ARTICLE

Optimal genotype determination in highly multiplexed SNP data

Martin Moorhead¹, Paul Hardenbol¹, Farooq Siddiqui¹, Matthew Falkowski¹, Carsten Bruckner¹, James Ireland¹, Hywel B Jones¹, Maneesh Jain¹, Thomas D Willis¹ and Malek Faham^{*,1}

¹ParAllele BioScience, South San Francisco, CA, USA

High-throughput genotyping technologies that enable large association studies are already available. Tools for genotype determination starting from raw signal intensities need to be automated, robust, and flexible to provide optimal genotype determination given the specific requirements of a study. The key metrics describing the performance of a custom genotyping study are assay conversion, call rate, and genotype accuracy. These three metrics can be traded off against each other. Using the highly multiplexed Molecular Inversion Probe technology as an example, we describe a methodology for identifying the optimal trade-off. The methodology comprises: a robust clustering algorithm and assessment of a large number of data filter sets. The clustering algorithm allows for automatic genotype determination. Many different sets of filters are then applied to the clustered data, and performance metrics resulting from each filter set are calculated. These performance metrics relate to the power of a study and provide a framework to choose the most suitable filter set to the particular study.

European Journal of Human Genetics (2006) 14, 207–215. doi:10.1038/sj.ejhg.5201528; published online 23 November 2005

Keywords: high throughput genotyping technologies; association study; genotype calling; genotyping conversion rate; genotyping call rate; genotyping Mendelian concordance

Introduction

Large-scale association studies to elucidate the genetic basis of common disease have been advocated for years.¹ High-throughput custom genotyping technologies are now available to achieve this goal.^{2–5} These genotyping technologies generate data on an unprecedented scale. Given the size of the data sets generated, automatic and robust algorithms for genotype determination are necessary. Ultimately, these technologies are evaluated by three key performance metrics: assay conversion, call rate, and genotype accuracy. Assay conversion rate is the fraction of all attempted markers that successfully generate geno-

types. Call rate is the proportion of samples for which genotypes are called for a converted marker. Genotype accuracy is the fraction of calls that are correct in that the alleles of both chromosomes are correctly determined.

These performance metrics change with the genotype filtering criteria, ultimately affecting the power of a genetic study. Low stringency filtering criteria include markers that produce marginal data and as a result pass incorrect genotypes, while excessively stringent clustering parameters results in throwing out potentially useful markers. In some studies, it may be that a higher conversion rate is preferable even at the expense of lower accuracy, or *vice versa*.

There is little material published describing genotype calling algorithms for these high-throughput technologies⁶ or assessing the impact of their performance metrics on the power of association studies. Herein, we address this in the context of the analysis of Molecular Inversion Probe (MIP) technology data.

*Correspondence: Dr M Faham, ParAllele Bioscience, 7300 Shoreline Court, South San Francisco, CA, USA. Tel: +1 650 228 0354;

Fax: +1 650 228 7437; E-mail: malek@p-gene.com

Received 29 March 2005; revised 29 September 2005; accepted 12 October 2005; published online 23 November 2005

MIP is a high-throughput genotyping technology capable of genotyping more than 10 000 SNPs in one reaction.^{2,4} In order to analyze this tremendous volume of data, we have implemented an algorithm based on an expectation–maximization (E/M) clustering algorithm to call the genotypes automatically. This algorithm may be relevant to any method where the genotype determination is carried out through assessment of the relative intensity of signals from the two SNP alleles.

Since the data sets generated by MIP are large enough to be effectively resistant to overfitting, an additional analysis module (multifiltering) was designed to apply many sets of parameter filters to a given data set. Performance metrics are calculated for each of these sets. This process then transforms the complex multidimensional space of many interacting filters into the simpler three dimensional space of performance metrics. This mapping is specific to each custom set of SNPs. The three performance metrics relate to the power of a study. Using the perspective of the power of a study provides a reasonable guidance for determining the appropriate genotyping filter set to use in a specific study.

