# GeneChip® CustomSeq® Custom Resequencing Array Design Guide

P/N 701263 Rev. 4

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

# Contents

Chapter 1

# Array Design Process Overview

Chapter 1

# Introduction

This section provides you with a step-by-step overview of the GeneChip® CustomSeq® Resequencing Array design process. The design process consists of five major steps:

1. Preliminary Design Review
2. Sequence Selection and Primer Design
3. Design Request
4. Array Design
5. Post Array Design

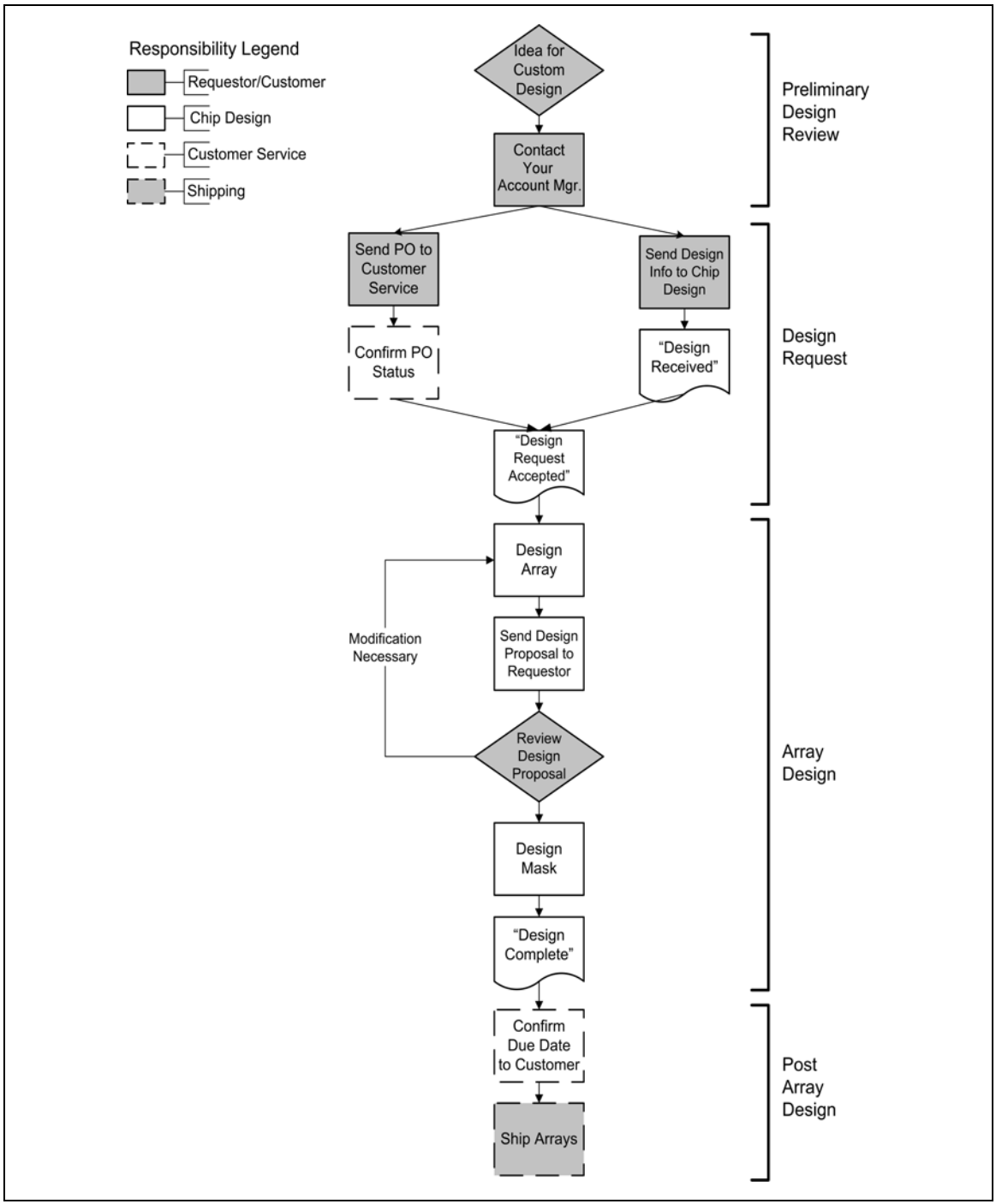
Responsibility Legend
Requestor/Customer
Chip Design
Customer Service
Shipping
Idea for Custom Design
Contact Your Account Mgr.
Send PO to Customer Service
Send Design Info to Chip Design
Confirm PO Status
"Design Received"
"Design Request Accepted"
Design Array
Modification Necessary
Send Design Proposal to Requestor
Review Design Proposal
Design Mask
"Design Complete"
Confirm Due Date to Customer
Ship Arrays
Preliminary Design Review
Design Request
Array Design
Post Array Design

**Figure 1.1**

## Preliminary Design Review

To begin the process for ordering a custom resequencing array, please contact your Affymetrix Account Manager. The Account Manager will review your request with both the Field Application Specialist (FAS) and the Custom Array Design Team.

