Published in final edited form as:
Curr Protoc Hum Genet. 2011 April ; 0 7: Unit7.18. doi:10.1002/0471142905.hg0718s69.

Targeted Sequencing Using Affymetrix CustomSeq Arrays

Polakit Teekakirikul[1,2], **Stephanie Cox**[2], **Birgit Funke**[1,2,3,4], and **Heidi L. Rehm**[1,2,3,4]

[1]Harvard Medical School, Boston, Massachusetts

[2]Partners HealthCare Center for Personalized Genetic Medicine, Cambridge, Massachusetts

[3]Brigham and Women's Hospital, Boston, Massachusetts

[4]Massachusetts General Hospital, Boston, Massachusetts

Abstract

This unit provides a basic protocol for oligo hybridization-based sequencing technology and resulting data analysis specific to the Affymetrix GeneChip CustomSeq Resequencing Array platform. All steps and critical aspects related to array design, experimental protocols, data management, and base-calling algorithms are addressed. This unit is particularly appropriate for sequencing targeted regions of the genome of up to 300 kilo-bases. The basic technology is most suitable for detecting substitution mutations, unless targeted indel probes are added.

Keywords

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

INTRODUCTION

DNA sequencing has become a pivotal step in elucidating the genetic basis of human disease and clinically significant genetic variation. As we gain more understanding of genes, pathways, polymorphisms, and allelic interactions with respect to their roles in human disease, cost-effective, high-throughput DNA resequencing technology is clinically required for affected individuals and their family members. With the advent of recent molecular methods for constructing large scale, sequence-specific microarrays, robust techniques for gene resequencing have emerged, based upon comparing hybridization intensity to a series of probes interrogating all possible DNA bases at successive positions.

Internet Resources

http://www.affymetrix.com/support/technical/byproduct.affx?product=cseq
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's 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.

An overview of the principle of oligo hybridization-based resequencing is provided in *UNIT 7.17*. This unit is its sequel, which describes a Basic Protocol for using the technology of oligo hybridization-based resequencing specific to the Affymetrix GeneChip CustomSeq Resequencing Array platform. Detailed information about this technology is provided on the Affymetrix Web site (see Internet Resources), and includes array design guides, protocols, and other support materials. Therefore, this unit will focus more on critical aspects of the design, execution, and analysis of these arrays that may not be comprehensively addressed in available materials.

At present, 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 a relatively well understood variant structure.

CRITICAL ASPECTS OF DESIGN, EXECUTION, AND ANALYSIS OF AFFYMETRIX GeneChip CustomSeq ARRAYS

An overview of the custom sequencing array assay steps is provided in Figure 7.17.2. The Affymetrix GeneChip system consists of a probe array, hybridization oven, fluidics station, scanner, and computer workstation. PCR is used to select areas of interest from the genomic DNA. Targeted regions of interest are PCR-amplified and then pooled to obtain one sample. The pooled, amplified DNA is subsequently fragmented into smaller pieces of DNA, which are end-labeled with a biotinylated nucleotide analog, using terminal deoxynucleotidyl transferase (TdT), and then hybridized to the array. The staining process consists of an initial stain with a fluorescent streptavidin-R-phycoerythrin conjugate (SAPE), which binds the biotins. There is also a secondary stain for signal amplification where biotinylated anti-streptavidin antibodies are bound to the initial SAPE molecules and subsequently stained with a second SAPE addition. After washing, the amount of fluorescence is monitored in a scanner. The fluidics station and the scanner are operated from the workstation with GeneChip operating software (GCOS). A diagram of a pipeline employing these systems and files can be seen in Figure 7.18.1.

This software also controls image acquisition and database management. The scanner is a wide-field, epifluorescent, confocal microscope that uses a solid state laser to excite fluorophores bound to hybridized nucleic acids. During the scan process, a photomultiplier tube collects and converts fluorescent values into an electronic signal, which is then converted into the corresponding numerical values. These numerical values represent the fluorescent intensities, which are stored as pixel values that comprise the image data file (DAT file). The raw image from a DAT file is further processed to generate a CEL file. Next, GeneChip sequence analysis software (GSEQ) allows the user to perform sequence analysis of the data from a CEL file to produce the final sequence calls for every position on the chip—A,C,T,G, M(A and C), R(A and G), W(A and T), Y(C and T), S(C and G), K(G and T), and N (no-call)—along with quality scores representing data confidence. The user can then manipulate the data through customized computer scripts to further refine the output and facilitate follow-up of variant calls and no-calls.

NOTE: All reagents are stable for one year unless otherwise noted. All equipment and reagents are stored at room temperature unless otherwise noted.

Materials

AmpliTaq Gold and buffers (Applied Biosystems cat. no. 4311858; store at −20°C)

2.5 mM $MgCl_2$

5 M Betaine (Sigma cat. no. B0300)

Dimethyl sulfoxide (DMSO; Sigma cat. no. D9170)

10 mM dNTP mix: dilute 100 mM PCR-grade dNTP stock solutions (Roche cat. no. 1969064; store at −20°C), and combine equal volumes of the four 10 mM dNTPs

10 μM synthesized oligonucleotide primer mix: dilute 100 μM stocks solutions (Integrated DNA Technologies; store at −20°C), and combine, e.g., two to three primer pairs per mix

25 ng/μl high-quality genomic DNA (~10 μg), e.g., prepared from human peripheral blood samples (*UNIT 14.4*)

Human genomic DNA (positive control; Clontech cat. no. 6550-1; store at 4°C)

Molecular biology–grade water (HyClone)

Low-range quantitative DNA ladder (Invitrogen cat. no. 12373-031)

QIAquick PCR purification kit (Qiagen cat. no. 28104)

GeneChip Resequencing Assay kit (Affymetrix cat. no. 900447; store at −20°C) includes:

Fragmentation reagent (DNase I) and buffer

DNA labeling reagent

Terminal deoxynucleotidyl transferase (TdT) and buffer

130× oligo control reagent

2 and 4% (w/v) agarose gels: general purpose 2% (e.g., E-gel 96, Invitrogen cat. no. G7008-02) and high-resolution 4% (e.g., Invitrogen cat. no. G6200-04)

5 M tetramethylammonium chloride solution (TMAC; Sigma cat. no. T3411)

1 M Tris·Cl, pH 7.8 (Sigma cat. no. T2913; *APPENDIX 2D*)

1% (v/v) Tween-20 (Sigma cat. no. P1379)

50 mg/ml acetylated bovine serum albumin (BSA) solution (Invitrogen cat. no. 15561-020; store at −20°C)

10 mg/ml herring sperm (Promega cat. no. D1811; store at −20°C)

GeneChip Hybridization Oven 640 (Affymetrix)

GeneChip Fluidics Station 450 (Affymetrix)

GeneChip Scanner 3000 7G (Affymetrix)

GeneChip operating software (GCOS) version 1.4 (Affymetrix)

GeneChip sequence analysis software (GSEQ) version 4.0 (Affymetrix)

Library files for custom arrays (Affymetrix)

1.5-ml microcentrifuge tubes

96-well semi-skirted plates (Fisher Scientific cat. no. 08-408-250)

384-well skirted plates (ABGene cat. no. AB-1111)

Adhesive PCR film (ABgene)

Benchtop centrifuge with plate adaptors (e.g., Eppendorf 5810R)

GeneAmp PCR system 9700 (Applied Biosystems)

Gel electrophoresis equipment, e.g., d E-gel PowerBase, Version 4 (Invitrogen cat. no. G6200-04) *or* Agilent 2100 Bioanalyzer

