AFFYMETRIX®

Technical Note

GeneChip® CustomSeq® Resequencing Array Base Calling Algorithm Version 2.0: Performance in Homozygous and Heterozygous SNP Detection

Large-scale comparative sequencing projects require a rapid, accurate, and cost-effective method for variant detection and genotyping. The new generation of Affymetrix GeneChip CustomSeq® Resequencing Arrays combine higher density arrays, an improved sample prep protocol, and an improved base calling algorithm to enable analysis of up to 300 KB of unique DNA sequence on a single array. Customers have employed these arrays in a variety of contexts, from whole-genome sequence variation detection in haploid bacterial genomic DNA, to resequencing entire gene families derived from diploid human genomic DNA.

Introduction

The GeneChip® Sequence Analysis Software (GSEQ) incorporates an improved algorithm, Resequencing Algorithm Version 2.0 (RA v2.0), to provide automatic sequence and genotype calls from hybridization intensity data obtained from GeneChip CustomSeq® Resequencing Arrays. This technical note describes RA v2.0 and its performance on two data sets representing homozygote and heterozygote SNP detection for 8 µm feature size arrays. These data demonstrate the ability of resequencing arrays to generate high-quality sequence information at >90 percent call rate and >99.9 percent accuracy. In addition, it describes analysis methods that can be applied as post RA v2.0 filters that improve performance for heterozygous detection. After filtering, an overall accuracy of 99.98 percent was achieved on a specific 300 KB design representing diploid sequence.

Algorithm Overview

Development of an automated and highly accurate method of heterozygote SNP detection is a challenging task. Previous algorithms developed for resequencing arrays, such as the resequencing Algorithm Version 1.0 (RA v1.0) and the ABACUS algorithm developed by Cutler, *et al.*[1], have demonstrated highly accurate and reproducible base calling from resequencing arrays with 20 µm features. RA v2.0 is an alternative base calling method, built upon the principles of RA v1.0 and ABACUS, which implements critical enhancements and improves robustness for making base calls from 8 µm feature arrays. The RA v2.0 consists of three major steps. The first step uses a set of preprocessing filters to detect poorly behaving probe sets in a specific sample; the second step consists of the base calling method; the third step evaluates the local hybridization pattern around each base across all samples in an analysis batch, in order to identify unreliable calls that are potential false positives. A detailed description of the algorithm was published in the 2006 Conference Proceeding from EMBS[2].

Algorithm Parameters

RA v2.0 uses several parameters in making a base call. Please refer to the Affymetrix GeneChip® Sequence Analysis Software User's Guide Version 4.0 (GSEQ 4.0) for a complete description of algorithm parameters. Two of the user-definable parameters, the Genome Model and the Quality Score Threshold (QST), must be set to the appropriate value prior to analysis.

