Li *BMC Genetics* (2016) 17:90
DOI 10.1186/s12863-016-0398-x

BMC Genetics

METHODOLOGY ARTICLE Open Access

A new model calling procedure for Illumina BeadArray data

Gengxin Li

Abstract

Background: Accurate genotype calling for high throughput Illumina data is an important step to extract more genetic information for a large scale genome wide association studies. Many popular calling algorithms use mixture models to infer genotypes of a large number of single nucleotide polymorphisms in a fast and efficient way. In practice, mixture models are mostly restricted to infer genotypes for common SNPs where their minor allele frequencies are quite large. However, it is still challenging to accurately genotype rare variants, especially for some rare variants where the boundaries of their genotypes are not clearly defined.

Results: To further improve the call accuracy and the quality of genotypes on rare variants, a new model calling procedure, named M-D, is proposed to infer genotypes for the Illumina BeadArray data. In this calling procedure, a Gaussian Mixture Model and a Dirichlet Process Gaussian Mixture Model are integrated to infer genotypes.

Conclusions: Applications to Illumina data illustrate that this new approach can improve calling performance compared to other popular genotyping algorithms.

Keywords: Dirichlet Process Gaussian mixture model, Gaussian mixture model, Genotype, HapMap, Single nucleotide polymorphism, Rare variants

Background

Genome-wide association studies (GWAS) have been designed to discover many causal genetic variants contributing to human diseases [1, 2]. The success of GWAS relies heavily on the International HapMap Project where millions of single nucleotide polymorphisms (SNPs) have been widely identified on SNP arrays [3, 4]. With the rapid development in biotechnology, a leading producer, Illumina [5], is capable of offering SNP arrays with tremendously wide coverage of genetic variants in a fast and cost efficient way. A number of high dimensional intensity data are generated by this manufacturer, and various powerful genotyping algorithms are imperatively needed to accurately infer genotypes. Recently, several popular calling algorithms have been designed for Illumina platform, such as: BEAGLE with BEAGLECALL software [6], CRLMM [7, 8], GenCall [9], GenoSNP [10], and Iluminus [11]. In general, Illumina chip catalogs millions of SNPs and processes a large number of parallel samples, and the genotyping algorithms for the Illumina data is of the main interest.

With the application of single base extension (SBE) biochemistry technology [12], the Illumina data measures the pair of intensities with two alleles (*A* and *B*) at every SNP for each individual. Typically, a SNP with alleles *A* and *B* makes three possible genotype clusters, named *AA*, *AB*, and *BB*, and all possible genotypes of each SNP are called by various genotyping algorithms. One strategy is the population-based approach through which genotypes of all individuals within a SNP are inferred at one time, but its calling performances highly depend on the size of population. Thus, this method is not applicable for rare SNPs with low minor allele frequency (MAF). Another approach, GenoSNP, is designed to infer all SNP genotypes within one individual simultaneously, and is referred to as a SNP-based calling method. The applicability of this algorithm [10] relies on the assumptions that response features of all probes are similar. Compared to the population-based method, it would be unnecessary to collect a large number of samples for rare SNP calling due to the availability of high density SNPs. Unfortunately, this method leads to a larger proportion of SNPs breaking

Correspondence: gengxin.li@wright.edu
Department of Mathematics and Statistics, Wright State University, 3640 Colonel Glenn Hwy, 45435 Dayton, USA

the Hardy-Weinberg (HW) principle which violates the assumption that commonly occurs in practice.

Most of the predominant calling algorithms employ the mixture models [13–15] to infer three genotype clusters. In particular, the mixture models developed from the population-based strategy work well for common SNPs but gradually lose their effectiveness for rare variants. To improve calculation accuracy, the mixture models need a sufficient number of observations in each genotype cluster to precisely estimate parameters. However, rare SNPs always contain a small number of individuals in one or two genotype clusters, and some rare SNPs with extremely small values of MAF may lose one or two clusters. This phenomenon creates two problems: (1) the number of components for rare SNPs is uncertain; (2) the boundaries of some genotype clusters are not clear for rare SNPs with sparsely populated observations. The problem about developing better inference for rare SNPs motivates the use of the Dirichlet Process (DP) Gaussian Mixture Model (GMM) [16–18]. One popular application of DP is clustering in the fields of brain imaging, information retrieval and genetics. To successfully perform a cognitive task, DP has been applied to analyze activation structures in functional magnetic resonance imaging [19]. DP has also been used to model relationships among documents in the field of information retrieval [20, 21]. For better understanding of ancestry history in the genetic study, DP was smoothly adopted to identify the sets of haplotypes corresponding to subpopulation [21]. Due to its good characteristics in clustering, this paper extends DP model to the genotyping area. Specifically, a DP prior plays a critical role in clustering data through defining a mixture model with a variable number of components. More importantly, its clustering and discreteness properties allows an easy partitioning of the data into different groups, even though some observations lack clear cluster membership. Besides, empirical studies have showed that GenoSNP can improve the genotyping quality for rare variants through calling a large number of SNPs within one individual. However, the genotype clusters implemented by GenoSNP may be in a shift away from their expected positions, which could result in many SNPs breaking the HW principle [5]. For a DP Gaussian Mixture Model (DP-GMM), its model selection procedure is based on a rich-gets-richer phenomenon [17], which indicates that the cluster with an extremely small number of observations is still toughly estimated. A reference SNP selection step [22] is incorporated here to infer genotypes of rare SNPs with extremely low MAF, and this new method may solve the HW principle problem.

In this paper, a new model calling procedure (M-D) is an approach that is made up of two models and one SNP selection procedure, namely Gaussian Mixture Model, DP Gaussian Mixture Model, and reference SNP selection. In brief, this method partitions SNPs into three groups in terms of the SNP's MAF and the sample size of each cluster. In this method, three models are applied in three groups individually. The performance of M-D is evaluated through comparison with other genotyping algorithms for Illumina BeadArray data.