If your design request is straight forward, your Account Manager and FAS will work with you to prepare the sequence submission documents and submit a Purchase Order. If your design idea requires modification to our design standard, our technical group is available to help. If necessary, a conference call can be arranged with the Affymetrix Custom Array Design Team.

This initial review process ensures that we provide you with an array design that meets your needs.

## Sequence and Primer Selection

The first step to designing a custom array is for you to identify the sequence(s) of interest. Once genomic region(s) are selected for analysis, the sequence is converted to the proper format and quality checked (see *Sequence Selection* on page 11 and *Primer Selection and Verification* on page 14). Next, PCR primers are designed to amplify the region(s) and amplicons are tested for adequate target amplification. It is important to ensure that you have adequate biological material to detect specific sequences before submitting a design for manufacture. Therefore, we ask customers to check all amplicons prior to sequence submission. Frequently, this step goes hand-in-hand with the preliminary design review. Any questions or concerns you have encountered during this process can also be reviewed with the Custom Array Design Team. Please refer to the *GeneChip® CustomSeq® Resequencing Array Protocol* (P/N 701231) for a full description of the protocol.

## Design Request

Next, sequences are submitted to Affymetrix by filling out the Design Request Form and sending the formatted sequence and instruction files to Chip Design (see Chapter 3 on page 19 for details). Once the design information is received, the Chip Design Group will send you a "Design Received" message[1].

To start the Affymetrix design process, a Purchase Order must be sent to our Customer Service Department. Affymetrix Customer Service will review the purchase order to verify that it is complete. Your Purchasing Agent will be contacted for any necessary clarifications. Once everything is complete, a confirmation message is sent to your Purchasing Agent.

When both the completed Purchase Order and the design information have been received, the Chip Design Group will send you a "Design Request Accepted" message. This message signifies that array design will begin on your GeneChip CustomSeq® Resequencing Array.

## Array Design

The Affymetrix Chip Design Group begins the array design process by assigning a specific chip designer to your design. This individual is your contact person throughout the design process. If the designer has questions, he/she will send you a "Design Clarification" message.

If everything is okay with the design, you will receive a "Design Proposal" message from Chip Design. This message summarizes the number of bases that are sequenced on the array. Please respond by indicating whether or not you approve the design. Once you accept the design proposal, we will move forward with the mask design.

Upon completing the mask design, we will send you a "Design Complete" message. This message signifies that the chip design is complete and your arrays will be manufactured. No changes can be

[1] *All e-mail messages described in this document will have the following information in the Subject line:* ***<Array description>-<Array name>: <e-mail description>****. The array description is provided by you in the Design Request Form. The array name is a unique code used to identify your design. For example, the subject line of a "Design Request Accepted" message will look something like "Rat Discovery-RATDISC1r510777: Design Request Accepted."*

made to the design after this time.

We send this series of communications to update you on the status of your array design as it moves through the design process.

## Post Array Design

After your arrays are manufactured and quality control tested, they are shipped to you along with Library Files containing information specific to your design. Technical questions should be directed to your FAS, while all other questions should be directed to your Account Manager.

Chapter 2

# Resequencing Design Standards and Considerations

Chapter 2

# Sequence Selection

The Affymetrix CustomSeq® program allows researchers to design custom resequencing arrays containing unique content. Sequence of interest may be downloaded from any number of public or proprietary databases and converted to FASTA format. Due to the variation in sources, customers are responsible for sequence integrity and if applicable, obtaining licenses to any genes or sequences on the array.

## ARRAY CAPACITY

The following table **approximates the number of bases** that fit on the array based on the feature size and array format (this is in addition to the standard Affymetrix controls we tile on the array).

**Table 2.1**
Array Capacity

| Array Format | Maximum Capacity |
|---|---|
| 49 | 303,366 |
| 100 | 117,254 |
| 169 | 47,974 |

Please use the following formula to calculate the number of bases you can sequence on a single array. This equation accounts for the first and last 12 flanking bases required for each discrete fragment tiled on the array (see *Start* on page 28). Please note that the total number of bases submitted in your sequence file should include the flanking bases.

$$\text{\# of bases sequenced} = \text{Maximum Capacity}^{*} - (\text{\# of non-contiguous fragments X } 24)$$

*(or total amount of sequence submitted if less than Maximum Capacity)*

## AMBIGUOUS SEQUENCES

Any ambiguous sequences (Ns, or other ambiguous IUPAC codes) submitted to Chip Design will be randomly replaced with A, G, C, or T. If possible, please refer to additional or alternative sources of sequence and manually edit the ambiguous bases prior to submission. Non-IUPAC codes in sequences will not be accepted.

