. Leading manufacturers, such as Affymetrix and Illumina, now offer SNP genotyping arrays that can interrogate over a million SNPs on a single assay on a genome-wide scale with sufficiently high signal-to-noise ratio to enable highly accurate genotype calls to be made with the appropriate statistical genotyping tools.

Current SNP genotyping algorithms, such as BRLMM (Affymetrix Inc., 2006), Birdseed (Affymetrix Inc., 2007) and CHIAMO (Marchini *et al*., manuscript in preparation; Wellcome Trust Case Control Consortium, 2007) for the Affymetrix platform and GenCall (Illumina Inc., 2005) and Illuminus (Teo *et al*., 2007) for the Illumina platforms, typically employ a genotype calling strategy that is reliant on the availability of data from a large collection of individuals. Given the data from a collection of individuals, these genotyping algorithms interrogate each SNP in turn, clustering the allele-specific probe intensities into the three classes representing the three genotypes. The size of the reference population required depends on the minor allele frequency (MAF) of the SNPs of interest. For example, for SNPs where the MAF is <10%, it would be necessary to have a reference population with much more than 100 individuals in order to expect data representing all three genotype classes (AA, AB and BB) at each SNP; for SNPs that have MAF 1% it would be necessary to have a reference population with 10 000 individuals in order to have at least one data point in all three genotype classes; and moreover one would typically require >10 samples in a class in order to estimate the distribution parameters accurately. As high-throughput genotyping techniques are rapidly progressing, more and more SNPs can be genotyped on a single array. In order to increase genomic coverage and probe uniformity across the genome more SNPs with low minor allele frequency are included in these arrays. The size of the reference population required by the current genotyping algorithms will increase for those SNPs with ever-decreasing minor allele frequency.

The motivation for a population-based strategy is that probe intensities vary on a SNP-by-SNP basis. This is due to differences in binding affinities arising from a number of factors, which include probe sequence content, that cause cluster centres and characteristics to vary from SNP to SNP. A variation on the population-based strategy uses model parameters pre-computed from a reference population to derive predictive models that do not necessitate a full re-clustering each time a new test sample is obtained. However, there are some practical limitations to this approach, since model parameters must be recalculated each time the SNP content of a genotyping array is modified or a new genotyping array is produced.

*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.

‡Present address: UMR 203 INRA INSA-Lyon BF2I, Biologie Fonctionnelle, Insectes et Interactions, F-69621 Villeurbanne Cedex, France.

It would be desirable if each probe type on a genotyping array had the exact same response characteristic regardless of which genomic region was being queried. In this scenario, it would be unnecessary to cluster the probe intensities for a population of individuals at each SNP and, instead, the probe intensities for every SNP *within* a single individual could be clustered. This approach would have the advantage of avoiding problems with SNPs with very small minor allele frequencies by allowing information to be borrowed from across SNPs (rather than across individuals) to determine genotype cluster positions. Although probe responses are rarely homogeneous in practice, we have discovered that within-class variation is lower compared to inter-class variation on the Illumina Infinium SNP genotyping array. This enables high-quality SNP genotyping *within* a sample without the need for a reference population. The performance of the method would be independent of the size of the study and it can be used as a quality metric against other available SNP genotyping algorithms. Our method relies on the observation that the inter-class variation can be maximized by accounting for dye-specific and bead-specific effects on the Infinium assay.

2 METHODS

2.1 Data

The Illumina Infinium SNP genotyping array consists of hundreds of thousands of beads. Each bead harbours a set of 50mer oligonucleotide probes designed to hybridize to a specific genomic region that is adjacent to the SNP of interest. Using a two colour single base extension (SBE) chemistry (Steemers *et al*., 2006) each bead is able to assay both SNP alleles. The array has, on average, 20 beads per SNP thus giving 20 pairs of allele-specific intensity measurements per SNP. These are averaged to produce a single summary pair of allele-specific intensity values for each SNP, which we use for clustering. Beads are divided into a number of sets called 'beadpools'. Each beadpool consists of beads that are manufactured at the same time and are physically located at similar positions on the microarray. For each beadpool, we perform quantile normalization (Bolstad *et al*., 2003) in order to correct dye-specific biases due to the two colour system. Clustering is done on the log scale of the intensities $\{\log_2(x+1), \log_2(y+1)\}$ for each beadpool separately.