Methods

Illumina BeadArray data

The Illumina Omni BeadArray chip collects over one million SNPs per sample, and increasingly covers the newly identified variants. In the probe design, every beadtype that is capable of assaying two SNP alleles represents a SNP [12]. A large number of beadpools that include millions of beadtypes results in the ultimate production of the Illumina microarray. Here, Illumina data measures the pair of raw intensity at each beadtype for every sample, and the genotype clusters are estimated at this scale.

Statistical models

Model I: Gaussian mixture model (GMM)

The pair of raw intensity $\mathbf{x}_{is} = (r_{is}, g_{is})$ for the ith individual at the sth SNP is the basic measurement. Within one SNP, all subjects' intensity data may fall into three genotype clusters corresponding to three genotypes (AA, AB, BB) and one null component which collects the abnormal raw intensity measurements. Model I is a Gaussian Mixture Model [23] that is applied to the basic measurement $\mathbf{x}_{is}$. In principle, this model assigns each pair of raw intensities $\mathbf{x}_{is}$ to one of the components with probability π_{ks} for $k = 1$, 2 or 3. The relevant latent genotype class is measured by an indicator variable z_{is} generated from a multinomial distribution ($Mult_3$) where $z_{is} = 1$, 2 or 3. Then this Gaussian Mixture Model can be expressed as:

$$\begin{aligned} & z_{is} \sim Mult_3(1, \pi_{1s}, \pi_{2s}, \pi_{3s}) \\ & \ell(\mathbf{x}_s|\mathbf{\Theta}_s, \mathbf{z}_s) = \prod_{i=1}^{n_s} \prod_{k=1}^{3} \Psi(\mathbf{x}_{is}|\boldsymbol{\mu}_{ks}, \Sigma_{ks})^{I(z_{is}=k)} \\ & \boldsymbol{i = 1,\ldots, n_s, s = 1,\ldots, S, k = 1, 2 \text{ or } 3} \end{aligned} \tag{1}$$

where n_s is the total number of individuals observed at the sth SNP, and S is the total number of SNPs. Ψ denotes a normal density with mean $\boldsymbol{\mu}_{ks}$ and variance-covariance matrix Σ_{ks} in the kth component at the sth SNP; all pairs of raw intensity within the sth SNP are measured by $\mathbf{x}_s$=($\mathbf{x}_{1s}$, $\mathbf{x}_{2s}$,..., $\mathbf{x}_{n_s s}$); the unknown parameters of the GMM is denoted by $\mathbf{\Theta}_s$=($\boldsymbol{\pi}_s$, $\boldsymbol{\mu}_s$, Σ_s) where $\boldsymbol{\pi}_s$=(π_{1s}, π_{2s}, π_{3s}), $\boldsymbol{\mu}_s$=($\boldsymbol{\mu}_{1s}$, $\boldsymbol{\mu}_{2s}$, $\boldsymbol{\mu}_{3s}$), and Σ_s=(Σ_{1s}, Σ_{2s}, Σ_{3s}).

The maximum likelihood estimates of the parameters are inferred [23]. For the indicator variable $z_{is} = k$, the $(t+1)$th iteration is estimated by

$$f_k\left(\mathbf{x}_{is}; \mathbf{\Theta}_s^t\right) = \frac{\pi_{ks}^t \Psi\left(\mathbf{x}_{is}; \boldsymbol{\mu}_{ks}^t, \Sigma_{ks}^t\right)}{\sum_{u=1}^{3} \pi_{us}^t \Psi\left(\mathbf{x}_{is}; \boldsymbol{\mu}_{us}^t, \Sigma_{us}^t\right)} \tag{2}$$

The iterative estimates of mean $\boldsymbol{\mu}_{ks}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{ks}$ are expressed as,

$$\boldsymbol{\mu}_{ks}^{t+1} = \frac{\sum_{i=1}^{n_s} f_k\left(\mathbf{x}_{is}; \boldsymbol{\Theta}_s^t\right)\mathbf{x}_{is}}{\sum_{i=1}^{n_s} f_k\left(\mathbf{x}_{is}; \boldsymbol{\Theta}_s^t\right)} \tag{3}$$

$$\Sigma_{ks}^{t+1} = \frac{\sum_{i=1}^{n_s} f_k\left(\mathbf{x}_{is}; \boldsymbol{\Theta}_s^t\right)\left(\mathbf{x}_{is} - \boldsymbol{\mu}_{ks}^{t+1}\right)\left(\mathbf{x}_{is} - \boldsymbol{\mu}_{ks}^{t+1}\right)^T}{\sum_{i=1}^{n_s} f_k\left(\mathbf{x}_{is}; \boldsymbol{\Theta}_s^t\right)} \tag{4}$$

Two measurements Posterior Rate (PR: p_{is}^k) and the Average Posterior Rate (APR: p_s) for the sth SNP are adopted to assess the quality of SNP calling [22]. Specifically, PR quantifies the strength of every individual's cluster signal, and APR gives the average strength of all individuals at the sth SNP [22].

$$PR: p_{is}^k = \frac{P(\mathbf{x}_{is}|k)\pi_{ks}}{\sum_{u=1}^{3} P(\mathbf{x}_{is}|u)\pi_{us}}$$

$$APR: p_s = \frac{\sum_{k=1}^{3}\sum_{i=1}^{n_{ks}} p_{is}^k}{\sum_{k=1}^{3} n_{ks}}$$

Note that $P(\mathbf{x}_{is}|k)$ is a conditional probability of the ith individual given that this subject is assigned to the kth cluster, and n_{ks} is the sample size of the kth cluster at the sth SNP.

