BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1714–1720
doi:10.1093/bioinformatics/btu107

Genetics and population analysis

Advance Access publication February 23, 2014

iCall: a genotype-calling algorithm for rare, low-frequency and common variants on the Illumina exome array

Jin Zhou[1], Erwin Tantoso[2], Lai-Ping Wong[2], Rick Twee-Hee Ong[2], Jin-Xin Bei[3], Yi Li[3], Jianjun Liu[3], Chiea-Chuen Khor[2,3] and Yik-Ying Teo[1,2,3,4,5,*]

[1]Department of Statistics and Applied Probability, [2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, [3]Genome Institute of Singapore, Singapore, [4]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore and [5]Life Sciences Institute, National University of Singapore, Singapore

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Next-generation genotyping microarrays have been designed with insights from 1000 Genomes Project and whole-exome sequencing studies. These arrays additionally include variants that are typically present at lower frequencies. Determining the genotypes of these variants from hybridization intensities is challenging because there is less support to locate the presence of the minor alleles when the allele counts are low. Existing algorithms are mainly designed for calling common variants and are notorious for failing to generate accurate calls for low-frequency and rare variants. Here, we introduce a new calling algorithm, iCall, to call genotypes for variants across the whole spectrum of allele frequencies.

Results: We benchmarked iCall against four of the most commonly used algorithms, GenCall, optiCall, illuminus and GenoSNP, as well as a post-processing caller zCall that adopted a two-stage calling design. Normalized hybridization intensities for 12 370 individuals genotyped on the Illumina HumanExome BeadChip were considered, of which 81 individuals were also whole-genome sequenced. The sequence calls were used to benchmark the accuracy of the genotype calling, and our comparisons indicated that iCall outperforms all four single-stage calling algorithms in terms of call rates and concordance, particularly in the calling accuracy of minor alleles, which is the principal concern for rare and low-frequency variants. The application of zCall to post-process the output from iCall also produced marginally improved performance to the combination of zCall and GenCall.

Availability and implementation: iCall is implemented in C++ for use on Linux operating systems and is available for download at http://www.statgen.nus.edu.sg/~software/icall.html.

Contact: statyy@nus.edu.sg, zhoujin@nus.edu.sg

Received on October 25, 2013; revised on February 15, 2014; accepted on February 17, 2014

1 INTRODUCTION

Statistical algorithms have automated the process of calling the genotypes in large-scale microarray genotyping where up to 5 million variants can be assayed simultaneously. This process has been predominantly applied to probes that query two possible allelic outcomes at a genomic variant (generically defined as allele A and allele B), and the degree of hybridization to each of the two alleles is reflected by the fluorescence intensity. Translating both sets of allelic hybridization intensities thus allow discrete decisions to be made with respect to whether the genotype of a sample at a particular single nucleotide polymorphism (SNP) should be AA, AB or BB. Occasionally, when the hybridization intensities do not offer sufficient support for one of the discrete calls, a NULL call can be made, which is subsequently treated as missing.

Numerous algorithms have been developed to call genotypes for earlier generations of microarrays on both the Affymetrix and Illumina platforms. They can be broadly classified into five categories: (i) single-sample single-SNP calling such as the Dynamic Model (Di *et al.*, 2005), where only the intensity measurements at each SNP are considered, disregarding whether there were other samples genotyped at the same time; (ii) multi-sample single-SNP calling such as the GenCall (in the proprietary software BeadStudio and GenomeStudio) and illuminus (Teo *et al.*, 2007), where the intensity measurements at each SNP across multiple samples are jointly considered in a cluster analysis framework to learn about genotype cluster characteristics before making the calls; (iii) single-sample multi-SNP calling such as GenoSNP (Giannoulatou *et al.*, 2008), which does not rely on parameters derived from multiple samples; (iv) hybrid calling such as optiCall (Shah *et al.*, 2012), where a prior distribution is generated using multi-sample and multi-SNP data even though the actual calling is performed within each SNP; and (v) linkage disequilibrium-aware calling such as MAMS (Xiao *et al.*, 2007; Yu *et al.*, 2009) and the approach by Browning and colleagues (Browning and Yu, 2009), where the signal intensities from multiple SNPs are jointly evaluated using the correlation structure between SNPs to improve the accuracy of the calls.