2.2 Statistical model

Let $\boldsymbol{x}_i = \{\log_2(x_i+1), \log_2(y_i+1)\}$ be the pair of summary log intensities for the i-th SNP. We model the distribution of the probe intensities using a four-component mixture of Student t-distributions which can be described in a hierarchical form,

$$p(z_i|\boldsymbol{\theta}) = \prod_{m=1}^{4} \pi_m^{I(z_i=m)} \tag{1}$$

$$p(u_i|z_i,\boldsymbol{\theta}) = \prod_{m=1}^{4} \mathcal{G}(u_i|\nu_m/2,\nu_m/2)^{I(z_i=m)} \tag{2}$$

$$p(\boldsymbol{x}_i|u_i,z_i,\boldsymbol{\theta}) = \prod_{m=1}^{4} \mathcal{N}(\boldsymbol{x}_i|\mu_m,u_i\Lambda_m)^{I(z_i=m)} \tag{3}$$

where $\mathcal{G}$ denotes the Gamma distribution, $\mathcal{N}$ denotes the Normal distribution, $\boldsymbol{\pi}$ are the mixture proportions, $z_i \in \{1,2,3,4\}$ is an indicator variable for the latent genotype class, $I(\cdot) \in \{0,1\}$ is an indicator function, u_i is a latent scale variable and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mu, \Lambda\}$. We fix $\nu_m = 4$ for $m = 1,2,3,4$. Each mixture component corresponds to either one of the three genotype classes AA, AB and BB or a null class to capture outliers. The hierarchical model exploits the representation of the Student t-distribution as a scaled infinite mixture of normal distributions,

$$\mathcal{S}(\boldsymbol{x};\mu,\Lambda,\nu) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\mu,u\Lambda)\mathcal{G}(u|\nu/2,\nu/2)du \tag{4}$$

with location parameter μ, scale matrix Λ and ν degrees of freedom.

We use conjugate priors throughout to enable fast, analytical integrations. The prior for the mixture weight is given by a Dirichlet distribution

$$p(\boldsymbol{\pi}|\boldsymbol{\kappa}) \propto \prod_{m=1}^{4} \pi_m^{(\kappa_m-1)}, \tag{5}$$

and a normal-Wishart prior used to define the location and scale parameters for each genotype mixture component allowing our model to maintain identifiability

$$p(\mu_m,\Lambda_m) = \mathcal{N}(\mu_m|m_0,\eta_0\Lambda_m)\mathcal{W}(\Lambda_m|\gamma,S_m) \tag{6}$$

where $\mathcal{W}(\Lambda|\gamma,S)$ is the Wishart distribution. The location and scale parameters of the null class are fixed and set to values to make the distribution relatively flat over the feature space.

2.3 Model inference

Two methods for posterior inference are examined. The first approach is a standard Expectation Maximization (EM) algorithm (Dempster *et al*., 1977; Peel and McLachlan, 2000) and the second approach adopts a strategy based on an variational Bayes EM (VB-EM) (Archambeau and Verleysen, 2007; Beal *et al*., 2003) algorithm.

2.3.1 Expectation Maximization The EM algorithm computes a maximum *a posteriori* (MAP) model fit by iteratively computing expectations of the latent parameters and maximizing the expected complete data log-likelihood

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \sum_z \int p(z,\boldsymbol{u}|\boldsymbol{x},\boldsymbol{\theta}^{(i)}) \log p(z,\boldsymbol{u},\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{u}. \tag{7}$$

It can be shown that each successive iteration of the EM algorithm monotonically increases the posterior probability (Dempster *et al*., 1977). If the posterior distribution is unimodal then as $i \to \infty$, $\boldsymbol{\theta} \to \boldsymbol{\theta}_{MAP}$, else $\boldsymbol{\theta}$ tends to a local mode in the posterior distribution. If multimodality is suspected, it is typical to run the EM algorithm several times with random initializations in order to verify that it has not converged to a single local mode. Genotype calls are obtained by finding the genotype with the maximum probability conditional on the MAP parameter estimates, $g_i = \arg\max_g p(z_i = g|\boldsymbol{\theta}^{MAP})$.