Biomek FX Robot (Beckman Coulter), recommended

2-ml screw cap tubes

Rotary vacuum evaporator (SpeedVac DNA 110, Savant)

96-well semi-skirted plates (Fisher Scientific cat. no. 08-408-250)

37°C, 49°C and 95°C digital dry block heater (e.g., VWR cat. no. 12621084) with thermometer (0°C to 110°C)

Conical centrifuge tubes of appropriate sizes (e.g., 15- or 50-ml Falcon tubes, Fisher Scientific)

Additional reagents and equipment for carrying out agarose gel electrophoresis (*UNIT 2.7*), determining DNA concentration (*APPENDIX 3D*), and performing GeneChip hybridization, washing, and staining steps according to Affymetrix protocols (see Internet Resources)

Select sequences and design arrays

The steps for designing an array are detailed in the Affymetrix GeneChip CustomSeq Resequencing Array Design guide (see Internet Resources). Several more detailed aspects of the design process are discussed in the Commentary section and *UNIT 7.17*. Critical steps of sequence selection and array design are outlined below.

1 Retrieve reference sequences from the NCBI databases for the regions of interest.

2 Run a homology search among regions to be sequenced to predict likelihood of cross-hybridization. If necessary, separate highly homologous sequences on different arrays to prevent cross-hybridization.

> NOTE: *Affymetrix will also assess cross-hybridization during the design submission process, although a general up-front assessment may be useful.*

3 *(Recommended)* Prior to submission of a design request, develop and test primer design and PCR amplification using the sequence to be interrogated to ensure that the up-front PCR can be performed robustly. See the Commentary section (Critical Parameters and Troubleshooting) for more recommendations regarding target amplification.

> Primers should be placed outside the sequence to be interrogated on the array.

4 Using all available online databases (e.g., dbSNP, locus-specific databases) and control sample sequencing, determine the spectrum of common variation across the regions of interest.

> To avoid allele dropout of rare variants through poor hybridization due to the presence of a nearby common variant, redundant tiling of sequences with the minor allele of a common variant allows detection of nearby rare variants that may occur on the minor allele background. The additional sequencing probe sets (e.g., triplicate set) containing the minor allele can correct hybridization when the minor allele of a common variant presents.

5 If probes for interrogating specific genotypes will be included (e.g., reported indel mutations), assemble a database of these variants.

> Genotyping probes add extra sensitivity to the chip, particularly for detecting insertion and deletion mutations.

6 Create sequence files in FASTA format (see *UNIT 6.8*) before submitting to Affymetrix. Include in the FASTA file all sequences for which probes will be designed. In addition, create an instruction file in a tab-delimited text format (`*.txt`) to provide a summary of the start and end positions for each contiguous fragment tiled on the array. Generally, the submitted files include the following:

Sequence FASTA file for resequencing

Sequence FASTA file for minor allele tiling

Sequence FASTA file for genotyping

Instruction file for resequencing

Instruction file for minor allele tiling

Instruction file for genotyping

Map files for locating minor allele tiling and genotyping probes in resequencing tiling.

7 Perform quality control on these files to make sure they contain the expected content as follows:

Check deletions, insertions, and first exons of each gene.

Check at least one minor allele and one genotype for each gene.

Ensure that the correct variants are in the right files and none are missing.

Check that the probes are the right length.

8 Fill in the electronic Resequencing Design Request form, which can be downloaded from Affymetrix Web site. Send this request form along with all files to chipdesign@affymetrix.com.

The Affymetrix GeneChip CustomSeq Resequencing Array team will review the submission and communicate with user to clarify details of design.

Once the array is designed at Affymetrix, all design files will be returned for approval.

9 If any modification is necessary, send updated files and notification to Affymetrix to redesign the array. Spot check a variety of probes throughout the design files to ensure that no systematic errors have occurred.

Within the resequencing file, every probe is listed with the four possible bases. Note that the probes within these files are either in reverse complement and reverse sequence (for the sense) or reverse sequence (for the antisense).

10 If all files are completely correct, confirm with Affymetrix to start manufacturing the chips.

Affymetrix will then perform mask design and complete all steps as well as confirm due date to ship the array.

11 Set up Affymetrix equipment (including, hybridization oven, fluidics station, and scanner) for processing the Affymetrix GeneChip CustomSeq Resequencing Array platform. Install GCOS and GSEQ software required for operating the resequencing assay and for data analysis, respectively. Load library files for each custom array.

Amplify target DNA

12 Prepare the PCR master mixes for 15-μl PCR reactions for all amplicons and positive and negative controls in 1.5-ml microcentrifuge tubes as follows:

0.2 μl of 5 U/μl AmpliTaq Gold

1.5 μl of 10× AmpliTaq Gold PCR buffer

1.5 μl of 2.5 mM $MgCl_2$

3.0 μl of 5 M betaine

0.75 µl DMSO

0.15 µl of 10 mM dNTP mix

2.0 µl of 10 µM oligonucleotide primer mix

1.0 µl of 25 ng/µl high-quality genomic DNA prepared for sequencing, *or* human genomic DNA (positive control) *or* molecular biology–grade water (negative control)

4.9 µl molecular biology–grade water.

The quantity of DNA required for sequencing depends on the number of PCR reactions needed for each array, but 10 µg is typical.

Ensure the reactions include positive controls (human genomic DNA) and negative water controls for all primer sets.

Generally, we use two to three primer pairs for each multiplex PCR reaction for all amplicons and positive and negative controls.

13 Dispense 15-µl reaction mixes to the wells of 384-well plates, seal the plates with adhesive PCR film, and centrifuge the plates briefly (1 min) at maximum speed on a benchtop centrifuge.

14 Carry out PCR in the GeneAmp thermocycler, using the following amplification cycles.

Initial step:	10 min	95°C	(denaturation)
30 cycles:	30 sec	95°C	(denaturation)
	30 sec	60°C	(annealing)
	30 sec	72°C	(extension)
Final step:	10 min	72°C	(final extension)
Hold:	indefinite	8°C.	

15 Analyze 3 µl of the PCR products by agarose gel electrophoresis (e.g., see *UNIT 2.7*) to confirm general success and a clean negative control. Include a low-range quantitative DNA ladder.

We use a 2% agarose E-gel 96 (Invitrogen) which can be loaded robotically and run in 5 minutes.

Prepare DNA for array protocol

16 *(Optional)* Normalize PCR products. See *UNIT 7.17*.

We find that the recommended normalization step is not required if using robust reproducible short-range PCR assays. For most products, we use 4 µl of a 15-µl PCR reaction. For repeatedly weaker products, we use 5 to 6 µl.

17 Pool 4 to 6 μl of the PCR products of all amplicons for each sample (genomic DNA begin tested or positive or negative controls) into a 2-ml screw-cap tube.

As this step can be labor intensive, we recommend using robotics, e.g., a Biomek FX robot with a span 8 head. Close observation of robotic operation throughout the program is encouraged as some errors may occur. Typical errors include: tips falling from the probes onto the deck, tips not being released into waste, and liquid not being transferred from the 384-well plate to the 2-ml screw-cap.

18 Purify the pooled PCR product using, e.g., a QIAquick PCR purification kit and the manufacturer's protocol.

More than one column may be needed depending on pooled PCR volume.

19a *For solutions >50 μl:* Evaporate excess fluid down to ~50 μl using a rotary vacuum evaporater.