GENOME MODEL VALUE

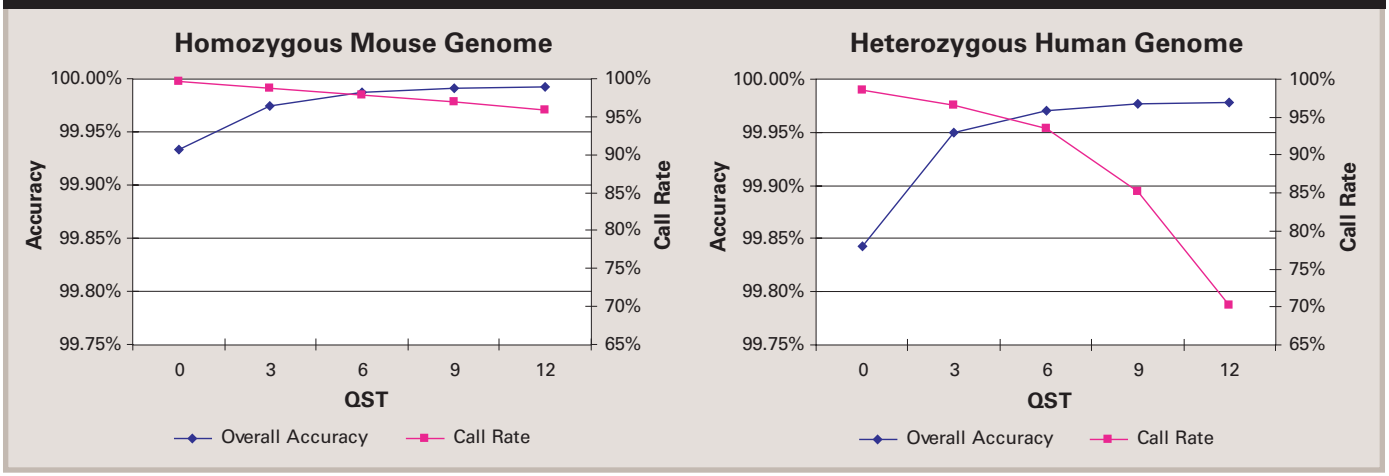
The Genome Model Value may be set to either 0 for diploid/heterozygous systems or 1 for haploid/homozygous systems. The Genome Model Value parameter determines which of the 11 possible genotype models the algorithm considers in making a call. At a Genome Model Value of 0, the algorithm considers all 11 possible models, 4 homozygous, 6 heterozygous, and a no call, in determining any particular base

[1] Cutler, D.J., *et al.*, High-throughput variation detection and genotyping using microarrays. *Genome Research*, **11**:1913-1925, (2001).

[2] 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 30-September 3, 2006. New York City, New York.

[3] https://www.affymetrix.com/support/downloads/manuals/gseq_user_guide.pdf

Figure 1: Impact of Quality Score Threshold on Call Rate and Concordance.



call. At a Genome Model Value of 1, the algorithm assumes that all positions are homozygous and does not consider the 6 heterozygous models. It is essential to set the appropriate Genome Model Value prior to analysis of data.

QUALITY SCORE THRESHOLD

RA v2.0 calculates a Quality Score (QS) for each position on the array based on the confidence of that particular call. The QS is a measure of the difference between the likelihood of the best fitting model and the second best fitting model. The lower the quality score, the lower the confidence of that particular call. The Quality Score Threshold (QST) is a tunable parameter that allows the user to set a minimum threshold that each position must meet in order for the algorithm to make a base call. All positions with QS below a given QST will automatically result in no call.

The impact of QST on data from two unique 300 KB GeneChip CustomSeq Array designs representing homozygous and heterozygous systems was assessed. As expected, when the QST value was increased, the overall accuracy increased and the number of bases called decreased (Figure 1). For the haploid/homozygous analysis, there was only a slight decrease in call rates (<2 percent) as the quality score threshold was increased above 6. Based on these data, a default QST of 12 was selected for maximum accuracy when analyzing homozygous data. By contrast, the call rate and accuracy of the diploid/heterozygous analysis was more sensitive to QST, resulting in a significant reduction in call rate at QST greater than 3, with a relatively small gain in overall accuracy. Based on this observation, QST=3 was selected as the default value for diploid analysis. The increased sensitivity of call rate and accuracy to QST in diploid analysis is largely a consequence of the fact that haploid data are fit to five possible homozygous genotype models, whereas diploid data are fit to 11 possible (homozygous or heterozygous) genotype models. The QS is calculated as the ratio of the likelihood of the best fitting model to the next best fitting model, and because diploid analysis considers the six heterozygous models in addition to the homozygous models, diploid analysis, in general, results in lower QSs.

The values described above are meant to provide guidance for baseline QST values. The trade-off between call rate and accuracy should be considered for each individual experiment. If a low false positive rate (high accuracy) is of greater importance than low false negative rate (high call rate), one may adjust the QST higher. Alternatively, if a low false negative rate is of greater importance than a low false positive rate, a lower quality threshold may be selected.