Model II: Dirichlet Process Gaussian mixture model (DP-GMM)

Model I is a fast and efficient genotyping model for SNPs having large values of MAF. In real experiments, many SNPs with low MAF may result in the disappearance of one or two genotype clusters. Also even though some SNPs with low MAF display three genotype groups, some clusters may lack sufficient data to support and recognize. In this case, Model II, DP Gaussian Mixture Model, is motivated by the need to carry out the model selection for SNPs with an uncertain number of genotype clusters [24]. Generally speaking, this is a nonparametric Bayesian method that potentially allows a flexible number of mixture components and also provides estimates for the mixture component parameters and the relevant mixing proportions.

A DP Gaussian Mixture Model [24] fits the pair of raw intensity $\mathbf{x}_{is}$ into K-component Gaussian Mixture Model with K approaching a large number. The model is expressed as,

$$\ell(\mathbf{x}_s|\boldsymbol{\Theta}_s, \mathbf{z}_s) = \prod_{i=1}^{n_s}\prod_{k=1}^{K} \Psi(\mathbf{x}_{is}|\boldsymbol{\mu}_{ks}, \Sigma_{ks})^{I(z_{is}=k)} \tag{5}$$
$$\boldsymbol{i = 1,\ldots, n_s, s = 1,\ldots, S, k = 1,\ldots, K}$$

where K is the total number of clusters. $\boldsymbol{\Theta}_s$=($\boldsymbol{\pi}_s$, $\boldsymbol{\mu}_s$, Σ_s) denotes the unknown parameters at the sth SNP where $\boldsymbol{\pi}_s$=(π_{1s}, ..., π_{Ks}), $\boldsymbol{\mu}_s$=($\boldsymbol{\mu}_{1s}$, ..., $\boldsymbol{\mu}_{Ks}$), and Σ_s=(Σ_{1s}, ..., Σ_{Ks}). Generally, the number of observations within the sth SNP (n_s) are partitioned into K components (n_{1s}, n_{2s}, ..., n_{Ks}) with relevant mixing proportions (π_{1s}, π_{2s}, ..., π_{Ks}). The distribution of n_{1s}, n_{2s}, ..., n_{Ks} follows a multinomial distribution and its probability mass function is written by,

$$p(n_{1s}, n_{2s}, \ldots n_{Ks}|\pi_{1s}, \pi_{2s}, \ldots, \pi_{Ks}, n_s) = \frac{n_s!}{n_{1s}!\, n_{2s}! \ldots n_{Ks}!}\prod_{k=1}^{K}\pi_{ks}^{n_{ks}} \tag{6}$$

where $n_s = \sum_{k=1}^{K} n_{ks}$ denotes the total number of individuals at the sth SNP. Then each pair of raw intensity for the sth SNP $\mathbf{x}_{is}$ has its own indicator z_{is} (i = 1, ...,n_s), and the distribution of indicator variables is expressed as,

$$p(z_{1s}, z_{2s}, \ldots z_{n_s s}|\pi_{1s}, \pi_{2s}, \ldots, \pi_{Ks}) = \prod_{k=1}^{K}\pi_{ks}^{n_{ks}} \tag{7}$$

The model can then be expressed as:

$$\begin{aligned}
&\boldsymbol{\pi_s}|\alpha \sim Dir\left(\frac{\alpha}{K}, \frac{\alpha}{K}, ..., \frac{\alpha}{K}\right)\\
&z_{is}|\boldsymbol{\pi_s} \sim Discrete(\pi_{1s}, \pi_{2s}, ..., \pi_{Ks})\\
&\boldsymbol{R}_{ks}|\nu, \boldsymbol{S} \sim W\left(\nu, \boldsymbol{S}^{-1}\right)\\
&\boldsymbol{\mu}_{ks}|m, r, \boldsymbol{R}_{ks} \sim N(m, r\boldsymbol{R}_{ks})\\
&\mathbf{x}_{is}|z_{is}, \boldsymbol{\Theta}_s \sim N(\boldsymbol{\mu}_{z_{is}s}, \mathbf{R}_{z_{is}s})
\end{aligned} \tag{8}$$

where α is the DP concentration parameter and can be thought as the inverse variance of DP. The distribution of the reciprocal of α follows a Gamma distribution with 1 degree freedom and mean 1. K is the maximum number of clusters, then $\boldsymbol{\pi_s}$ is distributed with a symmetric Dirichlet distribution with parameter $\frac{\alpha}{K}$. m and r are hyperparameters being the mean and relative precision of $\boldsymbol{\mu}_{ks}$, and the hyperparameters ν and $\boldsymbol{S}^{-1}$ are degrees of freedom and inverse mean of $\mathbf{R}_{ks}$ where $\mathbf{R}_{ks}$ follows a Wishart distribution with parameters ν and $\boldsymbol{S}^{-1}$, respectively.

The inference on Model II relies on the posterior distribution of each parameter conditional on all other parameters, then the parameters, hyperparameters and indicator variables are repeatedly sampled from their posterior distributions. In particular, the conditional posterior probabilities are proportional to the likelihood function multiplying priors. Then the posterior probabilities of the cluster indicator variable z_{is} conditional on all other variables are expressed as:

$$p(z_{is} = k|z_{-is}, \boldsymbol{\mu}_s, \mathbf{R}_s, \alpha, m, r, \nu, \mathbf{S}) \propto \begin{cases} \frac{n_{-i,ks}}{n_s - 1 + \alpha} N(\mathbf{x}_{is}|\boldsymbol{\mu}_{ks}, \mathbf{R}_{ks}) \\ \qquad \text{if } k \text{ is an existing cluster, and } n_{-i,ks} > 0 \\ \frac{\alpha}{n_s - 1 + \alpha}\int p(\mathbf{x}_{is}|\boldsymbol{\mu}_{ks}, \mathbf{R}_{ks}) p(\boldsymbol{\mu}_{ks}, \mathbf{R}_{ks}|m, r, \nu, \mathbf{S}) d\boldsymbol{\mu}_{ks} d\mathbf{R}_{ks} \\ \qquad \text{if k is a new cluster} \end{cases} \tag{9}$$

Note that $p(\mathbf{x}_{is}|\boldsymbol{\mu}_{ks}, \mathbf{R}_{ks})$ and $p(\boldsymbol{\mu}_{ks}, \mathbf{R}_{ks}|m, r, \nu, \mathbf{S})$ are the likelihood function and the joint function of parameters ($\boldsymbol{\mu}_{ks}$ and $\mathbf{R}_{ks}$), respectively. Once the optimal genotype clusters and their relevant component parameters are obtained, two measurements Posterior Rate (PR) and the Average Posterior Rate (APR) measuring the quality of the sth SNP can be calculated in the similar way.