Early generations of genotyping microarrays prioritized tagging SNPs identified from the International HapMap Project (Frazer *et al.*, 2007) that are selected on their ability to provide adequate coverage of the human genome in the HapMap populations, and are typically common SNPs with minor allele frequency (MAF) that exceeded 5%. The design of next-generation genotyping microarrays has been guided by the data from the 1000 Genomes Project and other whole-exome sequencing studies to increase genome coverage and to include low-frequency and rare variants (defined as SNPs with MAF

[*]To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

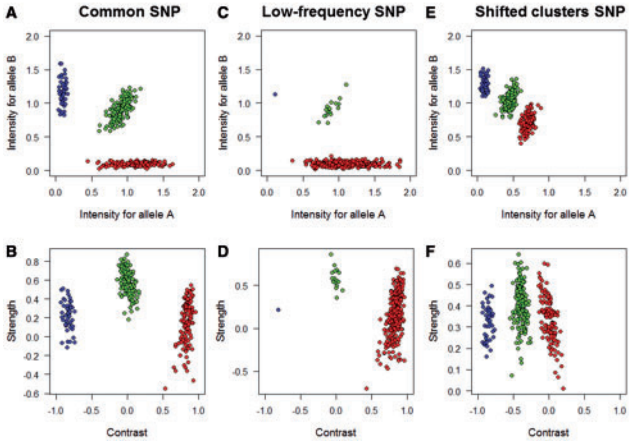

Fig. 1. Illustration of hybridization intensity profiles for three different SNPs at both the allelic intensities axes and the transformed contrast-scale axes. The three SNPs correspond to (i) a common SNP with MAF ≥5% (panels **A** and **B**); (ii) a polymorphic SNP with MAF <5% (panels **C** and **D**); (iii) a common SNP with shifted intensity clusters (panels **E** and **F**). In each panel, the assigned genotypes are colored accordingly as AA (red), AB (green) and BB (blue)

between 1 and 5%, and SNPs with MAF < 1%, respectively) that are often ancestry-specific (Mathieson and McVean, 2012).

Although genotype calling for these next-generation arrays similarly relied on translating the hybridization intensities (Fig. 1A and B), the lower allele frequency spectrum of the majority of these SNPs presents a significantly different challenge, where only a small fraction of the samples is heterozygous and there is usually no homozygous cluster for the alternate allele (Fig. 1C and D). This can thwart algorithms that perform multi-sample calling, as these algorithms, regardless of whether they are maximum likelihood-based or rely on Bayesian models, often set out to locate three genotype clusters. Shifts in the positions of the genotype clusters due to intrinsic hybridization chemistry for a fraction of the SNPs can compound the problem of multi-sample genotype calling (Fig. 1E and F).

In recognition of the challenges associated with calling the genotypes for rare SNPs, a novel 2-stage calling strategy (zCall) was introduced to post-process the genotype calls from a default calling algorithm such as GenCall (Goldstein *et al.*, 2012). This relied on calibrating the positions of the other two genotype clusters on the basis of the dominant homozygous cluster to improve the accuracy and call rate.

In practice, the genotypes for bulk of the SNPs can be accurately determined with straightforward rules in partitioning the distinctively different hybridization intensities. However, SNPs with lower MAF or with shifted intensities will not conform to these simple rules and instead require more sophisticated statistical strategies to accurately call genotypes. Here, we introduce a new genotype calling strategy for Illumina arrays, iCall, which performs multi-sample calling at a single SNP to improve accuracy across the full allele frequency spectrum. This algorithm adopts the classical three-component student's t-mixture model framework that illuminus adopts, but focuses on deriving appropriate penalties to find the best seeding parameters to initialize the expectation–maximization (EM) procedure to recognize the variety of situations where calling becomes difficult, such as when (i) the MAF is low; (ii) the total number of samples for joint calling is small; or (iii) the hybridization intensities deviate substantially from usual. Our method iCall is benchmarked against four of the most commonly used single-stage algorithms (optiCall, illuminus, GenoSNP and GenCall) as well as different iterations of two-stage calling with zCall, with a large dataset of 12 370 samples that have been genotyped on the Illumina exome chip, where 81 samples have been additionally whole-genome sequenced on Illumina HiSeq 2000 at a deep coverage of 30-fold. The sequencing calls for these samples were used to derive a set of gold standard calls for calculating calling accuracy. Our algorithm has been implemented in a C++ program that is available for public access.