19b *For solutions <50 μl*: Add molecular biology–grade water.

20 Determine DNA concentration using a spectrophotometer (e.g., see *APPENDIX 3D*).

Concentrations are usually between 50 and 200 ng/μl.

21 Plate 45 μl of each sample in a 96-well plate for the next fragmentation step.

Fragment and label DNA for array protocol

The following steps vary depending on the format of the chip. Below is an example using a 49-format chip (for types of formats, see Table 7.18.2).

22 Prepare the 6-μl fragmentation master mix for each 51-μl reaction as follows:

0.04 μl of 3 U/ml fragmentation reagent

3.28 μl fragmentation buffer

2.68 μl water.

Adjust the volume of the fragmentation reagent if the enzyme activity differs, also adjusting the volume of water to retain the 6-μl master mix volume.

23 Add the 6 μl of fragmentation master mix to the 45 μl of pooled PCR product in the 96-well plate (total volume 51 μl), and incubate 30 min at 37°C and 15 min at 95°C.

24 Determine the success of the fragmentation by performing high resolution agarose gel electrophoresis (e.g., see *UNIT 2.7*) of a 5-μl sample of the reaction mix on a 4% gel (e.g., 4% agarose E-gel include low-range quantitative DNA ladder) or by using an Agilent 2100 Bioanalyzer.

The electrophoresis should result in a smear of 20- to 200-bp fragments.

Fragmentation is one of the most variable 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.

In the case of underfragmentation, we have added two to three times as much as the initial fragmentation reagent to the reaction mix and re-incubated the underfragmented product. The volume of fragmentation reagent and incubation time are determined by experience.

Label DNA for array protocol

25 Prepare 18.3 μl (includes 5% extra) of labeling master mix for each reaction as follows:

12.6 μl of 5× TdT buffer

2.1 μl DNA labeling reagent

3.6 μl TdT.

NOTE: *During the labeling of DNA, the terminal deoxynucleotidyl transferase should not be removed from the freezer until ready to be added. It should be placed on ice or in an enzyme freezer block during use because this enzyme is temperature sensitive.*

26 Add 17.5 μl of the labeling master mix to 46 μl of each DNA sample (final reaction volume 63.5 μl), and incubate 2 hr at 37°C and 15 min at 95°C.

Hybridize, wash, and stain GeneChips

More details about the hybridization, staining, and wash steps section, including additional materials that may be required, are provided on the Affymetrix Web site (see Internet resources).

27 Using conical centrifuge tubes of an appropriate size, prepare 210 μl (including 5% extra) of prehybridization buffer master mix as follows:

2.1 μl of 1 M Tris·Cl, pH 7.8

2.1 μl of 1% Tween 20

205.8 μl water.

Add 200 μl of prehybridization buffer to each array, and prehybridize in the hybridization oven according to the manufacturer's protocol.

These volumes are for the 49-format chip.

28 Prepare, e.g., for each reaction in the 49-format chip, 168.0 μl (includes 5% extra) of hybridization buffer master mix as follows:

138.6 µl of 5 M TMAC

2.3 µl of 1 M Tris·Cl, pH 7.8

2.3 µl of 1% Tween 20

2.3 µl of 50 mg/ml BSA

2.3 µl of 10 mg/ml herring sperm DNA

1.8 µl of 130× oligo control reagent

18.4 µl water.

29 Add 160 µl of hybridization buffer to the labeled sample from step 26 (final volume 223.5 µl). Denature the sample 5 min at 95°C, and 5 min at 49°C, using a dry block heater with thermometer.

30 After prehybridization (step 27) is complete, remove the chips from the hybridization oven, discard the prehybridization buffer, and load 200 µl of the labeled sample (step 29) into the corresponding chip. Hybridize, and then stain according to the manufacturer's protocol.

After the hybridization step, we simply supply the volumes of master mixes required by the manufacturer's protocol, and additional steps (e.g., staining) are carried out by the fluidics station.

Scan GeneChips and output data

31 Scan the chips.

32 When the scan of each chip is complete, find the experiment in the image data tree, and drag and drop it into the right-hand window. Right click on the image and select image settings. Set the intensity to 30,000 and the color to Pseudo Color. Click OK (see Fig 7.18.2).

If the intensity of the image is not very bright, it can be adjusted by right clicking on the image. If the signal is still weak, the hybridization may have been poor.

This is a first check for evaluating technical performance. The call rate from the GSEQ software is a more detailed indication of the success of the experiment.

33 Use the files listed in Table 7.18.1 to output the data for a resequencing array experiment in the following steps.

Align the grid

Following the scan, GCOS software automatically generates average signal intensities of each probe by superimposing a grid on the scanned image in such a way that each square in the grid encloses a single probe cell. The goal of the grid alignment is for each square of the grid to delineate a single probe cell. However, manual grid alignment is required when the algorithm fails to perform automatic alignment, or sometimes when poor call rates are obtained

34a *If a grid has failed to align:* Go to step 35.

The error message "Failed to align grid" which may appear after the scan of the chip, prompts manual grid alignment.

34b *If a grid must be realigned:* First delete the original CEL file from the GCOS software as follows:

i. In the start menu under programs find the Affymetrix tab and choose GCOS Manager.

ii. When the GCOS Manager software appears, open the Resequencing folder under samples in the data tree on the left part of the screen.

iii. Click on the experiment with the CEL file to be deleted.

The experiment information should come up in the right-hand window.

iv. Right click on the CEL file and choose DELETE. Proceed to step 35.

Now the experiment has no CEL file and the grid can be realigned and a new CEL file created.

If any experiment has a poor call rate after the GSEQ batch analysis and/or a high rate of nonreference bases called for a chip that looks good upon visual inspection, then an examination of the grid alignment should be performed. An overhybridization white spot of signal in a corner region or something similar could also be a sign that the grid alignment did not perform accurately.

35 In the GCOS software, open the DAT file of the experiment requiring realignment and open the DAT file of an experiment in which the grid was automatically correctly aligned.

36 On the correctly aligned DAT file, click on the GRID button found in the lower right-hand corner of the DAT image. A checkered grid should appear. Keep this file open.

37 On the incorrectly aligned DAT file, click on the GRID button found in the lower right-hand corner of the DAT image. A rectangular box should appear. Use the cursor to pull the box's corners out to where they should roughly be (refer to the correctly aligned image for help).

38 Then highlight a corner and click IN (found on the lower right-hand corner of the DAT image, next to GRID). Align grid by using the correctly aligned DAT image as a guide. Do this for all four corners.

39 Examine the edges and make sure the lines are not going through the middle of any of the squares of signal.

40 Once correctly aligned, right click on the DAT file, and choose Realign SubGrids from menu.

41 Right click again and choose “Recalculate Cell Intensity.” The software has now made a CEL file for this experiment. Click SAVE.

Analyze data using GSEQ

GSEQ uses the Resequencing Algorithm Version 2.0 (RA v2.0), a base-calling method, to automate the generation of sequence and genotype calls from hybridization intensity data produced from 8 μm feature arrays. The RA v2.0 has been built upon the RA v1.0 and Adaptive Background Genotype Calling Scheme (ABACUS) developed by Cutler and colleagues (Cutler et al., 2001). A detailed description of the algorithm is provided in *UNIT 7.17* and the Affymetrix GSEQ User Guide (see Internet Resources). The RA v2.0 consists of three main steps: processing filters, base-calling method, and applying final reliability rules.