Methods

Clustering algorithm

Raw chip data for each sample is first normalized as described in Supplementary Figure S1. The E/M clustering algorithm is then applied on a marker-by-marker basis, that is, each marker is clustered independently (across all samples). For each marker, there are N data points to be clustered, where N is given by the number of samples measured. Each of these data points is really a vector of four numbers given by the signals in the four color channels for the different alleles (A, C, G, and T). We label the four channels according to the alleles of the SNP being clustered.

1. S_1 : The (alphabetically) first SNP allele, for example, C for a C/T SNP.
2. S_2 : The (alphabetically) second SNP allele, for example, T for a C/T SNP.
3. B_1 : The (alphabetically) first non-SNP allele, for example, A for a C/T SNP.
4. B_2 : The (alphabetically) second non-SNP allele, for example, G for a C/T SNP.

The transformation of the two-dimensional (2D) SNP allele signal space (S_1 vs S_2) into the other 2D space, S_{sum} vs x (contrast), is carried out using the following equations:

$$S_{\text{sum}} = S_1 + S_2$$

$$y = (2S_2/S_{\text{sum}})^{-1},$$

$$x = \sinh(2y)/\sinh(2).$$

After this transformation, we applied an E/M clustering algorithm.⁷ The actual clustering takes place in the x

(contrast) dimension only. Note that the sinh function used in the contrast space transformation is motivated by the desire to both increase the width of the homozygous clusters (usually found at $x=1$ and -1) so that they are more similar to the heterozygous cluster (usually found at $x=0$) and to make the homozygous cluster more normally distributed. Other transformations were investigated and found to be less useful.

We fit the N samples into K clusters (1, 2, or 3) represented by Gaussian functions. For the i th sample, we can write its un-normalized probability of belonging to the j th cluster as

$$f_{ij} = (\lambda_j/\sigma_j) \exp(-1/2((x_i - m_j)/\sigma_j)^2),$$

where m_j , σ_j and λ_j are, respectively, the mean, sigma, and weight of the j th cluster. The normalized probability of the i th sample belonging to the j th cluster can then be written as

$$z_{ij} = f_{ij} / \sum_{j=1}^K f_{ij}$$

so that $\sum_{i=1}^N \sum_{j=1}^K z_{ij} = N$, the number of samples.

Calculating these normalized probabilities of samples belonging to clusters in terms of the cluster parameters (m_j , σ_j , and λ_j) is referred to as the E step of the E/M algorithm. It is followed by the M step whereby new values for the cluster parameters are calculated from the z_{ij} values of the previous E step as follows:

$$\lambda_j = (1/N) \sum_{i=1}^N z_{ij},$$

$$m_j = (1/N\lambda_j) \sum_{i=1}^N x_i z_{ij},$$

$$\sigma_j^2 = (1/N\lambda_j) \sum_{i=1}^N (x_i - m_j)^2 z_{ij}.$$

Successive iterations of E + M steps are completed until convergence is reached as defined by the cluster parameters no longer changing by an amount greater than a predefined tolerance for the calculation (in our case 1.0E–5).

When three clusters are being fit to the data, there are effectively eight free parameters in the fit (the constraint $\sum_{j=1}^K \lambda_j = 1$ removes a free parameter). Empirically derived additional constraints (boundaries) are imposed on the means (m) and standard deviations (σ) of the three clusters; these constraints are particularly important for the identification of rare alleles. Initial values for the eight free parameters must be specified before the algorithm can be executed. For most markers, the E/M algorithm will converge appropriately using the simplest set of initial values as determined by the average properties of the clusters for ‘normal’ SNPs, that is, $m_1 = -1$, $m_2 = 0$, $m_3 = 1$, and $\lambda_{1,2,3} = 0.05$. However, there are some excep-

tions. In particular, some SNPs are monomorphic (one cluster) or have a sufficiently rare minor allele that there are no homozygous samples for the rare allele (in this case there are two clusters). These one or two cluster SNPs have been fit by starting the E/M algorithm with three clusters and then automatically 'dropping out' any cluster that is not supported by the data, that is, whose weight falls below $0.8/N$. If a cluster is not present in the data, it will drop out very quickly (usually in the first iteration) before it has a chance to drift across into the tail of another cluster. Figure 1c shows an example of a SNP that has been automatically fit with two clusters.