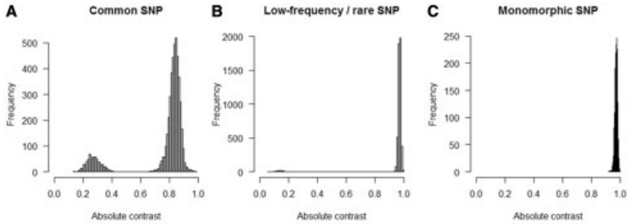

Fig. 2. Histograms of the absolute value of the contrast coordinates for 12 370 samples at three SNPs with different MAFs, corresponding to a (**A**) common SNP (MAF ≥ 5%); (**B**) low-frequency or rare SNP (0% < MAF < 5%); and (**C**) monomorphic SNP (MAF = 0%)

2 METHODS

2.1 Normalized intensity data

As with optiCall and illuminus, iCall uses the normalized hybridization intensities for the respective two alleles at each SNP that is generated from the proprietary software GenomeStudio as the input. We generally define the two alleles as A and B, and let (x_j, y_j) denote the normalized intensities for sample j at a specific SNP. iCall transforms the normalized intensities to the contrast-strength coordinate system (c_j, s_j), where the contrast c_j and strength s_j for sample j are defined, respectively, as

$$c_j = \frac{x_j - y_j}{x_j + y_j} \qquad \text{and} \qquad s_j = \log(x_j + y_j).$$

2.2 Identifying the parameters to initialize calling

The performance of the genotype calling can depend crucially on the set of initial calls used to seed the algorithm, especially if the mathematical framework for initializing the calls is similar to the framework for subsequent calling. For instance, if the initial set of calls already assumes the presence of only one genotype cluster, subsequent iterations of a calling algorithm will usually remain within the same domain space unless the empirical data provide a strong motivation to introduce additional genotype clusters. iCall adopts a separate framework to generate the initial set of calls by considering the information presented by the absolute contrast measurements, or $|c_j|$. When considered across multiple samples, the density profile of the absolute contrast can inform the potential locations of each genotype cluster (Fig. 2). A common SNP will usually yield a density profile with two distinct peaks (around 0 and 1 for the absolute contrast respectively), whereas a rare or low-frequency SNP will give a

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

profile with a small peak near 0 and a significantly larger peak around 1, and a monomorphic SNP will yield only one peak around 1. To model this, we consider two scenarios: (i) the first assumes a normal distribution for $|c_j| \sim \text{Normal}(\mu, \sigma^2)$, and this aims to capture the situation when the SNP is monomorphic; (ii) the second aims to identify the situation for a non-monomorphic SNP and assumes a two-component normal mixture model for $|c_j|$ such that

$$|c_j| \sim p \cdot \text{Normal}\left(\mu_1, \sigma_1^2\right) + (1-p) \cdot \text{Normal}\left(\mu_2, \sigma_2^2\right)$$

with $0 \leq \mu_1 < \mu_2 \leq 1$, and all the parameters $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ are estimated from the data within an EM algorithm framework. The first scenario is actually a special case of the second scenario where $p = \mu_1 = \sigma_1^2 = 0$. To identify which scenario is more appropriate for the observed data, we introduce a penalty as a function of $(\mu_2 - \mu_1)$ to calculate the penalized log-likelihoods for both scenarios (where in scenario 1, $\mu_1 = 0$, $\mu_2 = \mu$). Otherwise, the two-component mixture model will always yield a higher log-likelihood because of the better fit of the data into two normal distributions with smaller variances.

The penalized log-likelihood functions are calculated as

$$\sum_j \log\left(\phi\left(|c_j|; \mu, \sigma^2\right)\right) + n \cdot S(\mu)$$

and

$$\sum_{\mathrm{j}} \left\{ \log\left(\phi\left(|c|_{j\in\text{class}_1}; \mu_1, \sigma_1^2\right)\right) + \log\left(\phi\left(|c|_{j\in\text{class}_2}; \mu_2, \sigma_2^2\right)\right) \right\} + n \cdot S(\mu_2 - \mu_1)$$

for scenarios 1 and 2, respectively, where the penalty term $S(x) = \log\left(\frac{\psi(x|\text{meanlog}=0.4,\ \text{variancelog}=0.4)}{\int_0^1 \psi(y|\text{meanlog}=0.4,\ \text{variancelog}=0.4)\mathrm{d}y}\right)$, n represents the number of samples used for the joint calling, $\psi(\cdot)$ is the density function of a log-normal distribution with mean and variance of the distribution on the log scale equal to meanlog and varriancelog, and $\phi(\cdot)$ is the density function of a normal distribution. The intuition here is when the values for μ_1 and μ_2 are not significantly different, the calling algorithm prefers to combine the two components instead of forcing the presence of two clusters.