Model III: Dirichlet Process Gaussian mixture model with reference SNP selection (DP-Ref)

A DP Gaussian Mixture Model with reference SNP selection step (DP-Ref) combines the benefits of the population-based method with the SNP-based approach. In this context, the reference SNP selection plays an important role in determining the effectiveness of Model III. A reference SNP is referred to as a good quality SNP providing sufficient information about three genotypes clusters, thus each SNP in the third group will be called with assistants of the carefully selected reference SNP. Practically, the final reference SNP is selected by a three-step procedure [22]. Through out this section, each SNP in the third group is denoted as the "T-SNP" that needs to be called with the support of a reference SNP, and the final reference SNP having good quality is defined as "R-SNP"

Step I. High MAF SNPs are selected as candidate reference SNPs. In fact, SNPs with large MAF (> 0.15) before the T-SNP are selected as R1-SNPs.

Step II. Good clustering property SNPs from R1-SNPs are further selected (denoted as R2-SNPs). This step requires three genotype clusters of R1-SNPs to contain at least 10 % of entire observations individually.

Step III. A SNP from R2-SNPs being remarkably similar to the T-SNP is selected (denoted as R-SNP). The resemblance between the T-SNP and each R2-SNP is measured by the cluster distance D_t [22]. For simplifying the calculation, two dimensional raw intensity vector $\mathbf{x}_{is}$ is projected to an univariate variable y_{is} [11], and the T-SNP and all R2-SNPs are classified into three genotype clusters in terms of this univariate variable.

$$y_{is} = \frac{r_{is} - g_{is}}{r_{is} + g_{is}}$$

$$\begin{cases} y_{is} \text{ if } y_{is} < 0.5 \\ y_{is} \text{ if } -0.5 \leq y_{is} < 0.5 \\ y_{is} \text{ if } y_{is} \geq 0.5 \end{cases} \tag{10}$$

Empirical studies show that the initial cutoffs dividing the univariate variable $\mathbf{y}_s = (y_{1s}, ..., y_{n_s s})^T$ into three clusters can be fixed as 0.5 and -0.5. The cluster label of each individual would be roughly determined by the above equation.

We select one SNP from the third group as the T-SNP, then the cluster measure (D_t) is to find the minimum distance between the T-SNP and R2-SNPs [22]. The SNP from R2-SNP gives the minimum distance will be the R-SNP. The cluster measure is calculated by,

$$D_t = \min_{d;d\in\Xi} \left\{ \sum_{k=1}^{3} trace \left\{ (\mathbf{x}_{kt} - \boldsymbol{\mu}_{kd})((\Sigma_{kt} + \Sigma_{kd})/2)^{-1}(\mathbf{x}_{kt} - \boldsymbol{\mu}_{kd})^T \right\} \right\} \tag{11}$$

Note that Ξ is the set of R2-SNPs selected for the T-SNP; $\mathbf{x}_{kt}$ and Σ_{kt} are the raw intensity vector and variance-covariance matrix in the kth cluster for the T-SNP; $\boldsymbol{\mu}_{kd}$ and Σ_{kd} are the mean and variance-covariance matrix of the dth R2-SNP; In brief, The final R-SNP will provide sufficient clusters information to assist the testing T-SNP.

A new augmented vector is generated by combining the T-SNP with the final reference SNP,

$$\mathbf{m}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_d \end{pmatrix}$$

where $d \in \Xi$, and the second model DP-GMM (Eqs. 6–9) will be applied to the combined raw intensities to identify the genotype clusters through the aid of the reference SNP.

Application of new model (M-D)

This section focuses on the application of M-D. Specifically, entire SNPs are classified into three groups, and an appropriate model is selected to fit in each group. The classification standard relies on the calculations of MAF and the sample size of each genotype cluster through Model I. The reason for choosing this model is that GMM can quickly estimate the SNP's MAF and the sample size of each genotype cluster. Other advanced models (Model II and III) will be applied to the selected SNPs with small MAFs. According to this calling procedure, any SNP will be classified by,

$$\boldsymbol{x}_s = \begin{cases} \boldsymbol{x}_s \in g_1 & \text{if MAF} \geq 0.05. \\ \boldsymbol{x}_s \in g_2 & \text{if MAF} < 0.05 \text{ and } b_1 \leq n_{ks} < b_2 \\ & \quad \text{for any one of clusters.} \\ \boldsymbol{x}_s \in g_3 & \text{otherwise.} \end{cases} \tag{12}$$

Note that n_{ks} is the sample size of the kth cluster at the sth SNP. The first group (g_1) collects SNPs with high MAF (≥ 0.05), and a large proportion of SNPs is in this group. The second group (g_2) includes SNPs with low MAF (< 0.05) and a certain number of subjects in either existing genotype clusters. In this study, b_1 and b_2 are fixed as 3 and 10 to determine the number of SNPs recruited in g_2. The last group (g_3) collects the rest SNPs with low MAF and a small number of observations in one or two genotype clusters. In fact, SNPs in g_1 can display three genotype clusters (one major homozygote, one minor homozygote and one heterozygote) with a large number of subjects in each cluster. The rest poor SNPs with low MAF are contained in g_2 and g_3 where some SNPs may

not display three genotype clusters, either one or two clusters disappear and the existing cluster may contain very few observations. In particular, the classification between g_2 and g_3 is not fixed, and scientists can easily manage the allocation of SNPs between g_2 and g_3 through adjusting the values of b_1 and b_2.