For some SNPs, the heterozygous cluster is not found at the center of contrast space ($x=0$, where the two allele signals are equal) but can be offset towards one or the other homozygous clusters. In this case, the E/M algorithm sometimes converges on a local minimum and does not reach the appropriate global minimum. We have developed an automated approach, which almost always converges on the global minimum. We perform 40

different E/M fits, each with different seed values that are chosen to cover the entire spectrum of cluster parameters found across large data sets. From these 40 fits, the best-fit solution is then chosen using a simple log-likelihood metric

$$L = \sum_{i=1}^N \log(f_{ij}),$$

where for each sample we use the f_{ij} probability with respect to the most likely cluster for that sample. Figure 1d shows an extreme example of a marker with a highly offset heterozygous cluster that has been correctly fit using our automated method.

Call confidence

The minimum probability ratio parameter determines the minimum allowed ratio of probabilities, for a data point, between belonging to the closest cluster and the next closest cluster. A secondary parameter, cluster probability, specifies the minimum probability allowed with respect to a data point's main cluster. We have found that minimum

Figure 1 Clustering plots for single markers. The transformation from (a) signal space (normalized signal 1, normalized signal 2) to (b) contrast space (S_{sum} , x) is shown for a typical marker with a well-centered heterozygous cluster at $x=0$ and homozygous clusters at $x=-1$ and $x=1$. (c) A marker with no rare homozygotes that has been automatically fit with two clusters using the adaptive features of our E/M clustering algorithm that drops out of the fit clusters that are not supported by the data. (d) Depicts a marker with a highly shifted heterozygous cluster that has been correctly fit through our automated strategy of using a wide range of seed conditions in 40 independent E/M cluster fits and then choosing the fit with highest log-likelihood. In all four plots, the genotype calls have been indicated by color as follows: red = homozygous in allele 1; green = homozygous in allele 2; blue = heterozygous; and black = no call.

probability ratio is a much better discriminator than cluster probability in that it removes inaccurate calls at a much smaller cost in terms of call rate and conversion. We have also combined these two concepts of probability ratio and direct probability into a single metric

$$\text{Call Confidence} = P_{\text{ratio}} \exp(-\alpha/2((x_i - m_1)/\sigma_1)^2),$$

where

$$P_{\text{ratio}} = (z_{i1}/z_{i2})/((z_{i1}/z_{i2}) + 1)$$

is used to transform probability ratio into a number that behaves like probability in that it has the range 0→1 and where the subscripts 1 and 2 refer to the most likely and second most likely cluster for the *i*th sample. The parameter α in the equation for call confidence is used to control the relative importance of probability (of being in the closest cluster) vs probability ratio (closest/next closest cluster). In practice, we set it at a value of 0.01 because we want call confidence to be mostly determined by the probability ratio since it is a better indicator of accuracy than probability.

This call confidence is computed for each call. The 20th percentile call confidence of all the calls for each marker is also calculated. This value, which is a property of the marker, is abbreviated as marker confidence.

Mendelian concordance

We calculate the total number (summing over all parents–child trios and all passed markers) of child chromosomes that are concordant or discordant with Mendelian inheritance law and then define Mendelian concordance as the number of concordant calls/(number of concordant + discordant calls).