2.3.2 Variational Bayes In VB-EM, a variational approximation to the posterior distribution is constructed and optimized using an EM approach so that the Kullback–Leiblier divergence between the true posterior distribution $p(\boldsymbol{\theta},z,\boldsymbol{u}|\boldsymbol{x})$ and the variational approximation $q(\boldsymbol{\theta},z,\boldsymbol{u})$ is minimized

$$KL(q,p) \equiv \int q(\boldsymbol{\theta},z,\boldsymbol{u}) \log \frac{p(\boldsymbol{\theta},z,\boldsymbol{u},\boldsymbol{x})}{q(\boldsymbol{\theta},z,\boldsymbol{u})} d\theta. \tag{8}$$

In order to obtain analytically tractable posterior approximations, the variational posterior is assumed to have a factorized form $q(\boldsymbol{\theta},z,\boldsymbol{u}) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{z,\boldsymbol{u}}(z,\boldsymbol{u})$. The VB-EM steps then consist of the following iterations:

$$q_{z,\boldsymbol{u}}^{(t+1)}(z,\boldsymbol{u}) \propto \exp\left[\int \log p(z,\boldsymbol{u},\boldsymbol{x}|\boldsymbol{\theta}) q_\theta^{(t)}(\theta) d\theta\right] \tag{9}$$

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp\left[\sum_z \int \log p(z,\boldsymbol{u},\boldsymbol{x}|\theta) q_{z,\boldsymbol{u}}^{(t+1)}(z,\boldsymbol{u}) d\boldsymbol{u}\right] \tag{10}$$

The integrals required for (9) and (10) can be calculated analytically for distributions which are members of the conjugate exponential family (Beal *et al*., 2003). The Student t-distributions and conjugate prior distributions used in our models fall into this family. Details of the expressions for the

Table 1. Comparison of call rates and accuracy on 120 HapMap samples genotyped on the HumanHap300Duo BeadChip

Method	Call rate (%)	False calls	No calls	Call accuracy (%)
GenCall	99.799	38 911	73 295	99.694
Illuminus	99.819	89 025	66 199	99.576
GenoSNP	99.660	88 249	124 613	99.419
GenoSNP-VB	100.000[a]	94 380	143	99.742

The null genotypes in HapMap were excluded from the analysis.
[a]To three decimal places.

VB-EM updates are given in Archambeau and Verleysen (2007). Genotype calls are obtained by finding the genotype with the maximum probability based on the variational approximation, $g_i = \arg\max_g q_z(z_i = g)$.

In contrast to the standard EM algorithm, the E-step of the VB-EM approach integrates over the approximate distribution of the model parameters rather than conditioning on its mode. This improves robustness to uncertainty in the model parameters.

2.4 Genotype calling methods

We compared our genotype calling methods, GenoSNP and GenoSNP-VB, which are based on the EM and VB-EM algorithms, respectively, with genotype calls from Illumina's proprietary algorithm, GenCall and Illuminus (Teo *et al.*, 2007) which are both population-based methods. We should note that the default settings were used for both GenCall and Illuminus. The hyper-parameters for GenoSNP-VB were set as follows: $\kappa_0 = 1.1$, $m_0 = [(9, 8, 6, 6); (6, 8, 9, 6)]$, $\eta_0 = 1$, $\gamma_0 = 1$, $S_0 = [(0.1, 0.0); (0.0, 0.1)]$.

