Published in final edited form as:

Curr Protoc Hum Genet. 2009 April ; 0 7: Unit-7.17. doi:10.1002/0471142905.hg0717s61.

An Overview of Custom Array Sequencing

Prachi Kothiyal^{1,2}, Stephanie Cox³, Jonathan Ebert¹, Bruce J. Aronow^{1,2}, John H. Greinwald^{1,2}, and Heidi L. Rehm^{3,4}

¹Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

²University of Cincinnati, Cincinnati, Ohio

³Partners Healthcare Center for Personalized Genetic Medicine, Boston, Massachusetts

⁴Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts

Abstract

This unit provides an overview of oligo hybridization-based resequencing and a wide range of considerations for implementing the technology and analyzing the resulting data. The specific technology discussed is the Affymetrix GeneChip CustomSeq Resequencing Array platform. Concepts related to array design, experimental protocols, and base-calling using existing algorithms are presented. Details that should be evaluated during development of sequence tiling, target amplification, and PCR protocols are addressed. An overview of the Affymetrix GeneChip Sequence Analysis Software (GSEQ) is provided, along with factors that influence base-calling coverage and accuracy. Also outlined are performance measures that can be used to characterize base-calling with resequencing arrays, as well as factors known to affect their performance. Limitations associated with detection of insertions and deletions (indels) are discussed, with empirical data from our experiments used to outline possible approaches to indel detection. Critical topics in the design, implementation, and analysis of targeted sequencing arrays not previously discussed in detail are highlighted.

Keywords

resequencing microarrays; CustomSeq; base-calling; sequence-specific hybridization; mutation detection

INTERNET RESOURCES

http://www.affymetrix.com/products/arrays/specific/custom_seq.affx

CustomSeq Array Resources provided by Affymetrix.

http://www.affymetrix.com/support/technical/datasheets/customseq_datasheet.pdf

GeneChip CustomSeq Resequencing Array Program Data Sheet provided by Affymetrix.

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

GeneChip Sequence Analysis Software (GSEQ) User Guide provided by Affymetrix.

http://www.affymetrix.com/support/technical/technotes/customseq_arraybase_technote.pdf

Technical note on CustomSeq Resequencing Array Base Calling Algorithm Version 2.0 provided by Affymetrix.

http://www.affymetrix.com/support/technical/other/customseq_design_manual.pdf

CustomSeq Resequencing Array Design Guide provided by Affymetrix.

INTRODUCTION

Identifying and cataloging individual and population-level DNA sequence variations is a critical step towards understanding the genetic basis of disease and clinically significant human variation. As more and more genes, pathways, polymorphisms, and allelic interactions are understood with respect to their roles in disease, the need for cost-effective high-throughput DNA resequencing for clinically affected individuals has risen. Based on the recent rapid evolution of methods for constructing very large-scale, sequence-specific microarrays, powerful techniques for allelic gene resequencing have emerged, based on differential hybridization to interrogatory probe arrays.

Looking forward, while several next-generation sequencing technologies are emerging, at present, each has significant limitations that affect their ability to be practically applied to specific genes, gene sets, or individual diseases. In contrast to massive resequencing projects, such as whole genome de novo or scaffold-based resequencing, oligo hybridization-based resequencing is best used in settings in which a moderate amount of sequence is being analyzed (10 kb to 300 kb) in a repetitive manner (e.g., disease-specific studies analyzing hundreds or thousands of samples) against an established reference scaffold with relatively well understood variant structure. Moreover, we view it as a critical exercise to define the series of issues that must be considered prior to making clinical decisions based on sequence analyses from an individual patient's genome.

This unit focuses on the technology of oligo hybridization-based resequencing, general approaches for designing and running arrays, and principles behind the analysis of data that result from these approaches. The specific technology that will be presented is the Affymetrix GeneChip CustomSeq Resequencing Array platform. Affymetrix provides information about this technology on their Web site (see Internet Resources), including design guides, protocols, and other support materials. As such, this unit will not cover all details since many are readily available, but it will instead give a general overview of all aspects of the technology and its implementation, as well as focus on critical aspects that arise in our experience in the design, execution, and analysis of these arrays, but are not addressed in available materials.

CustomSeq ARRAY OVERVIEW

In the current configuration of Affymetrix GeneChip CustomSeq Resequencing Arrays, hereafter referred to as CustomSeq arrays, a single array can be used to sequence up to 300 kb of double-stranded unique DNA sequence. CustomSeq arrays rely on allele-specific hybridization for determining the DNA sequence of interest.