Power calculation

For the power calculation, we assumed a multiplicative model for the interaction between the alleles. In this model the risk of individuals carrying no risk alleles, the risk allele in a heterozygous state, and the risk allele in a homozygous state are f_0 , $f_0 \gamma$, and $f_0 \gamma^2$, respectively, where γ is the genotypic relative risk (GRR). Power was calculated based on standard methods.⁸ For the calculation of the power given genotyping error, we assumed that the error is equivalent in both directions. In the presence of error rate e , a marker with minor allele frequency p (in the combined cases and controls) will have an observed allele frequency p^* , where

$$p^* = p(1 - e) + e(1 - p).$$

We treated the observed and true calls of an SNP A as if they are from two different SNPs A and B with linkage disequilibrium (LD) between them. We calculated r^2 between SNPs A and B using the following equation:

$$r^2 = (p(1 - e)pp^*)^2 / pp^*(1 - p)(1 - p^*).$$

We calculated the relative increase in sample size (from what is needed with perfect genotyping) by the inverse

of r^2 .⁹ Results from this method of estimating the effect of error is similar to estimates using previously used methods.¹⁰

We assumed that the true error rate is three times the observed Mendelian concordance.¹¹ This estimation depends on the allele frequency of the studied markers and is therefore valid as long as the study is not significantly biased towards rare or highly polymorphic SNPs. We do note that there are some systematic errors like cluster dropout that will be missed by Mendelian concordance measures.

Results

Genotype determination

A typical MIP genotyping study includes hundreds to thousands of DNA samples measured across up to 10 000 markers, and hence the total number of data points that need to be analyzed is on the order of 10^5 – 10^8 . We have implemented an algorithm that requires no manual intervention at the detailed level of a single call, sample, or marker. The algorithm performs data normalization followed by the application of a clustering algorithm for genotype calling as described in the Methods section.

Ultimately, calling genotypes is controlled by filters that are applied to the data as whole, that is, are applied identically to each marker and all its calls. These filters are either at the marker level where a marker is deemed invalid or at the genotype level where a specific sample is not called for a particular marker. Some of these filters are applied before using the clustering process in order to remove poor quality data that might confound it. These preclustering filters are at the genotype level and include the elimination of data from saturated features, features in blemished areas on the array, and features with low signal to noise.

Most of the filters are applied after the application of the clustering process. Individual genotype calls are identified as ‘no-calls’ based on a set of filters, the most important of which relates to the call confidence, which is described in the Methods section. After applying filters at the genotype level, the marker as a whole is evaluated, and either passed or failed. The marker is deemed ‘invalid’ if it does not meet all of the marker acceptance criteria. The marker confidence value (described in the Methods section) of a marker is a property of the marker and can be used to filter out markers that have values below a specific threshold. The other main marker acceptance criteria are a minimum call rate, an acceptable number of heterozygous calls given the allele frequency, and reasonable signal variance within each cluster of the marker.

The fraction of markers that pass the marker filters described above is referred to as the conversion rate of the marker panel. Using only these passed markers, we then calculate two other performance metrics for the panel.

Data call rate is the number of genotypes called across all the passed markers divided by the number of genotypes attempted ($\# \text{ Samples} \times \# \text{ Passed Markers}$). Concordance with Mendelian inheritance is computed using mother–father–child trio samples. Mendelian concordance is used as a surrogate for accuracy.

Multifiltering

For any MIP marker panel, we can obtain a variety of possible outcomes for the three performance metrics outlined by adjusting the filters described in the above section. These filters interact in complex ways since calls can fail more than one filter and call filters can affect marker filters. For example, if enough calls are filtered out due to insufficient call confidence, the marker is filtered out since the minimum call rate filter is engaged. To explore the filter parameter space in detail, we apply more than 1000 filter sets to the data and calculate performance metrics for each filter set. In this data set, a range of assay conversion of 81.1–87.8%, a range of Mendelian concordance of 99.55–99.90%, and a range of call rate of 95.9–99.8% were obtained. Figure 2 shows the trade-off between the three performance metrics resultant from the application of multiple interacting filters for this specific data set.