Homozygous Performance

DESCRIPTION OF HOMOZYGOUS DATA SET

To assess performance of resequencing arrays on haploid targets, or in circumstances where only homozygous genotypes are expected, we applied the algorithm to sequence data generated from a pure-bred mouse strain. A 300 KB CustomSeq Array was designed based on publicly available sequence for the C57BL6 (B6) mouse strain. These arrays were hybridized to PCR targets generated from genomic DNA from the DBA strain, prepared according to the standard protocols described in the *GeneChip® CustomSeq® Resequencing Array Manual*[3]. Each of three biological replicates was hybridized to three arrays randomly selected from a single manufacturing lot. The experiment was replicated on a second set of arrays selected from a different manufacturing lot, giving a total of 18 arrays in the performance assessment set. The resulting intensity data were analyzed as a single batch using GSEQ 4.0, with all parameters set to default values, except with the Genome Model set to 1, and the QST set to 12.

HOMOZYGOUS PERFORMANCE

To determine the accuracy of array-based variant detection in haploid samples, the DBA mouse array data were compared to data obtained by capillary sequencing. Over 30 KB of genomic sequence from the

Table 1: Homozygous Accuracy – Mouse Genome.

Average Call Rate	The number of bases called divided by the total number of bases possible.	95.92%
Overall Accuracy	For all bases where a call is made, the percentage that agrees with capillary sequencing.	99.99%
Overall Reproducibility	For a pair of technical replicate arrays (pairs of mouse samples in this case), concordance is computed for all sites where the two arrays make a call. Ns excluded.	99.99%
Homozygous SNP Call Rate	Percentage of calls made for all known SNP positions.	95.95%
Homozygous SNP Accuracy	Percentage of calls made as SNP when capillary sequencing called the base as a SNP. • Does not include Ns or SNPs within 9 bps of another SNP	100%
Homozygous SNP False Positive	Percentage of calls made as a SNP when capillary sequencing called the base a reference.	0.01%
Homozygous SNP False Negative	Percentage of no-calls made when capillary sequencing called the base as SNP. • Does not include SNPs that occur within 9 bps of each other • Calculated for individual genotypes, not SNP sites	4.05%
Homozygous SNP Reproducibility	Same as overall reproducibility, but for SNP sites only.	100%

DBA mouse was evaluated by capillary sequencing using a commercially available service. Among the 34,766 sites with high-quality dideoxy sequencing calls (Phred score greater than 60), there were 173 non-reference homozygous calls (SNPs). In assessing performance, all SNP sites that were within nine base pairs of each other were excluded; thus the SNP detection accuracy values reported here are representative of accuracy at isolated SNP loci. Table 1 summarizes the performance results.

Heterozygous Performance

DESCRIPTION OF HETEROZYGOUS DATA SET

To assess the performance of CustomSeq arrays on diploid targets, a 300 KB array was designed to query the non-repetitive sequence from a contiguous interval within an ENCODE region (www.genome.gov/10005107) on chromosome 4. Twenty-five LR-PCR assays were used to amplify approximately 250 KB of this interval from 16 CEPH DNA samples. The resulting 16 pooled PCR products were fragmented, labeled, and hybridized to arrays, and the arrays were stained and scanned, according to the standard protocol described in the *CustomSeq Resequencing Array Manual*. The intensity data from the resulting 16 scans were analyzed as a single batch using GSEQ 4.0, with all parameters set to default values, except with the Genome Model set to 0 and the QST set to 3.

CALL RATE ANALYSIS

The array data were compared to capillary sequence and genotype data obtained from the ENCODE project for the same genomic interval in the same 16 CEPH samples. Across the 16 samples, the array call rates ranged from 96 percent to 99 percent (average of 96.6 percent), whereas the capillary sequence coverage ranged from 67 percent to 81 percent, with an average of 74 percent (Stacey Gabriel, personal communication).

Whereas the CustomSeq arrays contain oligonucleotide probes with a substantial range of GC contents, the array hybridization must be performed at a single temperature. To assess the impact of probe GC content on performance, average call rates over various probe GC content bins were calculated. Figure 2 indicates that average call rates exceed 90 percent for probes with GC content up to 70 percent, and that average call rates drop substantially when probe GC content exceeds 70 percent. The overall GC content of the interval evaluated in this experiment was 34 percent. Regions with higher overall GC content will have a greater fraction of probes at the higher GC content range, which could result in a slight decrease in overall call rates compared to these data.