Figure 7.17.1 presents an overview of how probes are designed to query a sequence of interest. Every base to be sequenced is represented by probes that are 25 base pairs long and differ only at a single central position (Lipshutz et al., 1999). The array uses four oligomers per base for each of the two DNA strands. The four oligomers differ at a single central position that could be A, C, G, or T, and they query complementary bases on the DNA strand. The remaining 24 positions are the same for all four oligomers and are complementary to the reference DNA sequence being queried.

Many copies of each oligomer are synthesized by photolithography in discrete features on the array. Current CustomSeq arrays use a feature size of 8 μm , although development is ongoing for 5- μm based arrays. Feature size determines the number of individual 25-mer sequences that can be interrogated on the array, and a smaller feature size implies an increase in the array capacity.

Relative binding of the patient's DNA to each of the probes determines the nucleotide at a particular position. The sample is labeled with fluorescent molecules and the feature(s) with the maximum fluorescence on the CustomSeq array indicates the base(s) at that position. For diploid organisms, a single strong signal indicates homozygosity and two equally strong signals indicate heterozygosity.

Assay Overview

Figure 7.17.2 provides an overview of the resequencing steps. The Affymetrix GeneChip system consists of a probe array, hybridization oven, fluidics station, scanner, and a computer workstation. The assay leverages long-range or traditional polymerase chain reaction (PCR). PCR is used to select areas of interest from the genomic DNA, and these PCR products are then pooled and fragmented. Next, the DNA fragments are labeled with biotin and subsequently hybridized with the microarray. Finally, the chip is washed, stained, and scanned to measure the fluorescence intensities. The workstation with GeneChip Operating Software (GCOS) controls the fluidics station and the scanner. GCOS is also used for image acquisition and database management. The GeneChip Sequence Analysis Software (GSEQ) allows the user to perform sequence analysis of the data to produce the final sequence calls.

Array Design

The steps for designing an array are detailed in Affymetrix's Custom-Seq Array Design Guide. Briefly, the design process requires identification of the sequence of interest as the first step. Array capacity limits the total number of bases that can be accommodated on a single array and should be taken into consideration while selecting the sequences and flanking bases. Upon selection of genomic regions to be interrogated with the array, the sequences are converted into an acceptable format (FASTA) and checked for quality. PCR primers are then designed for amplification of regions of interest. Several more detailed aspects of the design process warrant additional discussion.

1. Familiarity with the target sequence is critical. In order to detect variation that occurs on multiple haplotypes, e.g., a mutation that occurs within 13 bases of a nonreference sequence variation, knowledge of all common gene variations is required. In this case, a series of resequencing probes can be designed to include the variant (substitution or indel) and thus be used to capture additional sequence variation over both scaffolds.
2. Redundant tiling is very useful. In our experience, most no-calls occur sporadically on the arrays, as opposed to repeatedly poor performing sequence specific probes. As a result, if target sequences are tiled in duplicate or triplicate, this problem is significantly reduced. We have assessed single, duplicate, and triplicate tiling. Each

additional tiling adds more sensitivity and specificity. Obviously, however, overall sequencing capacity per array is being traded. We are currently using triplicate tiling for each 25-mer sequence for all arrays used in clinical diagnostics.

3. Offset tiling, in which the interrogation position is shifted from the middle to the 5' or 3' end of the probe, is also useful in increasing sensitivity and specificity. This approach is taken in the design of Affymetrix's SNP arrays, in which three to seven offsets are used per SNP. Empiric data is most reliable in determining which probe will function the best. Since no such data is available during the initial design of most projects, we recommend that the additional tilings (if tiling in triplicate) use probes with the interrogation position at 9, 13 (center), and 17, the most distinct offsets.
4. To add increased sensitivity and specificity for previously identified variants in the target regions of interest, specific genotyping probes can also be tiled. This approach involves the use of as many as seven different probe sets per variant, placing the interrogation position at different offsets within the 25-mer, followed by analysis using genotyping software, e.g., BRLMM (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf) and Birdseed (Affymetrix; <http://www.broad.mit.edu/mpg/birdsuite/birdseed.html>). This approach also allows tiling of reported insertions and deletions, which are not easily detected using this technology.