Given the trade-off, one can pick filter sets that provide better data in the study of interest. The validity of this approach is driven by the large number of measurements, eliminating the risk of overfitting the data. The better filter sets in Figure 2 are those across the outer arc of the graph and without significant compromise in the call rate. With representatives of the better sets defined, power of the study can be examined for each of them to determine the best filtering set for the study of interest as is discussed below (Figure 5).

We examine in more detail three example filter sets corresponding to high, medium, and low stringency. The effect of the marker confidence filter and other filters on conversion can be seen in Figure 3a. We show in Figure 3b and c that there is a correlation between marker confidence and call rate and Mendelian concordance. The increased call rate of the low stringency filter set compared with the other two sets (Figure 3b) reflects the lower threshold accepted for the genotype level filter of call confidence. This causes a passed marker to have a higher call rate in the lower stringency filter set. Those marginal calls have compromised accuracy leading to decreased Mendelian concordance for the same markers (Figure 3c).

Effect of the performance metrics in association studies

As was described above, the three metrics conversion rate, call rate, and genotype accuracy can be traded off against each other. Here, we discuss the evaluation of the relative importance of these three metrics in different genetic scenarios.

Figure 2 Example of multifiltering. In this example, 1299 filtering sets were applied to the same data set of 12 234 markers clustered in 113 samples. Each of the filtering sets reflects the combination of multiple filters including call confidence, call rate, marker confidence, and maximum cluster variance. For each of these filter sets, we show the conversion rate on the X-axis and the Mendelian concordance rate on the Y-axis in this specific data set. The size of the spots denotes the average call rate with the larger size for the higher average call rate. We note that these filters are not fully independent as evident by the clustering of related points close to each other as well as on top of each other. The three black arrows point to the three filter sets that are examined further in Figure 3.

Conversion rate Given the presence of extensive LD in the genome,^{12,13} the effect of nonconversion of a marker relates to whether there is another genotyped marker with a high degree of LD with respect to the nonconverted marker or not. Among all the SNPs in the study, the one with the highest r^2 value with the nonconverted SNP estimates the loss of power. For example, if the maximum r^2 value with the nonconverted marker is 0.5 in the population of interest and this nonconverted marker is the disease-causing variant, then detecting this association requires double the sample size that would have been needed with data from the nonconverted marker.⁹

The computation of r^2 can be carried out from available data if the SNPs in the study are genotyped by the HapMap¹⁴ or in the more common case it can be estimated from the distribution of the maximum r^2 values of the converted markers in the study. This assumes that the distribution of the maximum r^2 of a nonconverted marker with any marker in the study is the same as that of the converted markers in the study. This assumption is only true if the nonconverted markers constitute only a small fraction of the total number of markers and if they have similar LD properties as other SNPs. Correlation between markers nonconversion and their physical proximity to each other, as is generated by repeat elements, constitute an example where our assumption is not accurate. One caveat is that the r^2 value between an SNP causing disease and another SNP in cases can be different than that in

controls, and hence the distribution of maximum r^2 for disease-causing SNPs can theoretically be different from that of other SNPs. However, the difference between r^2 values in cases and controls is generally modest when the allele frequency difference between cases and controls is not very large.

We expect that the drop in maximum r^2 (and hence the effect of nonconversion) is more precipitous as the marker density decreases. We have indeed observed that for markers at 10kb or lower density and common minor allele frequency (>10%) in the Caucasian population (data not shown). We have also observed that the effect

of nonconversion when the SNPs for a study are chosen by a tag selection algorithm based on r^2 between SNPs¹⁵ is more profound than with an even spacing approach (data not shown). This is likely to be a property of all tagging algorithms and is due to the fact that the objective of these algorithms is to minimize SNPs that highly correlate among themselves and therefore the effect of nonconverted SNPs becomes more severe.

Accuracy and call rate We have assessed the effect of allele frequency on the deterioration of power with error. For a given GRR, higher frequency alleles have better power than lower frequency alleles. We have 'normalized' this effect by using combinations of allele frequency and GRR that would have the same power (or same sample size necessary to get 80% power) in the absence of error. Examples of these combinations are alleles with 1 and 10% frequencies in the control population and GRR of 4 and 1.75, respectively.