3 RESULTS

We tested the performance of GenoSNP and GenoSNP-VB by comparing calls on 120 HapMap samples genotyped on the Illumina HumanHap300Duo genotyping array with genotypes obtained from the International HapMap Project database (The International HapMap Consortium, 2005). In total, 36 630 045 genotypes were available for comparison. The X chromosome was removed from the analysis to avoid any gender bias. Since males contain only one copy of chromosome X they will always be homozygotes for the SNPs on this chromosome. The null genotypes in HapMap were also removed.

GenCall (99.694%) and Illuminus (99.576%) yielded high-quality genotype calls with much >99% accuracy and call rates (Table 1). However, both versions of GenoSNP yielded similar genotyping performance (99.419% and 99.742%, respectively) whilst operating entirely on a within-sample basis. GenoSNP-VB gives slightly better performance suggesting that the use of VB, which accounts for uncertainty in the model parameters, provides more robust inference than GenoSNP's MAP analysis. Figure 1 shows that the genotype probabilities assigned by GenoSNP are also well calibrated with empirical error rates. This is important for downstream analyses such as imputation of genotypes for genome-wide association studies and error rate characterization.

Figure 2 illustrates the reason for the excellent genotyping performance of GenoSNP and GenoSNP-VB. After separating the SNP data into different beadpools, the intensity data is sufficiently well separated in feature space and the three genotype clusters easily discernible. It is, therefore, possible to perform highly accurate unsupervised clustering to generate genotype calls without requiring SNP-by-SNP processing and comparison with a reference population.

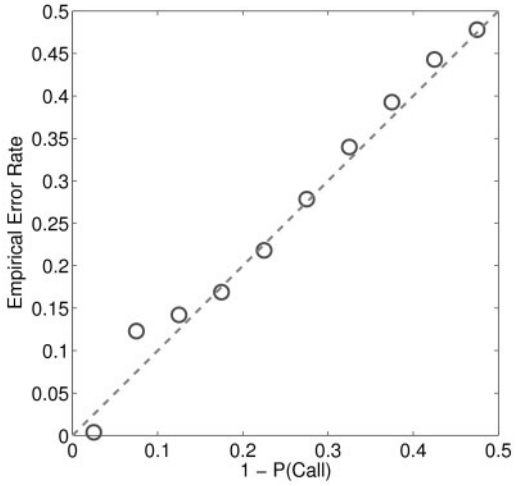

Fig. 1. GenoSNP genotype probabilities are well calibrated with empirical error rates.

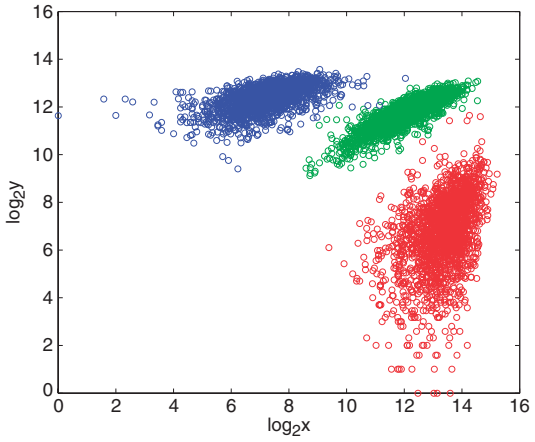

Fig. 2. Log allele-specific intensity plot of all SNPs in bead pool 1 for one HapMap sample. Each data point has been colour labelled using HapMap genotypes (AA - Red, AB - Green, BB - Blue, No Call - Black).

Table 2 presents the call rates and accuracy of the different methods broken down by zygosity. For the population-based methods, GenCall and Illuminus, heterozygotes are generally harder to call than homozygotes. On the other hand, GenoSNP appears to have better performance for the SNPs that are heterozygotes. GenoSNP-VB has high call rate and accuracy for both heterozygotes and homozygotes showing again the robustness of the VB approach.