Target Amplification

In order to generate enough target-enriched patient material for hybridization to the chips, selective amplification is needed for every genomic region being interrogated. Current approaches rely on locus-specific PCR for amplification. Long-range PCR may minimize the number of reactions required. However, large-scale sequencing of multiple different nuclear genes with dispersed exons still requires many long-range PCR reactions; long-range PCR is more expensive, less robust, and more variable in the quantities produced, therefore requiring quantification and normalization prior to hybridization.

In addition, in our experience, longer PCR products are more susceptible to underfragmentation, which then leads to poorer hybridization. As such, unless the DNA to be sequenced is contiguous (e.g., the mitochondrial genome), we recommend developing short-range PCR reactions that can be set up and cleaned up robotically in 384-well plates. Such reactions are highly reproducible, and, as such, set quantities can be robotically collected from each well, depending on the typical band intensity, without need for repeated quantitation and normalization. Development of short-range PCR assays is also useful for the development of sequencing assays for confirming variants identified on the array with capillary sequencing. The additional use of multiplex PCR design can decrease the total number of PCRs, although in our experience, going above triplex causes the quantity of each amplicon to be too low within a 15- μ l amplification reaction.

Genomic selection or targeted gene capture approaches are being developed for NextGen sequencing technologies. These methods are likely to also be useful for array-based sequencing to avoid the numerous upfront PCR reactions. These include the use of oligo

libraries, in solution, synthesized, or spotted on arrays, as well as home-made filter membranes. Each of these approaches is designed to select the DNA regions of interest through sequence-specific hybridization. While these approaches are still being optimized, the main challenge for use in CustomSeq arrays is the need to design capture probes that will be unique to the selective capture of the target regions being assayed, to avoid cross-hybridization from homologous sequences throughout the genome.

Array Protocol

After target amplification, the steps of the protocol (see Fig. 7.17.2) are mostly straightforward and generally work well when following the manufacturer's protocol. However, a few areas are worth discussing.

1. Pooling and normalization of PCR products: As mentioned above, when using short-range PCRs, we find that normalization is not required on each run, but it should be estimated based upon gel intensity of the product after examining several optimized PCR results. For most products, we use 4 μ l of a 15- μ l PCR reaction. For a repeatedly weaker products, we use 5 to 6 μ l. Pooling of products can be labor intensive, and we recommend the use of robotics, e.g., a Biomek FX robot with a span-8 head.
2. PCR clean-up and concentration: We use QIAquick PCR purification columns. If more than one column is needed because of upon pooled PCR volume, then eluates are combined. The eluate is then concentrated to the appropriate volume through the use of a SpeedVac, and the concentration of DNA is assessed by absorbance spectrometry.
3. Fragmentation: This is one of the most finicky steps to the protocol and must be monitored carefully. Overfragmentation and underfragmentation can cause problems. For example, we have observed that underfragmentation leads to weak hybridization, which in turn leads to reduced call rates. A fragmentation enzyme is part of the Affymetrix GeneChip Resequencing Reagent Kit. The identity of the fragmentation enzyme is proprietary, but we believe it is DNase I. We use 0.12 U of fragmentation enzyme in a reaction containing 2.5 to 5.5 μ g of pooled PCR products. After fragmentation is complete (smear of fragments from 20 to 200 bp), success is analyzed using either a high-resolution agarose gel (e.g., 4% HR agarose e-gel; Invitrogen) or an Agilent Bioanalyzer 2100. If the sample is underfragmented, more enzyme can be added.
4. Labeling, hybridization, washing, staining, scanning: These steps are well represented by the manufacturer's recommendations.

FILE FORMATS AND DATA OUTPUT

Figure 7.17.3 provides an overview of some of the critical files used or generated by the two software applications, GCOS and GSEQ, each of which are used during the course of a resequencing array experiment. The DAT file contains an image of the scanned probe array. The raw image from a DAT file is further processed to generate a CEL file, which contains averaged signal intensity values for each feature in numeric and visual formats. GSEQ

analyzes the signal intensity data from the CEL file to generate a CHP file with final sequence calls corresponding to each base position being probed. GSEQ automates the process of base calling using an algorithm with adjustable settings. It allows viewing of the sequence, exporting to FASTA and generation of a SNP summary report.

Grid Alignment

In order to generate average signal intensities of each probe (i.e., feature on the scanned array image), the software needs to superimpose a grid on the scanned image in such a way that each square in the grid encloses a single probe cell. Grid alignment is performed automatically by the software after the chip has been scanned. The goal of the grid alignment is for each square of the grid to delineate a single probe cell. Manual grid alignment is required when the algorithm fails to perform automatic grid alignment, or sometimes when poor data is obtained.