To compute the effect of error, we used a simple model where there is no error bias between the two alleles. As can be seen in Figure 4, the effect of error shows marked differences at different allele frequencies consistent with previous reports.¹⁶ A 3% error requires, for the detection of association with alleles at frequencies of 1 and 10%, a sample size increase of 100 and 20%, respectively. It is of note that this relative increase in sample size with error is independent of the number of samples required to detect the association assuming perfect genotyping. Therefore, the same conclusion is still true when larger sample sizes are required to detect weaker genetic effects.

Figure 3 Comparison of performance metrics in three filter sets. (a) We show here the results per marker for three filter sets. On the X-axis, we have sorted markers in each filter set by their marker confidence. Out of 12 234 markers, there were 10 744, 10 499, and 9920 markers that were converted in the low, medium, and high stringency filter sets, respectively. The average call rate was 99.27, 98.39, and 99.2% for the low, medium, and high stringency filter sets, respectively. The effect of marker confidence on conversion can be seen by the abrupt stop of the high stringency curve, as markers below a threshold of 0.97 are not considered converted in this filter set. Other filters clearly affect conversion as the low and medium stringency sets have the same marker confidence of 0.9, but the number of passing markers is different. Furthermore, these other marker filters cause the high stringency set to have lower number of passing markers even if we only consider those with marker confidence threshold of 0.97. (b) The relationship between marker confidence and call rate. (c) The relationship between marker confidence and Mendelian concordance. Even though the Mendelian concordance for the worst markers in the low stringency set is lower than 93%, the vast majority of markers have marker confidence above 98% (as shown in a) and therefore Mendelian concordance higher than 99%. The increased call rate of the low stringency filter set compared with the other two sets reflects the lower threshold accepted for the genotype level filter of call confidence. This leads a passed marker to have a higher call rate in the low stringency filter set. Those marginal calls have compromised accuracy causing decreased Mendelian concordance in the low stringency set.

Figure 4 The effect of accuracy on power across allele frequency range. Assuming perfect genotyping with 100% call rate and accuracy, we calculated the combination of allele frequency and GRR (under a multiplicative model) that would give 80% power (P -value of 0.05 after Bonforoni correction for 10 000 markers) to detect association in a study of 850 cases and 850 controls. The appropriate GRR clearly changes with the allele frequency. For example for 1 and 10% frequency alleles, the needed GRR to have 80% power to detect the association are 4 and 1.75, respectively. In this figure the allele frequency and GRR combination are used as pairs and simply denoted by the allele frequency on the X-axis. The Y-axis shows the number of samples required to have 80% power to detect the association. By definition, the number of samples required with perfect genotyping to detect the association is 1700 (750 cases and 750 controls) across the allele frequency spectrum. This 'normalization' facilitates the comparison of the differential power reduction across the allele frequency range. The effect of accuracy on the power is clearly dependent on the allele frequency. Alleles at lower frequencies (eg, 1%) are a lot more sensitive to error than those at higher frequencies (eg, 17%).

On the other hand, there is no allele frequency dependence on the magnitude of the loss of power due to missing data (call rate <100%). If we assume that there is no allelic bias in the invalid calls, then the effect of call rate on single marker analysis is straight-forward as it mimics the reduction in sample size. The effects of nonrandom missing data are discussed in Clayton *et al.*¹⁷

Example of power comparison in data generated with two filter sets

It is possible to reduce the three performance metrics resulting from each filtering set in Figure 2 to one number reflecting the power of the study. However, this assumes that all the markers are equivalent, with the inaccuracy equally distributed across all of them.

Instead, a more useful approach is to examine power across the spectrum of SNP quality.

As shown in Figure 3, there is a correlation between marker confidence and Mendelian concordance allowing the assessment of power in different bins of marker confidence. Therefore, to determine the most appropriate filtering set in a study, we assess the power of representatives of the better filter sets across different bins of marker confidence.