## REPETITIVE ELEMENTS

Repetitive elements and internal duplications lead to cross-hybridization and must be removed prior to sequence submission. To identify repeats you may wish to use RepeatMasker shareware available at the following URL:
http://repeatmasker.org/chi-bin/webrepeatmasker

To identify other types of repetitive regions/ large duplications you may wish to use Miropeats, available at:
http://www.littlest.co.uk/software/bioinf/index.html

1. Run RepeatMasker or similar programs to identify the region(s) of repeats.
2. If the region of repeat is < 25 bases, there is no need to remove the sequence as it can safely be tiled on the array.
3. If a repeat is > 25 bases then the repeat region should be removed from the sequence to be tiled by using the Instruction File (see *Instruction File* on page 26). The example below illustrates how this can be achieved.

In the following example a repeat region, indicated by the n's, is found between bases 200 and 300. To tile without this repeat region, the following instruction file would be required:

| **Sequence** | | | |
|---|---|---|---|
| 1 | 200 | 300 | 400 |
| GAAC...TTATnnnnnnnnnn...nnnnnnnnCTGTACCCAATTTG...AGGTAACTCTCT | | | |

| **name** | **alias** | **start** | **end** | **startSeq** | **endSeq** |
|---|---|---|---|---|---|
| Myseq | myseq-1 | 1 | 212 | GAAC... | ...nnn |
| Myseq | myseq-2 | 288 | 400 | nnnnn... | ...CTCT |

NOTE **The end sequence includes the first 12 bases of the repeat and the start sequence includes the last 12 bases of the repeat.**

4. Set up the Instruction File such that the "end" of a fragment is 12 bases after the beginning of the repeat region. This would make the base before the repeat region the last based to be sequenced. For example, if the repeat region starts at base 101, the last base you would want to sequence is base 100. This means that the last probe should go from bases 88 to 112, and 112 should be your "end" value in the instruction file.
5. Set up the Instruction File such that the "start" of the next fragment is 12 bases before the end of the repeat region. This would make the first base after the repeat region to be the first base to be sequenced. For example, if the repeat region ends at base 200, the first probe should be from bases 189 (value for the "start" column) to 213, so that base 201 is the first base to be sequenced.
6. When the sequences are submitted remember to replace all the masked sequences, replaced by n's, in both the Sequence File and the Instruction File with the original sequence. Any remaining n's will be treated as ambiguous sequence and replaced as described in *Ambiguous Sequences* on page 12. Repeat regions should be excluded through the use of "start" and "end" regions defined in the Instruction File only. This allows Affymetrix to produce the most specific probes.

### HOMOLOGOUS SEQUENCES

Highly homologous sequences also lead to cross-hybridization and compromised data quality for that specific sequence. Affymetrix recommends running a homology check on both your amplified and tiled sequences prior to submission. Highly homologous regions should be tiled on separate arrays. Alternatively, you may also screen out homologous sequences by excluding them on the design (using "start" and "end" column similar to the way you would exclude repetitive elements).

## Primer Selection and Verification

Primer selection and validation prior to submission are essential to ensure the array contains only sequence for which adequate biological material can be obtained. Primers should be designed external to the sequence interrogated on the array, but they do not need to be adjacent to the first base sequenced. A homology search should be run between the sequence of your amplified fragments and the sequence tiled on the array in order to limit cross-hybridization between similar sequences. To facilitate primer design, validated primers from the Perlegen database are available upon request. Please contact your FAS for access.

# Array Controls

## STANDARD CONTROLS

All resequencing arrays will have standard manufacturing and hybridization controls tiled on the arrays.

## PROCESS CONTROLS

Resequencing arrays will also contain DNA analysis control sequences to test for amplification from a synthetic construct. Please refer to the *GeneChip® CustomSeq® Resequencing Array Protocol* (P/N 701231) for full descriptions of the TagIQex controls.

## USER DEFINED CONTROLS

Additional controls may be tiled on the arrays. Please make sure to include these sequences in your array capacity calculation and add them to your sequence submission file.

## Minimum Array Order per Array Format

The table below defines the minimum number of arrays per synthesis for the corresponding array formats. Custom products must be ordered in full manufacturing lots.

**Table 2.2**

| Format | Minimum # Arrays |
|---|---|
| 49 | 40 ± 5 |
| 100 | 90 ± 5 |
| 169 | 160 ± 5 |

## Custom Resequencing Design Standards

Below are the standard design options that are allowed within your custom resequencing array design.

1. Target strandedness: forward and reverse
2. Feature size: 8 micron
3. Probe length: 25-mer
4. Design: Single array or multi-array set
5. Array format: 49, 100, 169

All designs will go through our internal Quality Control testing.