Figure 7.17.4A shows a small section of an array image after the software has performed successful grid alignment. Figure 7.17.4B shows a small section of an array with a misaligned grid. In this example, the grid has missed the first column of features, and in the bottom right, the grid is suboptimally aligned to capture the correct area of each probe cell. Therefore, manual grid alignment is needed to correct the problems. Manual grid alignment is best performed using a chip with successful grid alignment as a guide for aligning each of the four corners of the array. On occasion, misalignment can be limited to only small internal regions on the chip (poor call rates may signify such a problem), requiring careful analysis and alignment of subgrids.

GeneChip Sequence Analysis Software

GSEQ uses Resequencing Algorithm Version 2.0 (RA v2.0) to automate the generation of sequence and genotype calls from the intensity data. RA v2.0 has been built upon RA v1.0 and the Adaptive Background Genotype Calling Scheme (ABACUS) developed by Cutler and colleagues (Cutler et al., 2001). A detailed description of the algorithm is provided in the Affymetrix GSEQ User Guide. Figure 7.17.5 provides an overview of the steps involved in the generation of base calls in GSEQ.

Variant Detection and Performance Measures

At any given position, GSEQ can assign a homozygous call, heterozygous call, or a no-call to the base. For haploid organisms, the algorithm considers four homozygous models (A, C, G, T) and a no-call, whereas for diploid organisms, six heterozygous models (AC, AG, AT, CG, CT, TG) are also considered in addition to the above five models. Figure 7.17.6 shows examples of hybridization patterns corresponding to heterozygous and homozygous calls. When compared to the target reference sequence, each call could be a wild type (WT), a heterozygous variant (for diploid systems), a homozygous variant, or a no-call. A true variant could be falsely identified as a WT, a WT could be falsely identified as a variant, the base could be identified correctly as a WT or a variant, or it could be assigned a no-call.

Using true data (as determined by dideoxy sequencing) to determine call correctness, certain performance measures can be attributed to the chip, some of which are outlined in Table

7.17.1. These performance measures can be used as indicators to optimize the algorithm settings for base calling, some of which are discussed in the next section. The aim should be to minimize the false positive and false negative rates without jeopardizing the average call rate.

It is interesting to note that false positive (FP) and false negative (FN) miscall rates differ between variants and total bases. The value in the numerator is the same for both rates but the denominators vary. For total FP and FN rates, the denominator is total bases called by the array, while for variant FP and FN rates, the denominator is just the total number of variant calls/true variants. For example, if 25,000 bases are being called by an array and there are 30 true variants identified by an independent method, 20 false positives (WT called variant by array), 1 false negative (true variant called WT by array), and a total of 40 variant calls made by array, the rates would be:

Total FP rate of 0.08% (20/25000, the fraction of total calls that are incorrectly called variant by array)

Variant FP rate of 50% (20/40, the fraction of variant calls made by array that are incorrect)

Total FN rate of 0.004% (1/25000, the fraction of total calls that are incorrectly called WT)

Variant FN rate of 3.3% (1/30, the fraction of true variants that are incorrectly called WT by array).

GSEQ ALGORITHM PARAMETERS

GSEQ uses several parameters in making a base call. The Affymetrix GeneChip Sequence Analysis Software User Guide provides a complete description of the parameters. The genome model value and quality score threshold must be set correctly for the analysis.

The Genome Model Value is 0 for diploid organisms and 1 for haploid systems. For diploid organisms, the algorithm considers all the possible models per base position (four homozygous, six heterozygous, and a no-call) while determining the call. For haploid organisms, the algorithm evaluates only the four homozygous models to determine which one identifies the base call most reliably. In order to get valid results, it is important to set this value appropriately.

The quality score (QS) measures the difference between the likelihood of the best fitting model and the second best fitting model. Higher QS indicates a higher level of confidence for the call. GSEQ allows the user to set the QS threshold, where any position with QS below the threshold will be automatically assigned a no-call. Increasing the QS threshold (QST) results in more accurate results but the number of called bases decreases. Affymetrix provides guidelines for setting this threshold in the RA v2.0 technical note.

When a low false positive rate is preferred over a low false negative rate, the QST can be assigned a higher value. If true variants are known for the samples, these values can be assessed to obtain the optimal QST for a protocol. Once an optimal QST has been obtained,

the threshold can be set in GSEQ for subsequent analyses. Although Affymetrix's technical note suggests using a QST of 3, we have found that we can go down to a QST of 1 to increase the call rate without increasing our false negative rate. However, independent assessment of this threshold is suggested for each lab and their own experience.