The proposed new calling procedure is based on the partitions of SNPs.

$$\begin{cases} g_1: & \text{GMM} \\ g_2: & \text{DP-GMM} \\ g_3: & \text{DP-Ref} \end{cases} \tag{13}$$

In the first group, Model I (GMM) is applied to genotype SNPs. A sufficient number of observations are observed in three genotype clusters, which will greatly help the genotyping procedure identify the boundary of each cluster. In the second group, SNPs with low MAF, Model II (DP-GMM) can implement the model selection to search the appropriate number of clusters for each SNP, and DP's clustering and discreteness properties assures the optimum partition of observations, even for a small number of observations in a genotype cluster. In the third group, the number of genotype clusters for each SNP is uncertain and an extremely small number of observations are observed in either one or two clusters. In this case, applying a DP-GMM alone for clustering is not enough due to a rich-gets-richer phenomenon [17] where the larger genotype cluster can greatly attract sparsely populated observations that originally belong to another cluster. In view of this situation, the reference SNP strategy [22] is applied to help DP-GMM call rare SNPs (DP-Ref). More importantly, the selection of models can be determined through adjusting b_1 and b_2 in Eq. 12. For example, when b_1 takes a large value, a smaller proportion of rare SNPs may enter g_2 and more rare SNPs are allocated to g_3, thus GMM and DP-Ref will become major methods. If b_2 takes a large value, a larger proportion of rare SNPs may be assigned to g_2, then GMM and DP-GMM will become main methods. This flexible option provides more solutions for scientist who are interested in this genotyping method.

In this study, b_1 and b_2 are fixed as 3 and 10, then 88.6 % of SNPs are in g_1, 4.03 % and 7.37 % of SNPs will be assigned to g_2 and g_3, respectively. More importantly, DP Gaussian Mixture Model is powerful to infer the cluster containing a certain number of observations, thus Fig. 1 displays the genotyping results of three SNPs inferred by DP-GMM (rs1003505 MAF: 0.0479, rs1004262 MAF: 0.0404, rs1009148 MAF: 0.0439). For the extremely rare variants in g_3, DP-Ref is used to infer genotypes (rs10084633 MAF: 0.0166, rs1003945 MAF: 0.0118, rs1008185 MAF: 0), and the calling results are summarized in Fig. 2. To clearly illustrate the effect of the reference SNP on rare SNP calling in g_3, Fig. 3 displays how the reference SNP help rare SNP be genotyped. It is clearly seen that our model could actively infer genotypes of rare SNPs under the support of the reference SNP.

Results and discussion

Illumina BeadArray data description

The proposed method M-D is applied to an Illumina data consisting of 1 million SNPs and 3258 samples.

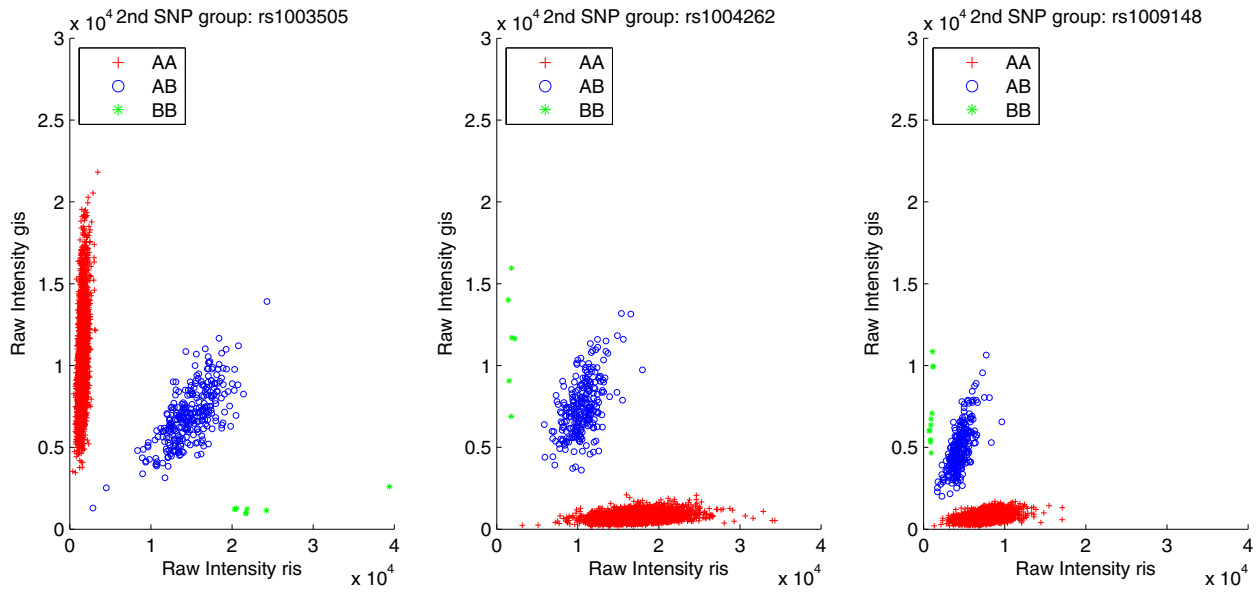

Fig. 1 Performance of the DP Gaussian Mixture Model on genotyping three rare SNPs (rs1003505, rs1004262, rs1009148)