The scenario with the higher log-likelihood is chosen to generate eight sets of location parameters to initialize the genotype calling in a three-component univariate Gaussian mixture model for c_j, where the eight sets are

$$\begin{bmatrix} -\mu & 0 & \mu \\ -1 & 0 & \mu \\ -\mu & 0 & 1 \\ -\mu & \mu & 1 \\ -1 & -\mu & \mu \\ -1 & -0.8 & \mu \\ -\mu & 0.8 & 1 \\ \mathrm{t}_1 & \frac{\mathrm{t}_1+\mathrm{t}_2}{2} & \mathrm{t}_2 \end{bmatrix}$$

if scenario 1 yields the higher log-likelihood, or

$$\begin{bmatrix} -\mu_2 & 0 & \mu_2 \\ -1 & 0 & \mu_2 \\ -\mu_2 & 0 & 1 \\ -\mu_2 & -\mu_1 & \mu_2 \\ -\mu_2 & \mu_1 & \mu_2 \\ -\mu_2 & -\mu_1 & \mu_1 \\ -\mu_1 & \mu_1 & \mu_2 \\ \mathrm{t}_1 & \frac{\mathrm{t}_1+\mathrm{t}_2}{2} & \mathrm{t}_2 \end{bmatrix}$$

if scenario 2 yields the higher log-likelihood, and t_1 and t_2 are chosen from the trimmed empirical distribution of c_j as

$$\mathrm{t}_1 = \frac{Q_c^{0.999} + Q_c^{0.001}}{2} - 0.95 \times \frac{Q_c^{0.999} - Q_c^{0.001}}{2}$$

and

$$\mathrm{t}_2 = \frac{Q_c^{0.999} + Q_c^{0.001}}{2} + 0.95 \times \frac{Q_c^{0.999} - Q_c^{0.001}}{2}$$

where Q_c^x denotes the $100x$ quantile value of the distribution of the empirical c values. This allows the initialization parameters to be guided by the observed values of c_j, which is particularly useful in the situation where the intensities for the genotype clusters are shifted significantly.

2.3 Initializing the genotype calling

In initializing the EM procedure, each of the eight sets of location parameters is used as the centers of a three-component univariate Gaussian mixture model with equal weights for the contrast measurement. The same standard deviation of 0.1 is assumed for the three genotype classes in the first three guided starts, and 0.05 $(Q_c^{0.999} - Q_c^{0.001})$ for the three genotype classes in the other five guided starts. Two penalized log-likelihood functions were calculated to select two sets of parameters among the eight to seed the EM procedure: (i) the first takes the form of $\sum_{k=1}^{3} \sum_{j\in\text{class}_k} \log\left(\phi(c_j^{(k)}; \mu_k, \sigma_k)\right) + n \cdot \frac{\mathrm{S}(\mu_2-\mu_1)+\mathrm{S}(\mu_3-\mu_2)}{2}$, where $S(x) = \log\left(\frac{\psi(x|\text{meanlog}=0.4,\ \text{variancelog}=0.4)}{\int_0^1 \psi(y|\text{meanlog}=0.4,\ \text{variancelog}=0.4)\mathrm{d}y}\right)$, with $\psi(\cdot)$ similarly representing the density function of a log-normal distribution and $\phi(\cdot)$ representing the density function of a normal distribution; (ii) the second penalized log-likelihood function shares the same form as the first function, but has an additional penalty calculated as the log-likelihood of the Hardy–Weinberg equilibrium (HWE) percentile. The intuition behind the two penalty terms is the first term penalizes on small distances between the heterozygous cluster and the two homozygous clusters, whereas the second term penalizes on genotype call configuration at a SNP that deviates further from the state of HWE. Of the eight guided starts, we identify the two guided starts (seed_1, seed_2) that yield the highest penalized log-likelihood without and with the additional penalty on HWE, respectively, and these two guided starts are subsequently used to seed the genotype calling. Note that the two sets of seeding start may be identical if the same guided start yields the highest penalized log-likelihoods in both calculations.