Sample Size

The sample size of a batch is a critical issue to be considered when running an experiment. Because GSEQ uses all samples to obtain the local hybridization pattern for each site and to detect significantly different hybridizations that correspond to true positive variants, individual chip performance heavily depends on a sufficiently large sample size. Call rates increase, and false negative and false positive rates improve when the sample size is increased. A detailed description of the study is provided in the RA v2.0 technical note. The analysis presented only analyzed sample sizes of up to 16. In our experience, one should use a sample size of at least 20, preferably 30 or more. Although gradual improvements continue to occur, the additional improvement in performance is less substantial beyond 20 samples.

GC Content

The GC content of a probe can affect sequence and region-specific hybridization efficiency and specificity. Affymetrix has performed a study to assess the impact of GC content on the array performance and the results have been published in the RA v2.0 technical note. Average call rates exceed 90% for probes with GC content up to 70%, but the call rates drop considerably when the probe GC content exceeds 70%. High GC content of a probe could thus lead to localized no-calls and missed variant calls, since single substitution has less impact on a probe with high GC content as compared to those with moderate or low GC content. In particular, stretches of repeated GC base pairs have been seen to perform quite poorly. The data should be assessed for high GC content of low-performance probes. Filtering calls attributed to high GC content can alleviate the false positive rate.

PROCESSING GSEQ CALLS FOR REDUCING FALSE POSITIVE CALLS

PCR failures, nearby SNP effects, cross-hybridizing probes, low sequence complexity probes, and non-biallelic calls have been identified as the major factors that contribute to false positive calls (RA v2.0 technical note). Affymetrix recommends that the calls be assessed for these conditions, and they should be filtered if any of these factors are found to be responsible. The technical note describes in detail the effect of these factors on chip performance. Some of the effects are described briefly below.

PCR Failure

Call rates can be examined as a function of the PCR amplicon in a sample, and all calls within amplicons that have an average call rate of less than a certain threshold (e.g., ~85%) should be converted to a failed region. Such regions can be sequenced with dideoxy sequencing to fill in the sequencing gaps.

Near SNP/Footprint Effect

If there is a homozygous variant, the performance of neighboring probes, which assume a wild-type base at the variant position, can be poor. Therefore, clusters of SNP calls (i.e., SNPs within 9 bases of each other) should be interpreted with caution. The variant call with the highest quality score should be retained, and the rest should be converted to no-calls. Confirmation of the sequence by an independent method is recommended. To avoid this problem, common SNPs should be determined in the sequence of interest before designing the array, and redundant stretches of resequencing probes should be tiled with the minor allele substituted. To make optimal automated use of these stretches of redundant minor allele sequence, the common SNPs can also be tiled using multiple offset probes that can be analyzed by genotyping software (e.g., BRLMM, Affymetrix) and a computer algorithm can be used to include or exclude the appropriate resequencing stretches depending on the genotype.

Cross-Hybridizing Sequences

False variants resulting from cross hybridization can be reduced by comparing the 25 bases surrounding each variant to the sequence of the hybridization target using BLAST (*UNIT* 6.8). According to a study conducted by Affymetrix, 6% of the false positives could be eliminated without compromising the true variants when any calls with 24 bases matching a homologous locus were eliminated. Unfortunately, this prevents accurate analysis of the homologous loci within a single array.

Low-Complexity Sequences

Variant calls at sites in probes with low sequence complexity (repetitive bases) have a higher likelihood of being incorrect. The LZW compression algorithm (Ziv and Lampel, 1978) is one of the methods that can be used for identifying low-complexity sequences. Calls can be removed at these sites by implementing a suitable threshold and a method for determining sequence complexity.

CUSTOMIZED FILTERS AND CustomSeq ARRAY TECHNICAL FOLLOW-UP

After the appropriate settings have been achieved to optimize call rates and accuracy, there will still be positions on an array that remain a “no-call” or variant call. A large portion (~70% in our experience) of the variant calls will be false positives. However, most of these are the result of consistent background hybridization signals at these positions. As such, a filter can be installed to screen out common false positives. For example, if the wild-type base at a given position is a G and there is often a signal at the A probe that leads to a G/A heterozygote call, then A/G heterozygous calls at this position can be excluded from follow-up. Similarly, we also found it necessary to exclude positions that frequently were not called by GSEQ.