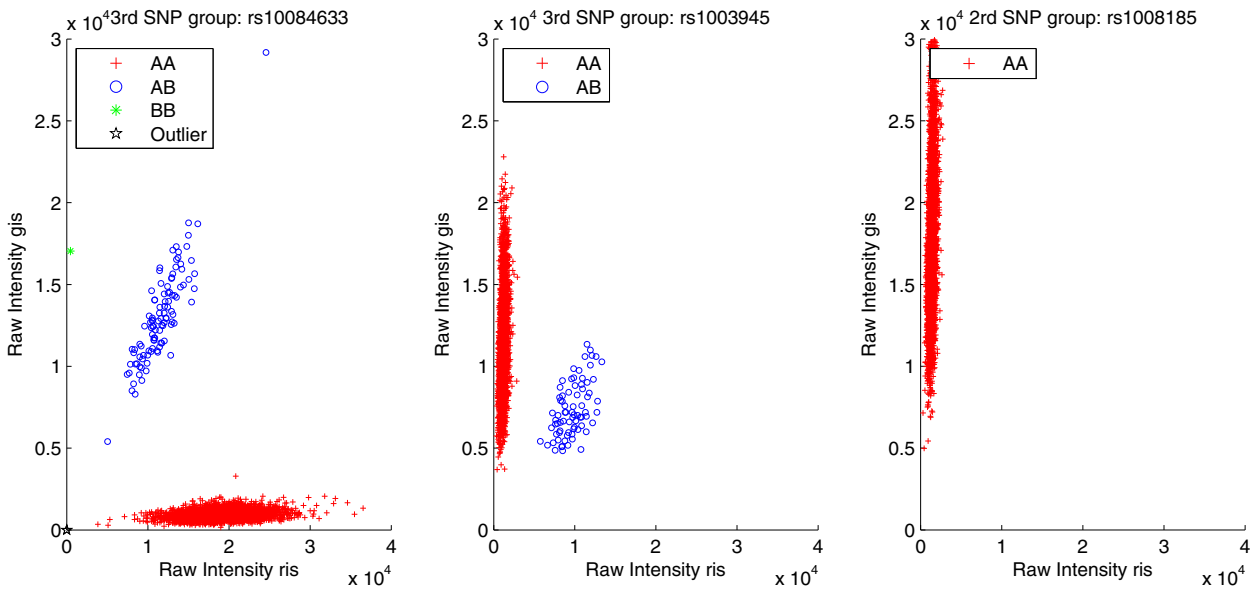

Fig. 2 Performance of the DP Gaussian Mixture Model with reference SNP selection on genotyping three rare SNPs (rs10084633, rs1003945, rs1008185)

Specifically, there are 38 different HapMap samples [3] measured multiple times to produce 141 repeated HapMap samples in this data. SNP calls from the chromosome 22 are analyzed. The performance of *M-D* is compared to those of GenCall representing a population-based method and GenoSNP standing for a SNP-based approach. The compatible cutoffs of all three calling algorithms are carefully selected, such as: GenCall score (GC score ≥ 0.15) is used to filter good quality SNPs; GenoSNP and *M-D* collect good quality SNPs and samples through the posterior probability (≥ 85 %).

Results

The performances of three calling algorithms are compared in terms of the call rate $\left(\frac{\text{genotypes that can be inferred}}{\text{genotypes that are supposed to be genotyped}}\right)$ and the concordance rate measuring the genotype agreement between any two algorithms. The overall comparison results are given in Table 1. It is clearly seen that

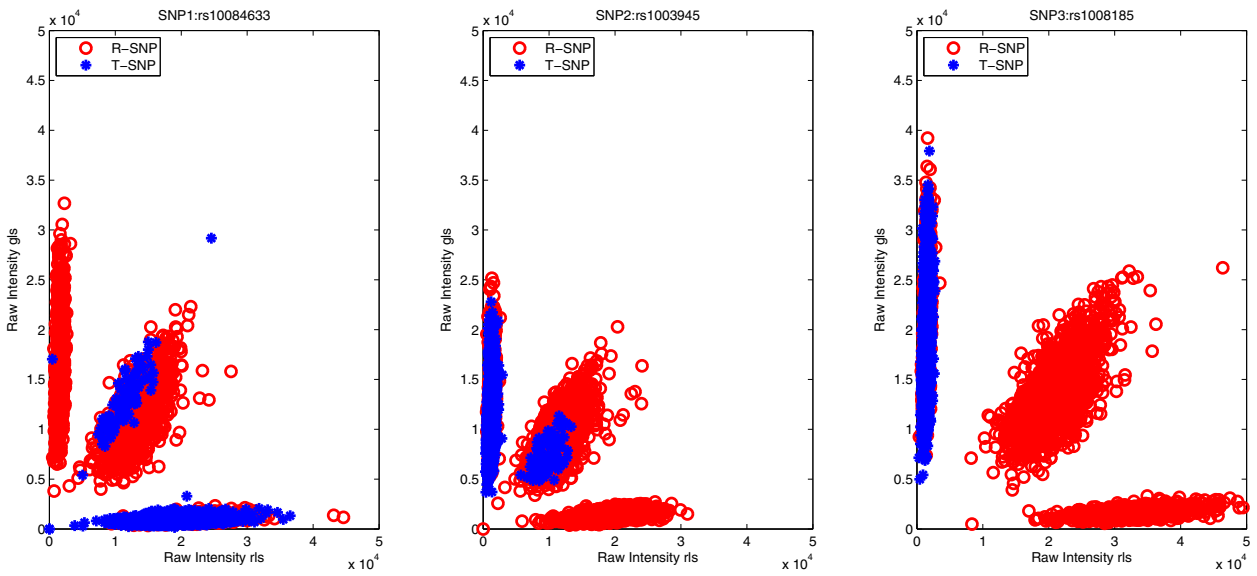

Fig. 3 Performance of the reference SNP selection on genotyping three rare SNPs (rs10084633, rs1003945, rs1008185)