2.4 Genotype calling

Each of the two sets of seeding starts is used to initialize the three-component bivariate truncated t-mixture model that illuminus adopts for $\boldsymbol{x_j} = (c_j, s_j)$, such that

$$F(\boldsymbol{x_j}) = \sum_{k=1}^{3} \lambda_k \phi\left(\boldsymbol{x_j} | M_k, \Sigma_k, \nu_k\right)$$

where

$$\phi_1\left(\boldsymbol{x_j}|M_1, \Sigma_1, \nu_1\right) = \frac{f\left(\boldsymbol{x_j}|M_1, \Sigma_1, \nu_1\right)}{1 - \int_{-\infty}^{-1} f\left(\boldsymbol{x_j}|M_1, \Sigma_1, \nu_1\right) dc}$$

$$\phi_2\left(\boldsymbol{x_j}|M_2, \Sigma_2, \nu 2\right) = \frac{f\left(\boldsymbol{x_j}|M_2, \Sigma_2, \nu_2\right)}{\int_{-1}^{1} f\left(\boldsymbol{x_j}|M_2, \Sigma_2, \nu_2\right) dc}$$

$$\phi_3\left(\boldsymbol{x_j}|M_3, \Sigma_3, \nu_3\right) = \frac{f\left(\boldsymbol{x_j}|M_3, \Sigma_3, \nu_3\right)}{1 - \int_{1}^{\infty} f\left(\boldsymbol{x_j}|M_3, \Sigma_3, \nu_3\right) dc}$$

with $f(\boldsymbol{x_j}|M, \Sigma, \nu)$ denoting the density function for $\boldsymbol{x_j}$ with location parameter M, variance–covariance matrix Σ at ν degrees of freedom. The calling algorithm mimics that of illuminus and similarly adopts an EM framework to yield two sets of genotype call configurations, each initialized from one of the two seeding starts. The default is to accept the genotype call configuration initialized with seed_1, except when the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

evidence against HWE is more significant in the configuration generated by seed$_1$ than the configuration generated by seed$_2$, in which case the genotype calls generated with seed$_2$ are accepted as the final calls. This minimizes the inadvertent miscalling that happens because of shifts in genotype clouds resulting in genotype calls that tend to deviate from HWE. As with illuminus, the posterior probability of each genotype class for a sample is calculated after considering the intensity profile of the sample relative to those of all available samples, and a valid genotype is assigned if the posterior probability exceeds a predetermined threshold (iCall uses a default of 0.8). If the posterior probability for the NULL category exceeds the threshold, or if none of the posterior probabilities exceed the threshold, a NULL genotype call is assigned and this is conventionally treated as a missing value.

2.5 Chromosomes X, Y and mitochondria

For calling genotypes at SNPs on the mitochondria and the non-pseudo-autosomal regions of the sex chromosomes, the genotype calling additionally requires information on the gender of each sample that determines the direction of hybridization inactivation. For SNPs on chromosome X, genotypes for females are determined in the same fashion as autosomal SNPs, whereas the genotypes for males will only be called as either AA or BB. For SNPs on chromosome Y, NULL calls will be produced for females and the calling only considers the intensity data for male samples and similarly yields genotype calls of either AA or BB. The situation is reversed for SNPs on the mitochondria, where NULL calls will be produced for males and the calling only considers the intensity data for female samples and produces calls of either AA or BB.

3 RESULTS

The performance of iCall was compared against four single-stage genotype calling algorithms: GenCall, optiCall, illuminus and GenoSNP. Intensity data were available for 12 370 samples that have been genotyped on the Illumina exome chip, of which 348 samples came from the Singapore Integrative Omics Project (iOmics) and 12 022 samples came from multiple complex disease studies that are being carried out at the Genome Institute of Singapore.

To compare the performance of different genotype calling algorithms, we need to derive a set of gold standard calls that we subsequently assumed to be perfect for benchmarking the genotype calls made by different algorithms. Of the 348 iOmics samples, 81 samples have been additionally whole-genome sequenced to a target coverage of 30-fold as part of the Singapore Sequencing Studies (http://www.statgen.nus.edu.sg/), and the sequence calls after quality checks for these samples were regarded as the gold standard calls that were subsequently used to benchmark the performance of the different methods. A total of 16 428 SNPs were present on the exome chip that overlapped with the polymorphic variants identified from the high-coverage sequencing. These SNPs were classified as common ($\text{MAF} \geq 5\%$, 13 542 SNPs), low frequency ($1\% \leq \text{MAF} < 5\%$, 1356 SNPs) and rare ($\text{MAF} < 1\%$, 1530 SNPs) according to the GenCall genotypes for all 12 370 samples.