42 To begin performing data analysis using GSEQ, go to START button on Affymetrix computer used for scanning the latest batch and select GSEQ software from Affymetrix Folder.

Apply processing filters

43 Open the cell intensity tree on the right-hand side of the screen to see the available CEL files. If the scanned experiment is not listed, go to tools and then filters. Refresh the filters with the relevant date to reveal experiments.

Set the base-calling method

44 Go to the run tab and select batch analysis. A new window will appear.

45 Adjust the resequencing algorithm settings under the tools menu, optimizing settings for each user. As an example, we suggest the settings in Table 7.18.2.

46 For subsequent analyses, check these settings before each run.

NOTE: *Automated overnight analysis is not possible due to the inability to ensure these settings.*

47 Close the algorithm settings box without saving unless changes are required.

Complete analysis

48 Drag and drop all the CEL files in the predefined control set, as well as the newly scanned chips from the left-side window into the right-side input window.

We recommend using at least 20 and preferably 50 to 100 chips in the control set.

49 Rename the report file with the batch number of the current batch (the software defaults to naming it as the first CEL file that is put into the batch analysis). The experiments in the previously specified data set will appear in red and the software will ask if you want to rewrite them. Click YES.

50 Click the middle button above the input window that has "ACGT" on it, and the analysis will start.

Alternatively, go to the edit menu and select start analysis from the drop down menu. It usually takes ~10 to 15 min to complete.

51 Once the analysis is complete and the report appears, quickly scan the call rates to ensure that there are no major issues.

52 Then open the analysis results tree on the left-hand side of the screen. Right click on the newly scanned experiments and select OPEN. Open all the experiments from the current batch, but not the previous experiments used in the batch analysis from the control set.

53 When all experiments are opened, click on EXPORT button, and export the table to the data storage folder and label it with the appropriate batch number.

This is a text file, which is necessary for further analysis of the resequencing data. More information about further analysis is described in UNIT 7.17.

COMMENTARY

Background Information

General considerations for array design—An array consists of oligonucleotides (25-mers), chemically synthesized at specific locations on a coated quartz surface. The precise location where each probe is synthesized is called a feature, and millions of features are contained on one array. By extracting and labeling nucleic acids from experimental samples, and subsequently hybridizing those prepared samples to the array, the amount of signal can be monitored at each feature. For resequencing arrays, each base of interest is queried by one or more sets of eight probes (four probes in the sense direction and four probes in the antisense direction).

Array capacity, which defines the total number of bases that can be sequenced, depends on array formats. Affymetrix currently provides three different formats including the 49-, 100-, and 169-formats, which enable the analysis of up to 300 kb, 100 kb, and 50 kb, respectively. Table 7.18.3 presents critical specifications for each flexible array format provided by Affymetrix Resequencing Array Program.

It should be noted that true capacity is slightly lower than the noted ranges due to some wasted space for designs containing discontinuous DNA sequences (typical of exon sequencing). The following formula is used to determine the number of bases sequenced on a single array.

The number of bases sequenced = maximum capacity (or total amount of sequence submitted if less than maximum capacity) − [the number of discontinuous fragments × 24].

This equation provides for an additional 24 flanking bases that are accounted for when using a shift mask design and discontinuous fragment sequencing.

Critical Parameters and Troubleshooting

Cross-hybridization—Repetitive elements and internal duplications can give rise to cross-hybridization. Thus, they should be removed prior to sequence submission. Also, highly homologous sequences leading to cross-hybridization and reduced data quality for that particular sequence should be identified. If necessary, highly homologous regions should be tiled on separate arrays and amplified using unique primer sequences.

Minor allele tiling—Familiarity with the target sequence is critical. In order to detect variation that could occur on different haplotypes (e.g., a mutation that occurs within 13 bases of a nonreference sequence variation), knowledge of all common gene variations is useful. In this case a redundant stretch of sequencing probes for 25 bases, centered on the minor allele, is tiled to account for all probes that would be affected by the minor allele. To create these probes, 49 base pairs (25 interrogating base pairs +24 flanking base pairs) are submitted to Affymetrix. Figure 7.18.3 illustrates the interrogation position of sequence within 13 base pairs of a sequence variant.

Redundant tiling—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 all arrays used in our clinical diagnostics laboratory.

Offset tiling—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 are 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 and 17, in addition to the traditional 13 (center). It should be noted that an additional four bases must be provided at the end of each fragment to be sequenced if tiling the 9 and 17 offset probes (see Fig. 7.18.4).

Genotyping probes—To add increased sensitivity and specificity for indels as well as other previously identified variants, specific genotyping probes can also be tiled. Genotyping probes can be tiled exactly like minor allele tiling probes, or a shorter stretch of probes can be used. We find that the most sensitive approach to genotyping is to tile probes to sequence the variant plus four bases on either side.

These nine probes, which appear to be sequencing nine bases, are identical to a series of nine offset probes to interrogate a single base as illustrated in Figure 7.18.5. These probes, along with the corresponding wild-type set from the standard sequencing probes, are then combined and analyzed by a genotyping software such as BRLMM (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf). The CDF file must be altered in order to point the BRLMM software to the appropriate probes on the array for analysis.

Target amplification—Target enrichment strategy is required for every interrogated region. Our currently preferred method is locus-specific PCR amplification. Long-range PCR can sometimes be used to reduce the number of reactions required. However, it requires quantification and normalization before hybridization because long- range PCR is less robust and more variable in product quantities. Moreover, longer PCR products are more susceptible to under-fragmentation, thereby leading to less effective hybridization. As such, it is recommended to develop short-range PCR reactions that can be generated and purified automatically in 384-well plates using robotic devices. Short range PCR assays can also be usefully applied to dideoxy sequencing assay development for validating variants identified on the array. In order to decrease the total number of reactions, PCR amplification may be multiplexed. Based upon our experience, we have generated PCR products in 6-plex in 15 μl and are able to obtain sufficient product from each amplicon. It should be noted that the use of gel analysis is not necessarily a useful step in evaluating the success of PCR, given that, in our experience, absence of a band on a gel may not predict failure of that amplicon during chip sequencing. However, we do use gels to ensure the overall success of the PCR process including a clean negative control.

Quality control—Negative and positive controls for the PCR master mix should be included with each batch. There are two positive controls used to assess PCR performance for the premixed master mix. These include commercially available human DNA and the IQ-EX control template and 1.0 kb primer pair (referred to as the Affymetrix Positive Control) included in the GeneChip Resequencing Assay Kit (stored at −20°C). If the negative control fails (e.g., contamination), the whole reaction should be repeated. If the Affymetrix Positive Control does not amplify, but the other reactions do amplify, then the experiment can be continued without consequence.

A hybridization control (the Oligonucleotide Control Reagent, stored at −20°C) is provided by Affymetrix in the GeneChip Resequencing Assay Kit and is added to every reaction prior to hybridization. This control binds to the outer corners of the GeneChip, and therefore facilitates alignment of the grid. If the Oligonucleotide Control Reagent is not present on the scan of the GeneChip, then the chip must be designated a failure.

In a clinical environment, new lots of arrays should be validated upon receipt to assess the quality. To validate the new lot, we use the same hybridization solution (stored at −20°C) after removal from a previous array. We follow steps for prehybridization for an array from the new lot. While the chip is filled with prehybridization solution, the previously hybridized sample is heated for 5 minutes at 95°C and 5 minutes at 49°C, and then added to a chip from the new lot. All remaining steps for hybridization and staining are processed. We then compare all variant calls, and if the concordance is 90% or greater and the call rate for the current array is 95% or greater, then the lot of arrays is considered to pass quality assessment.