SAMPLE SIZE

Because the algorithm computes a predicted

Figure 2. Call Rate as a Function of Probe GC Content.

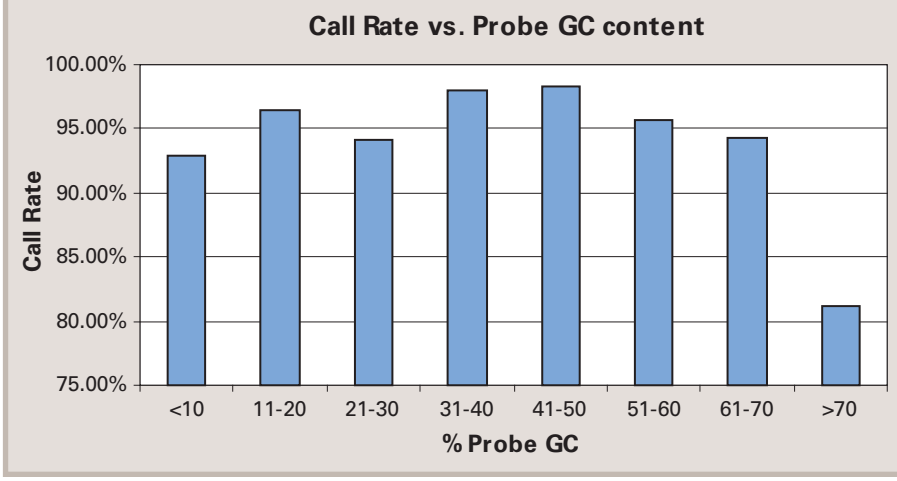
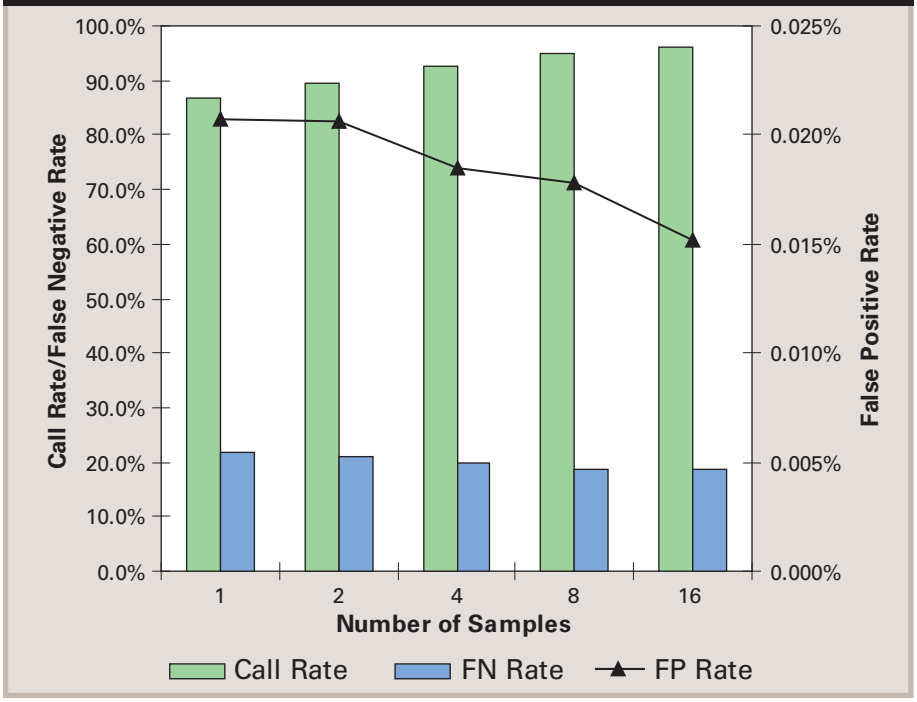


Figure 3: Performance as a Function of Sample Size.



background across all samples at each site analyzed in a batch (see Adaptive Background Section in GSEQ manual pp. 217-233), the performance was assessed as a function of sample size. A random set of samples was selected for each batch size. The call rates increased as sample size increased with improvement in both false negative and false positive rates (Figure 3).