To further understand the strengths and weaknesses of the available algorithms we compared the errors for all four methods (Table 3). We found that there are 47 675 genotypes that are called incorrectly by GenCall, Illuminus and GenoSNP-VB. Table 4 shows that approximately half of these common errors can be traced back to 1500 SNPs that resulted in an unsuccessful genotype calling in more than 10–15 of the 120 HapMap samples. These errors maybe the result of unusual probe hybridization characteristics or local

Table 2. Breakdown of call rates and accuracy by homozygotes and heterozygotes

	Homozygotes		Heterozygotes	
	Call rate	Call accuracy	Call rate	Call accuracy
GenCall	99.899	99.823	99.596	99.427
Illuminus	99.918	99.763	99.616	99.192
GenoSNP	99.504	99.264	99.981	99.738
GenoSNP-VB	100.000[a]	99.748	100.000	99.729

[a]To three decimal places (%).

Table 3. Breakdown of HapMap genotyping errors by GenCall, Illuminus and GenoSNP-VB

		GenCall		Illuminus	
		True call	False call	True call	False call
GenoSNP	True call	–	58 977	–	103 530
	False call	159 633	53 229	161 168	51 694
GenoSNP-VB	True call	–	61 398	–	105 911
	False call	43 715	50 808[a]	45 210	49 313[a]

[a]47 675 errors were made by all three methods.

Table 4. Breakdown of common genotyping errors made by GenCall, Illuminus and GenoSNP-VB by SNP

Errors per SNP	Number of SNPs	Total errors
≥ 1	23 494	47 675
≥ 2	6942	31 123
≥ 3	3393	24 025
≥ 4	2087	20 107
≥ 5	1466	17 623
≥ 10	533	11 733
≥ 15	307	9114
≥ 20	173	6859
≥ 25	127	5860
≥ 30	98	5079
≥ 50	41	2831
≥ 100	3	327

sequence structures that make these particular SNPs intrinsically difficult to genotype. Alternatively, there maybe errors in the HapMap genotypes (i.e. wrong allele labelling) that has also been noticed before in previous studies (Laframboise *et al*., 2007).

Although a number of genotyping errors were common to GenoSNP-VB, GenCall and Illuminus, the remaining genotyping errors were not common and, instead, genotypes may be called successfully by one approach but not the others. Figure 3 illustrates why GenoSNP successfully calls where Illuminus does not and vice versa. The within-sample approach of GenoSNP allows borrowing of information between SNPs so that accurate genotype cluster allocations for SNPs even with low minor allele frequencies can be achieved. In contrast, Illuminus clusters the population at these particular SNPs, but will encounter problems with model fitting and identifiability as all three genotype clusters may not be represented. GenoSNP is less successful when the hybridization characteristics of a SNP probe are significantly different to the others. Under this situation, the population-based approach of Illuminus provides greater flexibility whereas the within-sample approach of GenoSNP assumes a probe response homogeneity across SNPs which is inappropriate.

In order to assess the consistency of genotype calls, we analysed a sample that was genotyped three times on the Illumina Infinium HumanHap300 genotyping array and studied the concordance of the genotype calls between the three replicates (Table 5). We define a 'consensus call' as a SNP which is called in all three samples and 2× and 3× concordance as SNPs which are identically called in two or three of the samples, respectively. Any SNPs which are called differently in all three samples or contain at least one no call are listed in the column 'No Call'. Both versions of GenoSNP produce highly concordant calls and produces fewer non-concordant calls than GenCall. High concordance should not be mistaken for call accuracy but suggests low call variability, although this maybe at the expense of higher bias. This maybe a natural characteristic of within-sample approaches since, given a genotyping at a SNP for one sample, there is an increased probability of incorrect genotype calls occurring at the same position in other samples as the error is likely to be caused by the SNP possessing unusual hybridization characteristics.

GenoSNP and GenoSNP-VB are not only independent of the reference cohort size but also of the number of SNPs in the study, i.e. the SNP density of the arrays. In order to investigate that, we used the replicates of the same sample as above but choosing different number of SNPs for each one of them. We compared the resulting genotypes of replicates with subsets of size 100 431 SNPs and 205 608 SNPs to the genotypes of one replicate with all 318 238 SNPs. Both algorithms gave agreement around 99%. Since genotyping is performed for each beadpool separately, the chip density should not affect the genotyping results. The differences in agreement are mainly due to biases between the replicates that we addressed above and not caused by the differences in the number of SNPs interrogated.