Array limitations—Detection of insertions or heterozygous deletions (indels) poses a significant challenge in resequencing arrays because they are difficult to detect and are not called automatically by the software (Zimmerman et al., 2010). Insertions and deletions that are not tiled using genotyping probes are not well detected by this method because probes specific for all possible insertions and deletions cannot be tiled. A previously characterized indel can be interrogated using custom probes; however, a novel indel is difficult to predict. A reduced signal may indicate the presence of a novel indel but in our experience this only leads to detection one-third of the time (through follow-up of no-calls and variant calls). If specific probes are designed to detect known indels, sensitivity increases to approximately 95%. Algorithm development aimed at identifying possible deletions by detecting reduction in signal intensity is a theoretical approach, particularly in detecting 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.

Anticipated Results

Once the analysis is complete, call rates should be between 90% and 99%. Although all parameters have been appropriately set up to optimize call rates and accuracy, some positions on each experiment will receive a nocall or variant call. In our experience, ~70% of the variant calls will be false positives. Most of the consistent false-positive calls, as well as the repetitive no-call positions, are due to background hybridization signals at certain positions. As such, a data filter can be installed to remove common false positive variant calls and common no-calls at these positions. For other no-calls and variants calls unique to an experiment, an additional sequencing method such as dideoxy sequencing can be used to resolve the calls at these positions. In addition, improved computational algorithms can be constructed and used for reducing no-calls and the need to confirm variant calls with capillary sequencing. Such algorithms have been developed by JSI Medical Systems (http://www.jsi-medisys.de/html/products/SeqC/SeqC.htm) and TessArae (http://www.tessarae.com). However, we have not evaluated these software tools, and therefore cannot comment on their usefulness. *UNIT 7.17* provides more detailed information and discussion on customized filters and array technical follow-up.

Data received in the form provided by the Affymetrix GSEQ software can prove difficult to use in an efficient manner. Therefore, we recommend developing a data analysis pipeline to simplify the output of the experiments and facilitate any follow-up confirmation testing. We show a flow diagram of the software components and customized data analysis pipeline that we have built for clinical testing (Fig. 7.18.1). This pipeline incorporates the various customizations and filtering steps described in this unit. It also shows the steps required to arrive at a final set of confirmed variants after incorporation of follow-up capillary Sanger sequencing. A sample of a summary report from our pipeline can be seen in Figure 7.18.6.

Time Considerations

PCR, gel quality control, and pooling of DNA take 1 day. Note that robotics is used for shortening these labor-intensive and time-consuming processes. PCR product purification and concentration are performed in 1 day. Fragmentation, gel electrophoresis, and labeling are processed in 1 day, followed by hybridization overnight for 16 hours. Staining, scanning, grid alignment, and data analysis take 1 day. Subsequently, dideoxy sequencing and data review for amplicons requiring follow-up are performed over an additional 3 days. Assuming no additional dideoxy sequencing is required, the entire assay can be completed in one week.

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]

Zimmerman RS, Cox S, Lakdawala NK, Cirino A, Mancini-DiNardo D, Clark E, Leon A, Duffy E, White E, Baxter S, Alaamery M, Farwell L, Weiss S, Seidman CE, Seidman JG, Ho CY, Rehm HL, Funke BH. Anovel custom resequencing array for dilated cardiomyopathy. Genet. Med. 2010; 12:268–278. [PubMed: 20474083]

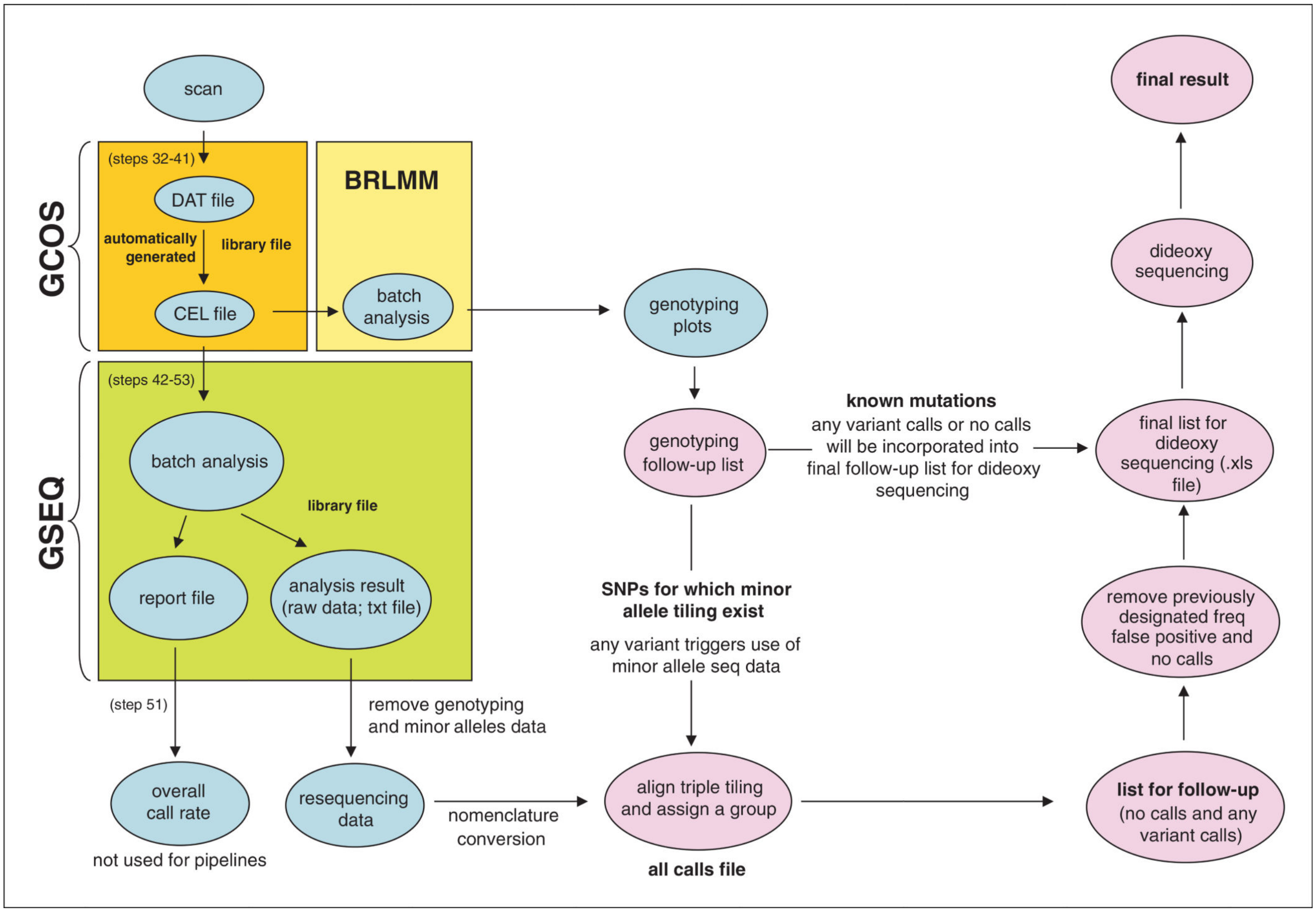

scan
GCOS
(steps 32-41)
DAT file
automatically generated
library file
CEL file
BRLMM
batch analysis
genotyping plots
GSEQ
(steps 42-53)
batch analysis
library file
report file
analysis result (raw data; txt file)
(step 51)
overall call rate
not used for pipelines
remove genotyping and minor alleles data
resequencing data
nomenclature conversion
genotyping follow-up list
known mutations
any variant calls or no calls will be incorporated into final follow-up list for dideoxy sequencing
SNPs for which minor allele tiling exist
any variant triggers use of minor allele seq data
align triple tiling and assign a group
all calls file
list for follow-up (no calls and any variant calls)
remove previously designated freq false positive and no calls
final list for dideoxy sequencing (.xls file)
dideoxy sequencing
final result

Figure 7.18.1.