DIPLOID ACCURACY

As part of the public ENCODE resequencing project, this region was analyzed by capillary sequencing for SNP discovery in 48 individuals, including the 16 CEPH individuals used to generate the array data described in this Tech Note. The newly discovered SNPs, as well as all known SNPs, were subsequently genotyped using multiple methods in all 48 individuals (Stacey Gabriel, personal communication). To measure performance, calls were compared for bases covered by both the ENCODE sequencing/genotyping data set and the array-based resequencing data set. Genotype data for known SNP loci in all 16 CEPH samples were used to assess array performance at homozygous and heterozygous positions (Table 2).

It is important to note that the ENCODE capillary sequence data set likely contains some errors. As a consequence, this analysis may be underestimating the accuracy of the array data. For example, some putative array false positives may, in fact, be ENCODE false negatives.

POST GSEQ RAv2.0 ANALYSIS

In an effort to understand the conditions that give rise to false negative and false positive calls, each discordant call was examined in detail. This analysis resulted in the identification of a set of five factors that accounted for the majority of false positive calls in this data set: PCR failures, nearby SNP effects, cross hybridizing probes, low sequence complexity probes, and non-biallelic calls. Next, a set of filters based upon these factors was developed that could be used to systematically remove false positive calls, thereby increasing accuracy. A detailed description of the filters is presented below.

– *PCR Failure*

Inefficient and failed PCR contributed significantly to No Calls and false positive calls, highlighting the importance of amplicon verification prior to array hybridization.

Table 2: Diploid Accuracy – Human Genome.

Call Rate	The number of non-N calls divided by the total number of calls.	96.56%
Overall Accuracy	Percentage of all calls (excluding Ns) that are concordant with ENCODE data.	99.95%
Call Rate at Variant Sites	Percentage of calls made for all known SNP loci including heterozygous and homozygous calls.	89.70%
SNP False Negative Rate	Percentage of variant positions in the ENCODE data that are called N or reference in the array data.	17.34%
SNP False Positive Rate	Percentage of reference positions in the ENCODE data that are called variant in the array data.	0.04%
Homozygous Accuracy	Percentage of homozygous variant positions in the ENCODE data with concordant array data (excluding array Ns).	96.91%
Heterozygous Accuracy	Percentage of heterozygous positions in the ENCODE data with concordant array data (excluding array Ns).	86.25%
Homozygous SNP False Negative	Percentage of mis-calls (No Calls and Ref calls) made for all known homozygous SNP positions in the ENCODE data.	9.12%
Heterozygous SNP False Negative	Percentage of mis-calls (No Calls and Ref calls) made for all known heterozygous SNP positions in the ENCODE data.	22.15%

All calls that originate from a failed amplification should be removed. These can be easily identified by examining call rate as a function of the PCR amplicon in a specific sample. In this data set, all calls resulting from amplicons with call rates below 90 percent were converted to No Calls, resulting in removal of seven amplicons out of a total of 400 PCR amplifications.

– *Near SNP/Footprint Effect*

The presence of a variant in the target results in an increase in hybridization signal from the complementary non-reference probe, but also results in a decrease in signal from the reference probes at positions surrounding the polymorphism, leaving a 10-20 bp footprint of low signal. This effect can cause the algorithm to call multiple polymorphisms close to one another, where one call is the true polymorphism and the others are false positives. To reduce false positive calls that surround a true variant call, all sites where multiple variant calls occurred within nine bases of each other were assessed. The variant call within a cluster with the highest quality score was retained, while the lower quality score variant calls were converted to Ns.