To illustrate the effects of copy number variation (CNV) on the performance of GenoSNP and GenoSNP-VB compared to the other available algorithms, we calculated the homozygosity rate in detected deletion regions. SNP genotyping methods are not designed to discover CNVs. However, SNPs in regions that are hemizygous for a deletion are expected to be called homozygous for the allele that is present. We used the deletion regions in the HapMap samples that have been identified by recent studies (Redon *et al*., 2006; Wang *et al*., 2007). For both datasets GenoSNP, GenoSNP-VB as well as GenCall and Illuminus produced a homozygosity rate of around 90%. The 'mis-classifications' represent a small number of SNPs and may also be related to the aberrant patterns of SNP genotypes in CNV regions. Even the 'gold standard' HapMap genotypes do not corroborate with the CNV data—they also have a homozygosity rate of around 90% in these regions.

Finally, we tested the performance of GenoSNP and GenoSNP-VB on a case-control study where the true genotypes are not known. We used data from 778 samples of coeliac diseases patients genotyped on the Illumina HumanHap300 genotyping array (van Heel *et al*., 2007). In total 247 017 334 genotypes were available

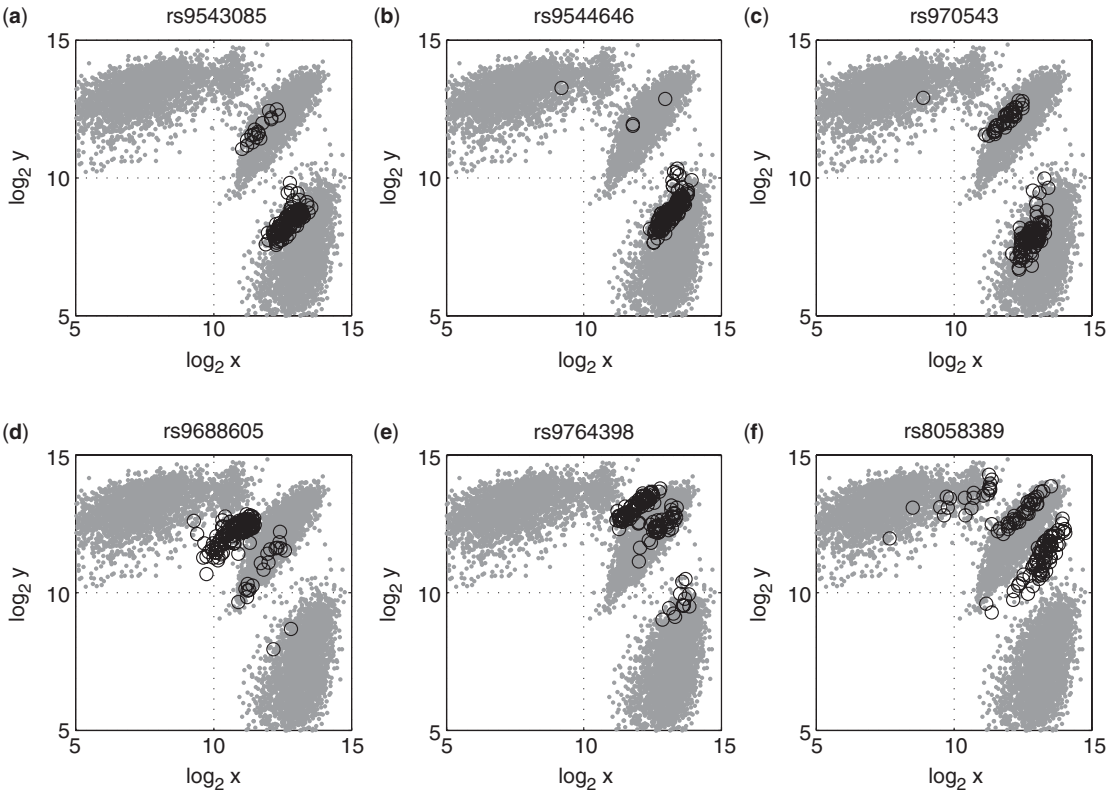

Fig. 3. Six examples showing Illuminus and GenoSNP-VB genotyping failures. The grey dots show log allele-specific probe intensities for all SNPs in bead pool 1 for one HapMap sample. The black circles are intensities for the 120 HapMap samples at the indicated SNP. (**a**–**c**) Illuminus fails for these three SNPs because the method is unable to properly fit a three-component mixture model due to the absence of the minor allele homozygote (low minor allele frequency). GenoSNP-VB is successful as it borrows information from across SNPs and is able to define sample-level genotype clusters. (**d**–**f**) However, GenoSNP-VB fails to call correct genotypes when the cluster centres for certain SNPs deviate considerably from the sample-level clusters.

Table 5. Comparison of call rates and consensus on a sample that was genotyped three times on the Infinium HumanHap300 BeadChip

Method	Consensus calls	Concordance (No. of samples)		No calls
		2×	3×	
GenCall	317 262	1092	315 057	1335
GenoSNP	317 494	857	316 635	9
GenoSNP-VB	317 494	721	316 771	9