Customized pipeline for data analysis. Abbreviations: BRLMM, genotyping software; CEL, cell intensity; DAT, data; GCOS, GeneChip operating software; GSEQ, GeneChip sequence analysis software; SNP, single-nucleotide polymorphism.

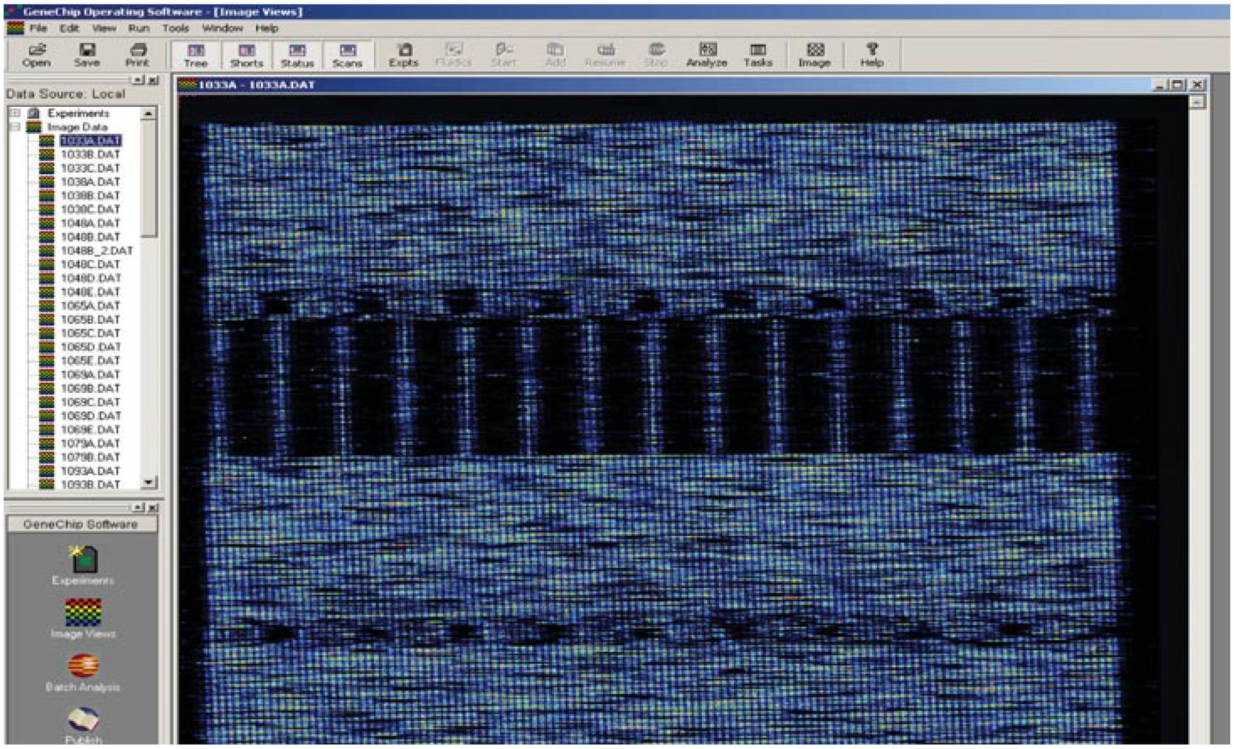


Figure 7.18.2.

DAT file shows an image of the scanned probe array.

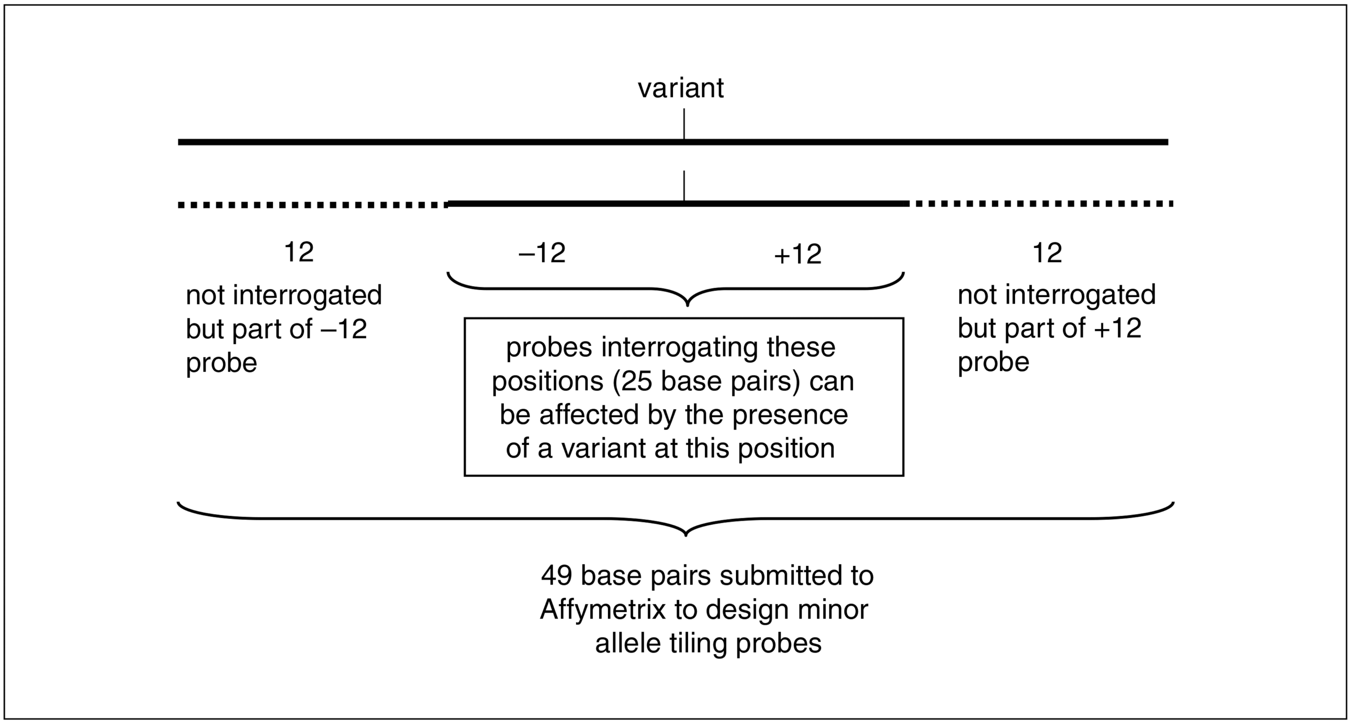

variant
12
not interrogated but part of –12 probe
–12
+12
probes interrogating these positions (25 base pairs) can be affected by the presence of a variant at this position
12
not interrogated but part of +12 probe
49 base pairs submitted to Affymetrix to design minor allele tiling probes

Figure 7.18.3.