Table 1 The comparisons of call rate and concordance rate among GenCall, GenoSNP and M-D

Algorithm 1	Algorithm 2	Call rate (%)		Concordance (%)
		Algorithm 1	Algorithm 2	
GenCall	M-D	96.71	99.71	99.93
GenoSNP	M-D	99.12	99.71	99.65
GenCall	GenoSNP	96.71	99.12	99.71

Note: The unit of Call Rate and Concordance Rate is percentage %; *M-D*: a new model calling procedure

genotypes inferred from M-D, GenCall and GenoSNP are highly consistent, and genotypes from M-D are more consistent with those inferred from GenCall (99.93 %), than those from GenoSNP (99.65 %). This is because M-D is a population-based method in a wide sense, and the model selection of a DP and the reference SNP selection step in M-D greatly improve its call accuracy and call rate (Table 1).

Because most samples in this Illumina data are collected from the hospital, true genotypes of these sample are not known totally, so the high agreement among 3 algorithms can not tell us which method performs best. Fortunately, 141 HapMap samples are contained in this data, and the true genotypes of these HapMap samples are explored by the HapMap project as a gold standard. Table 2 provides the comparison results between each discussed method and a gold standard in terms of the call accuracy and the call rate. In brief, M-D gives the best call accuracy and the largest call rate, followed by GenoSNP and GenCall. For example: the largest call rate is achieved by M-D (99.78 %), followed by GenoSNP (99.14 %) and GenCall (96.79 %). Moreover, M-D offers the best call accuracy (99.44 %), followed by GenoSNP (98.52 %), and GenCall (96.63 %).

Compared to the population-based methods (GenCall) and the SNP-based approaches (GenoSNP) [9, 10], the new model (M-D) is expected to perform better because it integrates a model selection step of DP and predominance of the population-based and the SNP-based strategies. In this study, SNPs are classified into 3 groups according to Eq. 12, and the comparison results corresponding to these 3 groups are summarized in Table 3. In brief, M-D gives the best call accuracy and largest call rate, followed by GenoSNP and GenCall. In particular, g_2 and g_3 collects whole rare SNPs, again, M-D still outperforms GenoSNP and GenCall on call accuracy and call rate.

Table 2 The comparisons of call rates and accuracy on HapMap samples for overall SNPs

Criterion	Item	GenCall (%)	GenoSNP (%)	M-D (%)
All SNPs	Call rate	96.79	99.14	99.78
	Accuracy	96.63	98.52	99.44

Note: M-D: a new model calling procedure; Call rate: the percentage of valid genotypes; Accuracy: the percentage of consistent genotype between each calling method and the gold standard

Table 3 Comparisons of call rates and accuracy on HapMap samples for three SNP groups

Class	Prop	Item	GenCall	GenoSNP	M-D
g_1	88.60 %	Call rate	96.59	99.13	99.77
		Accuracy	96.40	98.44	99.31
g_2	4.03 %	Call rate	97.62	99.56	99.75
		Accuracy	97.53	99.45	99.59
g_3	7.37 %	Call rate	96.60	99.14	99.70
		Accuracy	96.45	98.71	99.40

Note: M-D: a new model calling procedure; Call rate: the percentage of valid genotypes; Accuracy: the percentage of consistent genotype between each calling method and the gold standard; Class: indicates the three SNPs categories, such as: g_1, g_2 and g_3; Prop: indicates the percentage of SNPs which belong to three groups, respectively

Hardy-Weinberg Equilibrium (HWE) test is another important criteria to examine the quality of SNPs. In this Illumina data, most samples are from four populations: Hispanic African-American, non-Hispanic African-American, Hispanic European-American, and non-Hispanic European-American. The HWE test (P-value < 0.0001) is applied to four populations separately. The total number of SNPs failing the HWE test are summarized in Table 4. A SNP-based method, GenoSNP, considers all SNPs calls within a sample at a time to improve genotyping quality for rare variants, but a large number of SNPs corresponding to four populations break the HW

Table 4 Comparisons of Hardy-Weinberg Equilibrium test among GenCall, GenoSNP and M-D

Population	Num-Sample	Algorithm	# of failed SNPs
AA I	2005	GenCall	224
		GenoSNP	907
		M-D	422
AA II	83	GenCall	20
		GenoSNP	254
		M-D	80
EA I	867	GenCall	486
		GenoSNP	1024
		M-D	643
EA II	158	GenCall	40
		GenoSNP	348
		M-D	133

Note: AA I: African-Americans not of Hispanic Origin; AA II: African-Americans of Hispanic Origin; EA I: European Americans not of Hispanic Origin; EA II: European Americans of Hispanic Origin; Num-Sample: the number of subjects within each population; Algorithm: three algorithms in this table, that is, GenCall, GenoSNP, and M-D; # of failed SNPs: the number of SNPs fail the Hardy-Weinberg Equilibrium test within each population

principle. In contrast, GenCall applies the population-based strategy to call all individuals within one SNP, so the calling results are less biased, and a small number of SNPs fail the HWE test. M-D is also a population-based model in a wide sense, and the quality of SNP calls is much better than that from GenoSNP, a moderate number of SNPs break the HW principle. In summary, M-D performs well on genotyping rare variants and controlling the quality of SNPs.

Discussion

The principle of a DP Gaussian Mixture Model is to run a model selection procedure to explicitly estimate the number of components for rare variants. The concentration parameter measures the inverse variance of DP, which suggests that a larger concentration parameter implies an increasing number of components [17]. It brings a new problem of how to select the appropriate strength of the prior to control the number of components. In particular, this parameter is sensitive to SNPs where sparsely populated observations are in one or two components. There might be better ways to define this parameter to help the DP Gaussian Mixture Model more efficiently call genotypes for rare variants.