– *Cross Hybridizing Sequences*

If a target contains 25 out of 25 bases of sequence complementary to a tiled (reference or non-reference) probe, it typically results in a hybridization signal and a base call. Similarly, even imperfect (24 out of 25 bases) matches occurring at non-cognate positions on the array (due to, for example, repeated sequences in the target, or local duplications) can result in cross hybridization, which can lead to false positive base calls. To eliminate variant calls arising from cross hybridization in the target, the 25 bases surrounding each identified variant were compared by BLAST to the sequence of the hybridization target. If ≥ 24 bases matched a non-cognate locus in the target sequence, it was eliminated. This analysis eliminated 6 percent of all false positives, and none of the true positives.

Table 3: Reduction of False Positives by Post-GSEQ filters.

Summary of exclusions – position* sample specific cell counts	Calls removed	# FPs removed	% FPs removed	# TPs removed	% TPs removed
PCR Failure	19,519	168	31.28%	5	0.33%
Nearby SNPs – Footprint	252	167	31.10%	14	0.33%
Cross Hybridization Sites	64	33	6.15%	0	0.00%
Low Complexity Probes	128	1	0.19%	0	0.00%
Non-biallelic Calls	32	32	5.96%	0	0.00%

– *Low Complexity Sequences*

Probes with low sequence complexity can result in poor hybridization quality. Polymorphic calls at these sites will have a higher chance of being incorrect. LZW compression algorithm (Ziv and Lempel 1977) was applied with a threshold compressed probe sequence size of 13 to remove all calls at these sites to further reduce false positives. It is important to note that a different threshold or an alternative method for identifying low complexity sequences may be more effective in reducing false positives in other data sets.

– *Non-biallelic Calls*

Because the majority of SNP loci are bialleic and the presence of a true third allele is rarely observed, triallelic or tetra allelic loci are, in most experimental contexts, likely the result of spurious calls. Calls at such loci should be converted to N.

Table 3 describes the results of applying these five filters to calls from analysis at QST=3. All removed calls were converted to Ns, which resulted in a slight overall decrease in the call rate. The majority of false positives were removed by the failed PCR and nearby SNP effect filters, with a concomitant loss of a small fraction of true positives. Because some of the false positives were identified by multiple filters, the cumulative number of false positives eliminated was less than the sum of the false positives identified by the individual filters. Because of the limited sample size in this data set, we did not exclude sites not in Hardy Weinberg equilibrium, although clearly this filter could be useful, depending on the nature of the study.

A summary of overall performance before and after applying the post-GSEQ filters is described in Table 4. The 537 discordant calls that were categorized as false positives are reduced by 318 (219 remaining). The overall accuracy increases from 99.95 percent to 99.98 percent with a reduction in call rate of less than 1 percent.

Table 4: Accuracy

	Before Filters	After Filters
Call Rate	96.56%	95.98%
Overall Accuracy	99.95%	99.98%
False Positive Calls	537	219
False Positive Rate	0.040%	0.016%
True Positive Calls	1,498	1,479
SNP Call False Positive Rate	17.34%	18.52%
SNP Call False Negative Rate	8.18%	9.39%

Conclusions

The Resequencing Algorithm Version 2.0 (RA v2.0) is capable of generating high-quality sequence information from GeneChip® CustomSeq® Resequencing Arrays. Data from two distinct array designs demonstrated the algorithm's ability to automatically call >90 percent of bases from a single array hybridization. Analysis of the call rate as a function of probe GC content indicated that, for GC probe compositions ranging up to 70 percent, the call rates remained above 90 percent. For the homozygous data in this study, the algorithm demonstrated 99.99 percent sequencing accuracy and >99.99 percent reproducibility with low false positive (.01 percent) and negative rates (4.0 percent). For the more challenging diploid analysis, the automated calls resulted in >96 percent of the bases called with 99.95 percent accuracy. Additionally, diploid performance can be further improved to result in 99.98 percent accuracy by applying relatively simple post-GSEQ filters.

NOTE:

AFFYMETRIX, INC.

3420 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

For research use only. Not for use in diagnostic procedures.

Part No. 702331 Rev. 2