Minor allele tiling probes include the variant and ±12 base pairs. A total of 49 base pairs is submitted to Affymetrix to create probes to sequence the affected region.

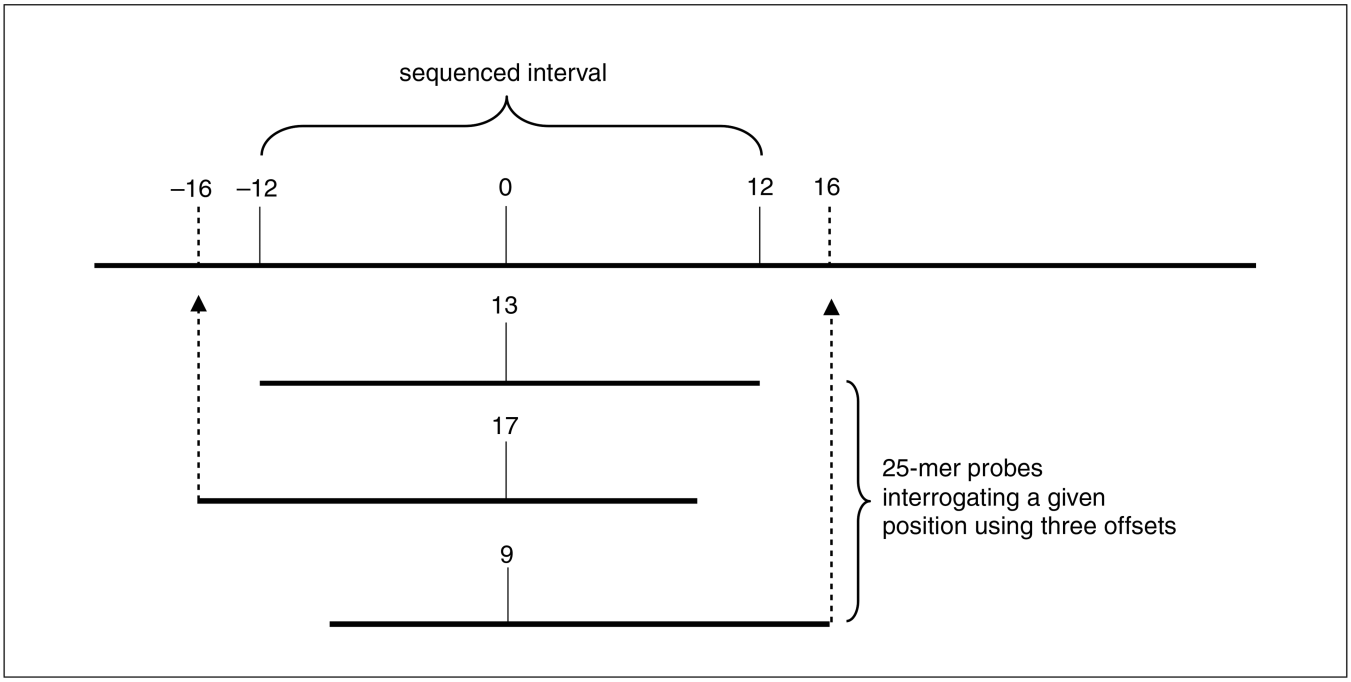


Figure 7.18.4.

Offset tiling uses additional probes with interrogating positions offset to the 5′ or 3′ end of the probe to increase sensitivity and specificity.

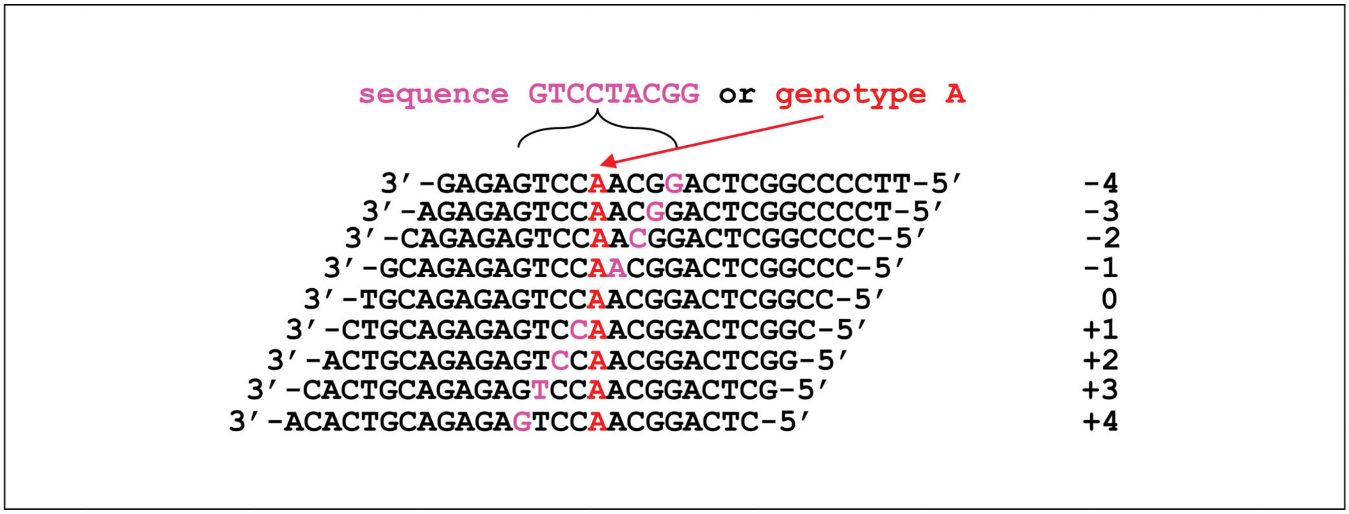
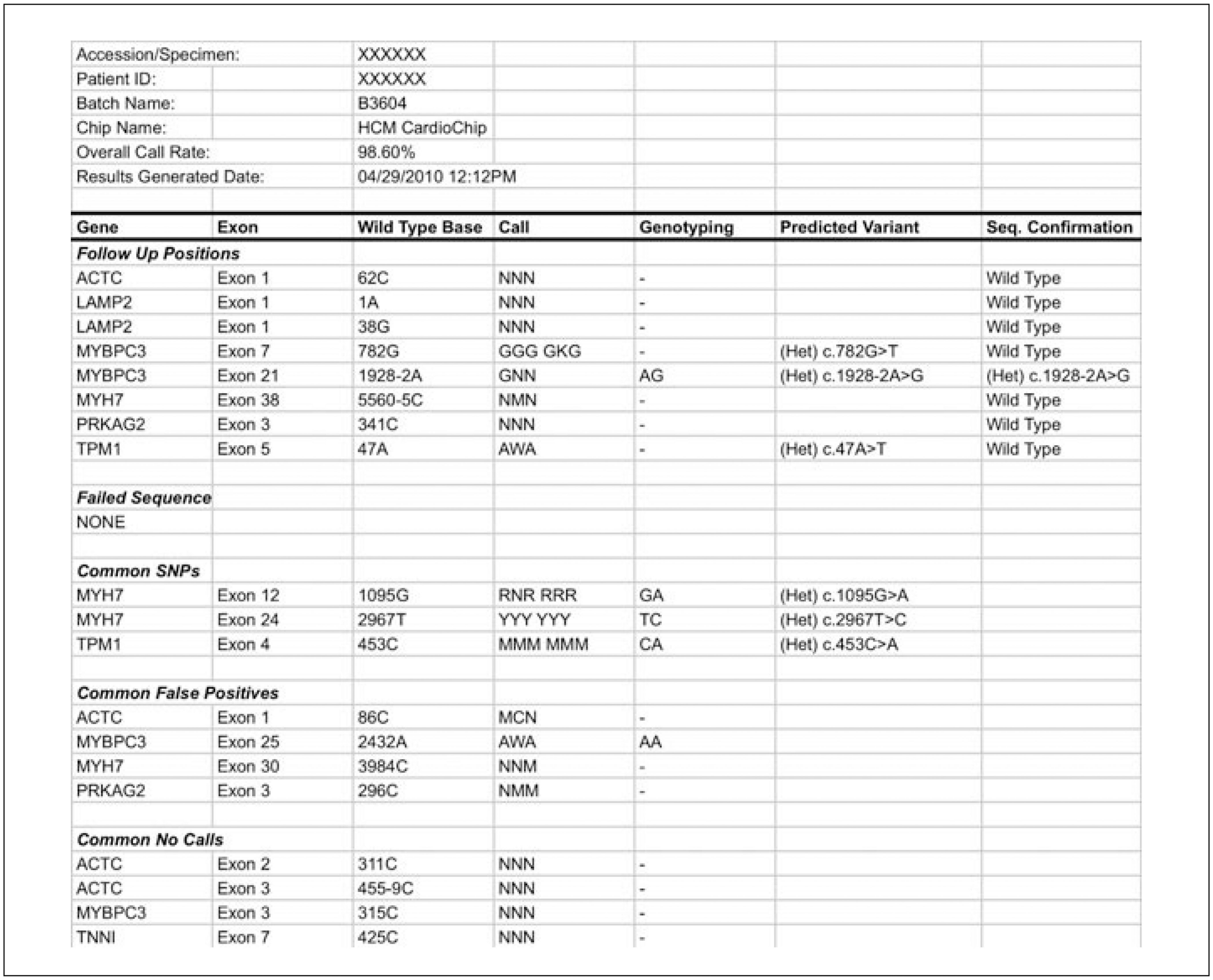