To evaluate how the number of samples available for joint calling impacts the algorithms, we thinned the dataset into four smaller sets with 500, 1000, 3000 and 5000 samples, which always included the 81 samples with gold standard calls. Note that GenCall genotypes were available for the 348 iOmics samples and 12 022 samples independently, and GenoSNP is a single-sample caller where the performance is not affected by the size of the available samples.

The performance of iCall, GenCall, optiCall, illuminus and GenoSNP is evaluated using five metrics: (i) call rate, defined as the percentage of valid genotype calls that are not assigned as NULL; (ii) concordance, defined as the percentage of valid genotype calls that are identical to the gold standard calls; (iii) overall concordance, defined as the percentage of genotype calls out of all possible calls that are identical to the gold standard calls, and is calculated as the product of the call rate and the concordance; (iv) minor allele concordance for rare and low-frequency SNPs, defined as the percentage of the heterozygous and minor allele-homozygous calls that are identical to the gold standard calls out of the total number of such calls made for these SNPs; and (v) missed minor allele call rate, defined as the percentage of the heterozygous and minor-allele homozygous calls that are not identified out of the total number of available minor allele calls in the gold standard. The last two metrics effectively evaluate the true-positive and false-negative rates for making a genotype call involving at least one minor allele at a rare or low-frequency SNP. The calculations of all five metrics are made using only the 81 samples for which there are gold standard calls available.

On the basis of call rates and concordance with the gold standard calls, iCall yielded the highest overall concordance rate and call rate regardless of the sample size (Table 1). We observed that GenCall yielded the highest concordance rate but tend to be more conservative at making calls, but still managed to deliver an overall concordance rate that was consistently higher than the performance by optiCall. The performance of illuminus and GenoSNP were comparatively less satisfactory, with GenoSNP yielding an overall concordance rate that was below 97%.

When evaluating the ability to correctly call genotypes carrying at least one copy of the minor allele that is present in the dataset at a frequency <5%, iCall consistently yields the highest accuracy and the lowest missed allele calls compared with GenCall, optiCall, illuminus and GenoSNP at low-frequency SNPs (Table 1). For example, iCall achieved a minor allele concordance rate of 97.140 and 97.168% at the sample sizes of 500 and 12 370, respectively, compared with optiCall at 96.932 and 97.033%, respectively, and the next-best performing algorithm (GenCall) at 97.083% on the basis of 348 samples. At rare SNPs, iCall similarly delivered the highest minor allele concordance rates across all sample sizes considered (at least 97.435%, with all other methods delivering concordance <97%). This suggests that whenever iCall made a call involving a minor allele, it was more likely to be correct than existing algorithms.

However, a high minor allele concordance can be achieved by a conservative algorithm that only calls the easy-to-call minor allele genotypes but misses out on most of the genuine minor allele calls. We additionally evaluated the extent that each caller is missing genuine minor allele calls. For low-frequency SNPs, iCall consistently exhibited the lowest missed minor allele call rate (with a maximum of 2.915%), compared with 2.958% for GenCall and 3.029% for optiCall with 12 370 samples. However, for rare SNPs, iCall was more conservative and made less minor allele genotype calls than optiCall, especially when the sample size is large (missed minor allele call rate of 3.280 and 2.967% for

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Comparison of iCall against optiCall, illuminus, GenCall and GenoSNP at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark

				Low-frequency SNPs				Rare SNPs			
Sample size	Call rate (%)	Concordance	Overall concordance (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)
iCall											
500	99.993	97.683	97.676	13 653	14 055	97.140	2.915	6161	6316	97.546	3.296
1000	99.990	97.683	97.673	13 653	14 055	97.140	2.915	6153	6315	97.435	3.422
3000	99.990	97.685	97.675	13 657	14 054	97.175	2.887	6163	6316	97.578	3.265
5000	99.988	97.685	97.673	13 656	14 055	97.161	2.894	6163	6316	97.578	3.265
12 370	99.986	97.686	97.672	13 658	14 056	97.168	2.880	6162	6317	97.546	3.280
GenCall											
348	99.983	97.688	97.671	13 647	14 057	97.083	2.958	6113	6315	96.801	4.050
optiCall											
500	99.987	97.667	97.654	13 617	14 048	96.932	3.171	6160	6378	96.582	3.312
1000	99.988	97.665	97.653	13 623	14 078	96.768	3.129	6175	6448	95.766	3.076
3000	99.985	97.675	97.660	13 621	14 044	96.988	3.143	6179	6463	95.606	3.014
5000	99.987	97.681	97.668	13 625	14 048	96.989	3.115	6180	6418	96.292	2.998
12 370	99.990	97.681	97.662	13 637	14 054	97.033	3.029	6182	6565	94.166	2.967
Illuminus											
500	99.805	97.652	97.462	13 036	14 203	91.783	7.303	5043	6436	78.356	20.844
1000	99.834	97.650	97.488	13 254	14 067	94.221	5.753	5120	6587	77.729	19.636
3000	99.873	97.645	97.521	13 496	14 052	96.043	4.032	5187	6700	77.418	18.584
5000	99.862	97.651	97.516	13 500	14 048	96.099	4.003	5035	6696	75.194	20.970
12 370	99.848	97.661	97.513	13 563	14 059	96.472	3.555	4772	6542	72.944	25.098
GenoSNP											
Single SNP	99.607	96.734	96.354	13 349	15 267	87.437	5.077	6059	7583	79.902	4.897

Among the 16 428 SNPs considered, 13 542 are common SNPs, 1356 are low-frequency SNPs and 1530 are rare SNPs. Within the gold standard, there are 1 222 885 valid genotype calls in total, which include 14 063 minor allele calls at low-frequency SNPs and 6371 minor allele calls at rare SNPs.

iCall and optiCall, respectively) although the genotype calls by iCall are much more likely to be correct (concordance of 97.546% by iCall versus 94.166% by optiCall). As the number of samples available for joint calling increases, optiCall appears to be more liberal at making minor allele calls, whereas iCall appears to be stable. GenCall, illuminus and GenoSNP consistently performed poorly when measured with these two minor allele metrics.

zCall is a post-processing caller that uses intensities and genotypes generated from a stand-alone caller as input data. We also compared the performance of GenCall+zCall, optiCall+zCall and iCall+zCall (Table 2). The results indicated that zCall always improved the genotype calls generated from all the three callers, with the greatest degree of improvement observed for GenCall genotypes. Across all SNPs, GenCall+zCall yielded marginally higher overall concordance rates (97.684 and 97.683% at sample sizes of 500 and 12 370, respectively) compared with those by iCall+zCall (97.681 and 97.683%, respectively). However, at low-frequency and rare SNPs, iCall+zCall outperformed GenCall+zCall, delivering higher concordance rates and lower missed call rates for minor allele genotypes. optiCall+zCall exhibited the same characteristics as optiCall, where it is more aggressive in calling minor allele genotypes but at the expense of making more erroneous calls.

4 DISCUSSION

We have introduced iCall, a method for calling genotypes that yields comparatively better performance than existing genotype calling algorithms, particularly in accurately calling the genotypes involving minor alleles at low-frequency or rare SNPs. One important aspect of genotype calling is that determining the genotypes accurately is straightforward for the majority of the SNPs, but there are SNPs where the MAF are considerably lower or when the hybridization profiles differ from the usual that require more robust considerations to accurately determine the genotypes. Our method improves on the framework of illuminus by using a series of penalty functions to identify the optimum parameters to seed the EM model. The availability of a large dataset that has been genotyped on the exome chip meant that we can evaluate the performance of existing algorithms across different sample sizes.

We have benchmarked the genotype calls obtained from different methods against a set of gold standard calls that was derived from deep sequencing. As a stand-alone caller, iCall performs the best in terms of delivering the most accurate genotype calls while minimizing the number of missed calls, particularly for genotypes involving minor alleles at low-frequency and rare SNPs. The better performance at low-frequency and rare SNPs

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 2. Comparison of iCall+zCall, GenCall+zCall and optiCall+zCall at 16 428 SNPs at different sample sizes for calling, where genotypes from whole-genome sequencing of 81 samples are used as benchmark