As an example, power was compared for the three filtering sets assessed in Figure 3. These sets are candidates

Figure 5 Sample size required for the detection of association with different markers in the study. The Y-axis shows the minimum sample size necessary to achieve 80% power to detect association (at P -value of 0.05 after Bonforoni correction for testing 10 000 markers) in the proportion of all of the 12 234 markers in the study shown on the X-axis. The markers in the X-axis are sorted by their marker confidence in each of the two clustering filters. Perfect genotyping with 100% call rate and accuracy in all the markers requires 1700 samples to achieve 80% power. Markers on the right side are those with lower accuracy and call rate and require higher sample size. (a) The case shown is for a causative allele of 20% frequency and a GRR of 1.54. Only a slight increase in sample size for most of the markers (left side of the graph) is necessary for the high, medium, or low filtering cluster filtering sets in comparison to perfect genotyping. Markers on the right side of the graph require larger sample size. The lower stringency filtering sets provide useful information on more markers. In this case, the low stringency set is the optimal among the three filtering sets as it provides information on more markers (on the right side of the graph) while not compromising the power of markers on the left side of the graph. Similarly, the medium stringency filtering set is better than the high stringency set. (b) The case shown is for a causative allele of 1% frequency and a GRR of 4. In this case, the lower stringency filtering set is the worst of the three as markers on the left side of the graph are significantly compromised. The medium stringency set provides marginally more information than the high stringency set for some markers on the right side of the graph and is therefore the optimal choice in this case.

for the optimal filtering set in this study as they are representative of the best sets present on the outer arc in Figure 2. We calculated the sample size necessary to detect the association for each bin of markers assuming causative

allele frequency of 20% (Figure 5a). Markers on the left of the graph with the highest marker confidence and associated accuracy come very close to achieving the power of perfect genotyping. Markers on the right side of the graph require higher sample sizes to detect the association. All three filter sets perform almost equally well for the markers with high marker confidence, but the low stringency filter set allows additional power to be gleaned from markers that failed the high or medium stringency filter set. Therefore, when considering a common (20%) causative allele, the lower stringency filtering set is optimal.

When one performs the same analysis assuming a causative allele frequency of 1%, the situation is different (Figure 5b). As was seen in Figure 4, the effect of error is more severe when interrogating alleles with lower frequency. In this case, low stringency filter sets have significantly decreased power even for high-performing markers as a small number of erroneous calls are allowed to slip through without much gain in power for the marginal markers whose utility is eliminated by higher error rate. A better strategy in this case is to use a medium or high stringency filter set, which restores most of the power for the high-performance markers with a minimal price in terms of loss of marginal markers. In this case, for rare (1%) causative alleles, the optimal choice is the medium stringency set as it provides information on marginally more markers than the high stringency filtering set.

Discussion

Genotype determination algorithms for highly multiplexed genotyping technologies need to be robust and scalable as well as flexible for the specific studies of interest. To accomplish this, we described a methodology comprising of a consistent automated genotype calling algorithm and the assessment of a large numbers of filter sets to compute for each set the three performance metrics: assay conversion, call rate, and genotype accuracy. This mapping of the complex multidimensional filtering space to the simpler 3D space of performance metrics is specific to each custom set of SNPs. The three performance metrics relate to power providing a reasonable approach for selecting the most powerful genotyping filter set. A single power value can be obtained from the performance metrics if one assumes that all SNPs are behaving in the same way. To avoid this oversimplifying assumption, we opted to compare the power between filtering sets across the full SNP quality spectrum.