sequence GTCCTACGG or genotype A
3′-GAGAGTCCAACGGACTCGGCCCCTT-5′ -4
3′-AGAGAGTCCAACGGACTCGGCCCCT-5′ -3
3′-CAGAGAGTCCAACGGACTCGGCCCC-5′ -2
3′-GCAGAGAGTCCAACGGACTCGGCCC-5′ -1
3′-TGCAGAGAGTCCAACGGACTCGGCC-5′ 0
3′-CTGCAGAGAGTCCAACGGACTCGGC-5′ +1
3′-ACTGCAGAGAGTCCAACGGACTCGG-5′ +2
3′-CACTGCAGAGAGTCCAACGGACTCG-5′ +3
3′-ACACTGCAGAGAGTCCAACGGACTC-5′ +4

Figure 7.18.5.

Probe design strategy for genotyping. Resequencing probes are placed with a minor/mutant allele at different offsets (9 offsets) to interrogate one genotype (A) and sequence 9 contiguous bases at the same time (GTCCTACGG).



Accession/Specimen:	XXXXXX
Patient ID:	XXXXXX
Batch Name:	B3604
Chip Name:	HCM CardioChip
Overall Call Rate:	98.60%
Results Generated Date:	04/29/2010 12:12PM

Gene	Exon	Wild Type Base	Call	Genotyping	Predicted Variant	Seq. Confirmation
Follow Up Positions						
ACTC	Exon 1	62C	NNN	-		Wild Type
LAMP2	Exon 1	1A	NNN	-		Wild Type
LAMP2	Exon 1	38G	NNN	-		Wild Type
MYBPC3	Exon 7	782G	GGG GKG	-	(Het) c.782G>T	Wild Type
MYBPC3	Exon 21	1928-2A	GNN	AG	(Het) c.1928-2A>G	(Het) c.1928-2A>G
MYH7	Exon 38	5560-5C	NMN	-		Wild Type
PRKAG2	Exon 3	341C	NNN	-		Wild Type
TPM1	Exon 5	47A	AWA	-	(Het) c.47A>T	Wild Type
Failed Sequence						
NONE						
Common SNPs						
MYH7	Exon 12	1095G	RNR RRR	GA	(Het) c.1095G>A	
MYH7	Exon 24	2967T	YYY YYY	TC	(Het) c.2967T>C	
TPM1	Exon 4	453C	MMM MMM	CA	(Het) c.453C>A	
Common False Positives						
ACTC	Exon 1	86C	MCN	-		
MYBPC3	Exon 25	2432A	AWA	AA		
MYH7	Exon 30	3984C	NNM	-		
PRKAG2	Exon 3	296C	NMM	-		
Common No Calls						
ACTC	Exon 2	311C	NNN	-		
ACTC	Exon 3	455-9C	NNN	-		
MYBPC3	Exon 3	315C	NNN	-		
TNNI	Exon 7	425C	NNN	-		

Figure 7.18.6.

Sample of a summary report generated from our data analysis pipeline. All positions requiring a confirmatory sequencing method are listed in the Follow Up Positions section. For each position, the gene, exon, and wild-type base are indicated. In the Call column, triplicate calls are shown, as well as minor allele tiling calls due to the presence of a nearby single-nucleotide polymorphism (SNP). When tiled, genotyping calls are noted next to the relevant base. The results of confirmatory Sanger sequencing are noted in the final column (Seq Confirmation). Non-wild-type base calls that are not followed up are listed in the subsequent sections including benign SNPs, as well as calls filtered out due to identification during test validation as a site of a common false positive or no-call.

Table 7.18.1

Critical Files Used in a Resequencing Array Experiment

File type	File description
GeneChip operating software (GCOS)	
Experiment file (`*.EXP`)	Information about the experiment name, sample, and probe array type.
Data file (`*.DAT`)	Contains an image of the scanned probe array.
Cell intensity file (`*.CEL`)	Derived from DAT file and contains signal intensity values for each feature in numeric and visual format.
GeneChip sequence analysis software (GSEQ)	
Analysis results file (`*.CHP`)	Generated from the signal intensity data from the CEL file. It lists final base calls, refseq base, corresponding quality scores, and heterozygosity for positions tiled on array.
Report file (`*.RPT`)	Includes probe array type, algorithm parameters, summary of final output, and fragment call rates.
Graphic file (`*.TIF`)	Graphic image of the array data.
FASTA exports	GSEQ automates the process of base calling using an algorithm with adjustable settings. It allows viewing of the sequence, exporting to FASTA files and generation of a SNP summary report.

Table 7.18.2

Suggested Settings for Resequencing Algorithms

Category	Setting
Filter Conditions Tab	
No signal threshold (probe signal/noise ratio)	1
Weak Signal Fold Threshold (mean/probe ratio)	20 (default)
Max signal-to-noise-ratio	20 (default)
Base Call Parameters Tab	
Genome model (0 = diploid, 1 = haploid)	0
Quality score threshold	1
Final Reliability Rules Tab	
Base reliability threshold across samples	0
Trace threshold	1 (default)
Sequence profile threshold	−0.175 (default)

Table 7.18.3

Specifications for Flexible Array Formats Available from Affymetrix[a]

Format	**49**	**100**	**169**
Sequence capacity	300 kb	100 kb	50 kb
Minimum lot per order	45 ± 5 arrays	90 ± 5 arrays	160 ± 5 arrays

[a] Derived from the Affymetrix Data Sheet: GeneChip CustomSeq Resequencing Array Program.