The DP mixture model incorporates the reference SNP selection step to take advantage of the population-based strategy and the SNP-based strategy for improving the missing rate and call accuracy for rare SNPs. The successful application of M-D is also based on the selection of the reference SNP across the genome. In practice, it is difficult to search the reference SNP from the entire genome due to the heavy calculation burden. In these cases, the instrumental SNPs before the testing SNP are picked out. When some probes break the assumption about identical probe responses for various SNPs, searching the best reference SNP is still challenging. In particular, the method about accurately measures the similarity between the testing SNP and the reference SNP still needs to be improved.

Conclusion

One classical genotyping approach is the population-based method, GenCall, and it requires a large number of observations to achieve a nice call accuracy. When the increasing number of rare variants are commonly identified on the large scale Illumina array, it is extremely difficult to successfully call genotypes for rare variants. A SNP-based method, GenoSNP, was designed to solve this challenging problem, but many more SNPs inferred from GenoSNP break the HW principle. In this paper, a new model calling procedure (M-D) is proposed to take benefits of a model selection step of a DP and the advantage of GenCall and GenoSNP to improve the quality of rare SNP calls. In brief, the new model calling procedure partitions SNPs into three classes in terms of MAF and the sample size of each cluster, and a DP Gaussian Mixture Model with or without reference SNP selection are applied to rare SNPs with low MAF. The finest performance of M-D is evaluated by comparing genotypes inferred by each discussed calling method to those from the HapMap project. Compared to GenCall and GenoSNP, M-D performs better on genotyping rare SNPs, and it also infers better quality of SNP calls than that from GenoSNP.

Abbreviations

APR, average posterior rate; D_t, cluster measure; DP, Dirichlet Process; DP-GMM, Dirichlet Process Gaussian mixture model; DP-Ref, Dirichlet Process Gaussian mixture model with reference SNP selection; GMM, Gaussian mixture model; GWAS, genome-wide association studies; HW, Hardy Weinberg; MAF, minor allele frequency; $Multi_3$, a multinomial distribution; M-D, a new model calling procedure; PR, posterior rate; R-SNP, final reference SNP; SBE, single base extension; SNPs, single nucleotide polymorphisms; T-SNP, a SNP in the third group

Acknowledgements

I sincerely thank Dr. Hongyu Zhao and Dr. Joel Gelernter for providing the raw intensity data and valuable advice.

Funding

This study was supported by National Institutes of Health Grants RC2 DA028909, R01 DA12690, R01 DA12849, R01 DA18432, R01 AA11330, and R01 AA017535, the Veterans Affairs Connecticut Mental Illness Research, Educational, and Clinical Centers, and Wright State University start-up.

Availability of data and materials

Dr. Gelernter Lab at Yale university generated the raw intensity data, and the availability of this data needs to get the lab's permission.

Authors' contributions

Conceived of the study and proposed the method: GL; Analyzed the method and interpreted the results: GL; Wrote the manuscript: GL.

Competing interests

The author declares that there is no competing interests.

Consent to publish

Not applicable.

Ethics approval and consent to participate

The study was approved by the Human Investigations Committees at Yale University, and all subjects signed written informed consent before participation.

Software

Software with a sample data set is available on request from the corresponding author (gengxin.li@wright.edu).

Received: 11 February 2016 Accepted: 16 June 2016
Published online: 24 June 2016

References

1. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genomewide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445:881–5.
2. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. Nature. 2007;447:661–78.
3. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–61.
4. Reich DE, et al. Quality and completeness of SNP databases. Nat Genet. 2003;33:457–8.

5. Ritchie ME, Liu RJ, Benilton S, Carvalho BS. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinforma. 2011;12:68.
6. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet. 2009;85(6):847–61.
7. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics. 2007;8:485–99.
8. Ritchie ME, Carvalho BS, Hetrick KN, Tavare S, Irizarry RA. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. Bioinformatics. 2009;25(19):2621–3.
9. Illumina Inc. Illumina GenCall Data Analysis Software. TECHNOLOGY SPOTLIGHT. 2005. http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf.
10. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics. 2008;24(19):2209–14.
11. Teo Y, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007;23:2741–6.
12. Steemers FJ, et al. Whole-genome genotyping with the single-base extension assay. Nat Methods. 2006;3(1):31–3.
13. Everitt BS, Hand DJ. Finite mixture distributions. London: Chapman & Hall/CRC; 1981. ISBN 0-412-22420-8.
14. Lindsay BG, Vol. 5. Mixture Models: Theory, Geometry, and Applications. Hayward: Institute of Mathematical Statistics; 1995.
15. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. "Section 16.1. Gaussian Mixture Models and k-Means Clustering". Numerical Recipes: The Art of Scientific Computing, 3rd ed. New York: Cambridge University Press; 2007. ISBN 978-0-521-88068-8.
16. Ferguson TS. A Bayesian analysis of some nonparametric problem. Ann Stat. 1973;1(2):209–30.
17. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat. 1974;2(6):1152–74.
18. Sethuraman J. A constructive definition of Dirichlet priors. Stat Sin. 1994;4:639–50.
19. Kim S, Smyth P, Stern H. A Bayesian mixture approach to modeling spatial activation patterns in multisite fMRI data. IEEE Trans Med Imaging. 2010;29(6):1260–74.
20. Teh YW, Jordan MI, Beal MJ, Blei DM. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. Advances in Neural Information Processing Systems. 2005;17:1385–92.
21. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet Processes. J Am Stat Assoc. 2006;101(476):1566–81.
22. Li GX, Gelernter J, Kranzler HR, Zhao HY. M^3: an improved SNP calling algorithm for Illumina BeadArray data. Bioinformatics. 2012;28(3):358–65.
23. McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley Series in Probability and Statistics; 2000.
24. Teh YW. Dirichlet Process. Technical Report: University College London.