There are 317 503 SNPs.

Table 6. Comparison of agreement and call rates on 778 samples of coeliac disease patients

Method 1	Method 2	Agreement (%)	Call rate (%)	
			Method 1	Method 2
GenCall	GenoSNP	99.73	99.55	99.99
GenCall	GenoSNP-VB	99.73	99.55	100
Illuminus	GenoSNP	99.39	99.76	99.99
Illuminus	GenoSNP-VB	99.39	99.76	100
GenCall	Illuminus	99.71	99.55	99.76

for comparison. Table 6 shows the percentage of agreement between every two methods for the SNPs that both methods call. GenoSNP and GenoSNP-VB are able to produce genotypes that are broadly identical to those from GenCall and Illuminus and the agreement is as good as that between GenCall and Illuminus.

4 DISCUSSION

We have developed a genotype calling algorithm, GenoSNP, for the Illumina Infinium SNP genotyping array that is able to call genotypes *within*-sample with comparable accuracy to other population-based genotyping algorithms for the platform. This capability provides researchers involved in studies of any scale with an independent genotyping tool to corroborate genotype calls from Illumina's own proprietary GenCall algorithm and other alternative population-based methods. It is particularly well suited to this task as the within-sample assumption is fundamentally different to that used in population-based approaches and therefore is a truly independent genotyping method. This maybe of considerable use in developing quality control procedures that utilize consensus calls from independent genotyping calling algorithms. Furthermore, whereas the performance of population-based approaches varies with sample size, the performance of GenoSNP is independent of the

size of the study. This stability of output can be useful in developing a reference in quality control metrics.

GenoSNP is easy to compute and the informatics burden is significantly reduced by not requiring a population. GenoSNP call probabilities are shown to be well calibrated enabling their use in downstream analyses such as phasing and genotype imputation. Furthermore, we demonstrate that robust Bayesian clustering using mixtures of Student t-distributions can be applied to genotyping using both standard EM and VB-EM methods. VB-EM carries no significant computational overhead than conventional EM and allows some uncertainty in the model parameters to be taken into account producing more robust inferences.