The methodology described here relies on the computation of Mendelian concordance as a surrogate for accuracy. Given the large number of SNPs, this can be robustly estimated using only a few trio samples with each new custom SNP set. One critical factor in the ability to use the multifiltering is that the optimization variable should be

distinct from the validation variable. For example, making cuts on markers that have a specific number of Mendelian errors compromises the ability to utilize this metric to estimate the accuracy. Furthermore, irrespective of this optimization ability, making cuts on variables like Mendelian errors can blind one to real biology of potentially great interest. Indeed by examining Mendelian errors we have identified several copy number polymorphism sites (data not shown).

We believe our approach is applicable to all highly multiplexed genotyping technologies. The applicability of this approach to optimal genotype calling for a particular study requires the consideration of three additional factors. First, we have studied the effect of performance metrics on statistical power only in single marker case-control analysis. The use of haplotypic or epistasis multimarker analysis is expected to amplify the effect of lower call rate or inaccuracy.¹⁸ Second, the focus on power does not capture other important values of data sets. For example, the generation of a minimal number of false positives due to technology artifacts is clearly a desirable feature and may dictate a different analysis approach.¹⁷ Finally, the optimization of conversion rate is influenced by an important characteristic of a genotyping technology which is the design success rate. The design success is the fraction of desired SNPs passing *in silico* criteria a specific technology may have. A trade-off between design success rate and the conversion rate of the designed SNPs is often possible. As the design rules become more stringent passing only markers that are informatically predicted to perform well in an assay, the conversion rate of these passed markers increases. With an infinite number of available and equally 'valuable' SNPs, the design success rate becomes irrelevant. In spite of the great effort in the public domain and the vast number of SNPs in the databases, SNPs are not infinite. More importantly the SNPs are not equally valuable. SNP choice for genetic studies is often made through LD-based tagging approaches,^{15,19,20} and/or they can be heavily biased towards functional variants, like nonsynonymous changes.²¹ Therefore, the trade-off between the design and conversion rates depends on the study of interest.

References

- 1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
- 2 Hardenbol P, Baner J, Jain M *et al*: Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 2003; **21**: 673–678.
- 3 Oliphant A, Barker DL, Stuelpnagel JR, Chee MS: BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* 2002; (**Suppl**): 56–58, 60–61.
- 4 Hardenbol P, Yu F, Belmont J *et al*: Highly multiplexed molecular inversion probe genotyping: over 10 000 targeted SNPs genotyped in a single tube assay. *Genome Res* 2005; **15**: 269–275.
- 5 Kennedy GC, Matsuzaki H, Dong S *et al*: Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003; **21**: 1233–1237.

- 6 Di X, Matsuzaki H, Webster TA *et al*: Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* 2005; **21**: 1958–1963.
- 7 Aitkin M, Rubin DB: Estimation and hypothesis testing in finite mixture models. *J R Statist Soc* 1985; **47**: 67–75.
- 8 Breslow NE, Day NE: *Statistical Methods in Cancer Research: Vol. 1 – The Analysis of Case–Control Studies*. Lyon, France: IARC Scientific Publications, 1980.
- 9 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
- 10 Gordon D, Finch SJ, Nothnagel M, Ott J: Power and sample size calculations for case–control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Heredity* 2002; **54**: 22–33.
- 11 Gordon D, Heath SC, Ott J: True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Heredity* 1999; **49**: 65–70.
- 12 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
- 13 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
- 14 Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
- 15 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
- 16 Kang SJ, Gordon D, Finch SJ: What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004; **26**: 132–141.
- 17 Clayton DG, Walker NM, Smyth DJ *et al*: Population structure, differential bias, and genomic control in a large scale, case–control association study. *Nat Genet* 2005; **37**: 1243–1246.
- 18 Kirk KM, Cardon LR: The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet* 2002; **10**: 616–622.
- 19 Weale ME, Depondt C, Macdonald SJ *et al*: Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 2003; **73**: 551–565.
- 20 Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF: Minimal haplotype tagging. *Proc Natl Acad Sci USA* 2003; **100**: 9900–9905.
- 21 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003; **33** (Suppl): 228–237.

Supplementary Information accompanies the paper on European Journal of Human Genetic website (<http://www.nature.com/ejhg>)