				Low-frequency SNPs				Rare SNPs			
Sample size	Call rate (%)	Concordance	Overall concordance (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)	Correct minor allele calls	Minor allele calls	Minor allele concordance rate (%)	Missed minor allele call rate (%)
iCall+zCall											
500	99.999	97.682	97.681	13 653	14 054	97.147	2.915	6167	6315	97.656	3.202
1000	99.998	97.683	97.681	13 654	14 055	97.147	2.908	6167	6315	97.656	3.202
3000	99.998	97.684	97.682	13 660	14 054	97.197	2.866	6167	6315	97.656	3.202
5000	99.998	97.684	97.682	13 658	14 055	97.175	2.880	6167	6315	97.656	3.202
12 370	99.998	97.685	97.683	13 661	14 056	97.190	2.859	6167	6315	97.656	3.202
GenCall+zCall											
500	99.998	97.685	97.684	13 654	14 054	97.154	2.908	6163	6314	97.608	3.265
1000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6162	6313	97.608	3.280
3000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
5000	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
12 370	99.998	97.685	97.683	13 654	14 054	97.154	2.908	6161	6312	97.608	3.296
optiCall+zCall											
500	99.999	97.660	97.659	13 628	14 040	97.066	3.093	6170	6369	96.875	3.155
1000	99.999	97.659	97.658	13 629	14 065	96.900	3.086	6186	6438	96.086	2.904
3000	99.999	97.672	97.671	13 626	14 041	97.044	3.107	6189	6414	96.492	2.857
5000	99.999	97.678	97.676	13 628	14 045	97.031	3.093	6190	6377	97.068	2.841
12 370	99.998	97.670	97.668	13 641	14 052	97.075	3.001	6189	6521	94.909	2.857

Among the 16 428 SNPs, 13 542 are common SNPs, 1356 are low-frequency SNPs and 1530 are rare SNPs. Within the gold standard, there are 1 222 885 valid genotype calls in total, 14 063 minor allele calls at low-frequency SNPs and 6371 minor allele calls at rare SNPs.

was similarly observed when iCall was incorporated as part of a two-stage calling process with zCall.

We have compared iCall against existing methods using two additional metrics that specifically focused on the ability to call the genotypes that involved at least one minor allele at rare and low-frequency SNPs. This is in line with the intended purpose of the exome microarray for finding low-frequency or rare SNPs that are associated with phenotypes. Measuring how accurately and sensitively a calling algorithm can call a heterozygous or minor allele-homozygous genotype is thus more important. After all, an algorithm that erroneously calls a rare SNP as major-allele monomorphic will have attained a concordance of at least 98%. In quantifying the association evidence at rare or low-frequency SNPs, it is common to pool allele counts across similar SNPs in a genomic region to assess allelic burden (Ionita-Laza *et al.*, 2013; Neale *et al.*, 2011; Wu *et al.*, 2010, 2011). Erroneously calling the presence of a minor allele genotype, or the failure to call a minor allele genotype when it exists can thus directly impact the power and false-positive rate of the association analyses.

Automated algorithms for calling genotypes have contributed to the success of large-scale genomic studies, and this is likely to continue with the continuous introduction of next-generation genotyping microarrays designed with knowledge gained from large-scale sequencing studies, querying up to 5 million SNPs across the genome or variants found specifically in the exons. Although these technologies provide the opportunity to investigate new hypotheses on the evolution of the human genome and the genetic etiology of diseases and traits, this can only happen if the content in the human genome can be accurately determined. We have introduced a calling algorithm that provides a better ability at accurately calling genotypes for rare and low-frequency SNPs, and consistently performs well at common SNPs.

Funding: This project acknowledges the support of the Saw Swee Hock School of Public Health, the Faculty of Science, the Yong Loo Lin School of Medicine, the National University Health System, the Life Science Institute and the Office of Deputy President (Research and Technology) from the National University of Singapore. L.P.W., R.T.H.O. and Y.Y.T. additionally acknowledge support from the National Research Foundation Singapore (NRF-RF-2010-05).

Conflict of Interest: none declared.

REFERENCES

Browning,B.L. and Yu,Z. (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.*, **85**, 847–861.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

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

Frazer,K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Giannoulatou,E. *et al.* (2008) GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics*, **24**, 2209–2214.

Goldstein,J.I. *et al.* (2012) zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics*, **28**, 2543–2545.

Ionita-Laza,I. *et al.* (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.*, **92**, 841–853.

Mathieson,I. and McVean,G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.

Neale,B.M. *et al.* (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.

Shah,T.S. *et al.* (2012) optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics*, **28**, 1598–1603.

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

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Xiao,Y. *et al.* (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **23**, 1459–1467.

Yu,Z. *et al.* (2009) Genotype determination for polymorphisms in linkage disequilibrium. *BMC Bioinformatics*, **10**, 63.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016