Our investigations have shown that the Illumina Infinium SNP genotyping technology possesses a high signal-to-noise ratio which produces high inter-class separation and sufficiently low intra-class variation to enable genotype clustering within-sample. We are currently investigating normalization and data transformation methods that would allow similar within-sample genotyping for Affymetrix SNP data. The within-sample genotyping capability is shown to be particularly advantageous for SNPs with very small minor allele frequencies where the existence of heterozygotes and minor allele homozygotes occur rarely. In these instances, a population-based approach would require a large number of samples in order to ascertain the cluster location belonging to the minor allele homozygote, whereas the within-sample approach allows the cluster locations to be determined by borrowing information across SNPs. However, as might be expected, the within-sample approach fails for SNPs whose hybridization characteristics are very different from other SNPs and possess highly shifted genotype clusters. Differences in hybridization characteristics are highly reproducible and suggest that these are related to factors such as probe sequence content. Predictive models of probe behaviour could allow SNP genotyping array manufacturers to identify probes that possess unusual hybridization characteristics and remove these from the final probe sets.

Within-sample genotyping also resolves many issues in large multi-cohort studies where variations in DNA quality between-and within-cohorts means that the definition of a 'reference' population is no trivial issue itself. For example, genotyping methods, such as CHIAMO (Marchini *et al*., manuscript in preparation), use complex hierarchical Bayesian clustering methods to resolve these problems by maintaining different clustering parameters for each cohort in the WTCCC, however, this approach requires a great deal of computational time. In contrast, by working within-sample, GenoSNP calls genotypes based on the characteristics of the sample of interest only and, therefore, variations at a population level are irrelevant.

GenoSNP could also make use of the information provided by samples ran simultaneously on the same array. The new Illumina Human610-quad DNA analysis Beadchip can have up to four samples per chip. GenoSNP could efficiently genotype all four samples at the same time by using the beadpool information and in this way accounting for the chip effect. GenoSNP can then be considered as a within-chip SNP genotyping method.

There are plethora of automated genotype calling algorithms currently in existence for the various commercial SNP genotyping platforms. Each generation of algorithms has taken increasingly complex model-based approaches that are able to handle many practical problems that have arisen from the advent of large-scale SNP genotyping projects. GenoSNP is peculiar in taking a more simplistic approach, seeking instead to fully exploit the high-quality data output from the Illumina SNP genotyping platform to perform *within-sample* genotyping. C++ and MATLAB source codes for GenoSNP are available from the website provided.

ACKNOWLEDGEMENTS

We would like to thank Y.Y. Teo and Taane Clark (Wellcome Trust Centre for Human Genetics, Oxford, UK) for providing early access to their Illuminus manuscript and software, Dan Peiffer (Illumina, Inc. San Diego, USA) for the HapMap genotyping data and GenCall genotype calls and David van Heel (Institute of Cell and Molecular Science, London, UK) for the coeliac disease data.

Funding: E.G. and C.Y. are funded by UK Engineering and Physical Sciences Research Council Life Sciences Interface Doctoral Training Studentships.

Conflict of Interest: none declared.

REFERENCES

Affymetrix Inc. (2006) BRLMM: an improved genotype calling method for the GeneChip human mapping 500K array set. Available at http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf (last accessed August 12, 2008).

Affymetrix Inc. (2007) Birdseed Algorithm – Affymetrix Genotyping Console Software 2.0. Available at http://www.affymetrix.com/products/software/specific/birdseed_algorithm.affx (last accessed August 12, 2008).

Archambeau,C. and Verleysen,M. (2007) Robust Bayesian clustering. *Neural Netw.*, **20**, 129–138.

Beal,M. *et al*. (2003) The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7*, Oxford University Press, Oxford, UK.

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

Dempster,A. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.

Illumina Inc. (2005) Spotlight,T, Illumina GenCall Data Analysis Software. Available at http://www.illumina.com/downloads/GenCallTechSpotlight.pdf (last accessed August 12, 2008).

Laframboise,T. *et al*. (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics*, **8**, 323–336.

Marchini,J. *et al*. (2008) A Bayesian hierarchical mixture model for genotype calling in a multi-cohort study. (In preparation).

Peel,D. and McLachlan,G.J. (2000) Robust mixture modelling using the t distribution. *Stat. Comput.*, **10**, 339–348.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Steemers,F. *et al*. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.

Teo,Y. *et al*. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

van Heel,D. *et al*. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.*, **39**, 827–829.

Wang,K. *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665.

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.