In our experience, using an array that sequenced 21,591 bases in triplicate, we had 46 false positive variant calls and 134 no-calls sites (all three probe sets were no-calls) that happened in 20% or more cases and were therefore filtered out and excluded from follow-up. For all other no-calls and variant calls, we feel that it is important to resolve the call at these

positions using an independent method, such as dideoxy capillary-based sequencing. Using this approach we typically need to sequence 15 out of 140 amplicons.

Reduction of no-calls and reduction of the need to perform confirmatory capillary resequencing can be attempted by designing improved computational algorithms. For example, for the detection of rare mutations in clinical disease settings, the use of an algorithm that compares a single patient data set to data from a group of controls is likely to improve call rates and false positive rates through a statistical comparison of signal strength for each probe. This approach is likely to resolve background signals that result in no-calls and false positive calls. Such an algorithm has been recently developed by JSI Medical Systems (<http://www.jsi-medisys.de/html/products/SeqC/SeqC.htm>), although evaluation of the software's performance is pending.

CustomSeq Arrays and Detection of Insertions and Deletions

Detection of insertions or heterozygous deletions (indels) poses a significant challenge in resequencing arrays, as they are difficult to detect and not called automatically by the software. A previously characterized indel can be inspected using custom probes. However, a novel indel is difficult to predict. A reduced signal may indicate the presence of a novel indel but the pattern is difficult to predict due to its dependence on the length of the indel, which is highly variable. Algorithm development aimed at identifying possible deletions by detecting reduction in signal intensity is a theoretical approach, particularly to detect deletions confined to within a PCR amplicon. Deletions larger than a PCR amplicon could be missed due to nonquantitative PCR and normalization of products prior to hybridization. In our experience, the rate of detection of novel small indels, without implementation of more sophisticated analysis, is ~63%. These were detected by follow-up of rare no-call positions and variant calls. If specific probes are designed to detect known indels, sensitivity increases to ~95%.

CONCLUSION

CustomSeq arrays offer a high-throughput sequencing method that can be used to identify nucleotide variations along contiguous or non-contiguous genomic regions. When employed on a large scale, microarray-based screening offers advantages due to its accuracy, efficiency, and cost-effectiveness. Although it currently must be used as a screening tool with variant calls and rare no-calls being confirmed through an independent method, the overall accuracy and sensitivity support its use in both research and clinical applications. While detection of insertions, deletions, and inversions remains a challenge for resequencing arrays, the approach provides a fast, automatable, and cost-effective method for repetitive resequencing of targeted regions of DNA.

LITERATURE CITED

- Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A. High-throughput variation detection and genotyping using microarrays. *Genome Res.* 2001; 11:1913–1925. [PubMed: 11691856]
- Lipshutz R, Fodor S, Gingeras T, Lockart D. High-density synthetic oligonucleotide arrays. *Nat Genet.* 1999; 21:20–24. [PubMed: 9915496]

Ziv J, Lempel A. Compression of individual sequences via variable-rate coding. *IEEE Trans Inf Theory*. 1978; 24:530–536.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 7.17.1. CustomSeq Array Design. Eight 25-mers are tiled to probe each individual base pair. A single position differs so that each of the four possible nucleotides can be queried on both the sense and antisense strands. Reproduced from the Affymetrix Data Sheet: GeneChip CustomSeq Resequencing Array Program.

Figure 7.17.2.

Schematic overview of the Affymetrix resequencing chip procedure. Image courtesy of Affymetrix.

Figure 7.17.3.
Affymetrix File Types. Adapted from Affymetrix GeneChip Sequence Analysis Software User Guide.

Figure 7.17.4.

Examples of an array image with and without correct grid alignment. Solid white lines show the grid boundaries. (A) Correct grid alignment. (B) Misaligned grid.

Figure 7.17.5.
Overview of Resequencing Algorithm Version 2.0.

Figure 7.17.6.
Examples of hybridization patterns. (A) A homozygous call. (B) A heterozygous call.

Table 7.17.1

Performance Measures for Base-Calling by GSEQ

Performance measure	Description
Average call rate	Total bases called ^a /total bases interrogated by array
Base call accuracy	Correct base calls/total bases called ^a
Total false positive rate	False variant calls/total bases called ^a
Total false negative rate	Variants falsely called WT/total bases called ^a
Variant false positive rate	False variants called/total variant calls ^a
Variant false negative rate	Variants falsely called WT/total true variants ^b

^aExcluding no-calls.

^bDetermined by an independent method.