## Custom Design Library Files

For each custom resequencing design you will receive a set of Library Files for your custom array. The Library Files are required for analysis in Affymetrix software. An installation program is also provided for each design, which loads the Library Files to your system.

The Library Files contain a list of all the probes tiled on your array and the location of each probe. In addition, the Library Files also include the standard analysis parameters we recommend for scanning and anlysis.

Chapter 3

# Submitting a Design Request

Chapter 3

# Submitting a Design Request

The Chip Design Group will begin your design after the Design Request Form, sequence and instruction files, and Purchase Order (PO) have been received.

## Purchase Order

A PO is a commitment to buy our products and services. A PO for a custom resequencing array consists of the following line items:

1. design fee[1]
2. order for the first lot(s) of arrays

The PO should be faxed to the Affymetrix Customer Service Group.

## Design Request Form

The Design Request Form provides Affymetrix with contact information and design parameters for your array design. The Design Request Form can be obtained online at: www.affymetrix.com/products/arrays/specific/custom_seq.affx

In the Design Request Form, please provide the following information:

### REQUESTOR INFORMATION

The Requestor Information fields provide the necessary contact information to notify the design requestor of the status of his/her design. If necessary, we will also contact the requestor for questions/clarifications about the design. The last two fields regarding Affymetrix contacts are optional; however, this information is helpful as it allows us to notify your Account Manager of your design status as well.

[1] *Please contact your account manager if you have any questions.*

# RESEQUENCING PROBE ARRAY INFORMATION

## Array Name

The array name may contain up to eight alphanumeric characters, hyphen, and underscore.

The final array name for a custom design is created in this fashion: <customized name>r<part number>

The customized name is provided by you in the "Array Name" field. "r" stands for resequencing. The part number is the number we assign to your arrays.

For example, if you have a custom design with "Rat1" as the Array Name, and we assign part number 510777 to your design, then the final array name will be: Rat1r510777.

If you have more than one design in your design request, each design will get its own part number.

## Array Description

Please provide a description for your design. It may contain up to 24 characters including spaces.

## Feature Size

Please select the standard option of 8 micron.

## Array Format

Please select the array format of your choice.

## Sequence File Name

Multiple Sequence Files may be specified by providing one Sequence File name per line.

## # of Sequences Submitted

The number of sequences submitted must equal the total number of

entries in all the design sequence files. This allows us to verify that the design data have been successfully transferred.

### Instruction File Name

Multiple Instruction Files may be specified by providing one Instruction File name per line.

### Space Optimization

Space Optimization authorizes the Chip Design Team to rearrange the order of fragments on an array to optimize space and enable the highest capacity.

### Repeats File Name

In addition to any standard libraries that might already exist for a given species, you may provide your own file containing repetitive sequences that you would like to use for checking sequence homology. Multiple files may be specified by providing one sequence file name per line.

### Cross-hybridization Threshold (%)

The cross-hybridization threshold should be an integer between 1 and 100. This threshold is used to set the threshold for cross-hybridization allowed within a single array design. If the percentage of probes that cross-hybridize with other sequence(s) over the total number of probes is greater than or equal to the cross-hybridization threshold, we will generate a recommended instruction file to tile the cross-hybridizing sequences on two different designs.

### Species

Please provide us with the species for your design. More than one species may be selected if it is a multi-species design.

If "Other" is selected in the Species field, please provide the name(s) of the species, separated by commas. You may provide your own repeat library as well.

## Sequence File

The Sequence File(s) should contain all the sequences from which probes will be selected. **Please remember to add 12 bases at the start and end of each fragment.**

The Sequence File must be in the FASTA format, where each raw sequence is preceded by a definition line. The definition line begins with a sign and is followed immediately with a name for the sequence. The Sequence File has the following additional characteristics:

- A ">" precedes the sequence name.
- The sequence name must be unique.
- The sequence name cannot be all numerical digits.
- The sequence name must correspond to the value in the "name" column in the Instruction File (for more information see *Instruction File* on page 26).
- All names defined in the Instruction File must exist in the Sequence File, and vice versa.
- The sequence name in the Sequence files may be up to 20 characters that are alphanumeric or special characters: +, @, $, %, ^, &, (, ), =, #, ~.
- The sequence name in the Sequence files may not end in an underscore followed by a single alphanumeric character. Example: xxxx_3 and zzzz_s sequence names are not allowed.
- The sequence name in the Sequence files may not end in "_at" or "_st" as these are Affymetrix identifiers.
- A comment may follow the sequence name; however, the information here will not be utilized by the Chip Design Group.
- No blank lines should exist between sequences.

### Example 1: Sequence File

>reference_seq

taaaccctgaaccctaaaccctaaaccctgaaccctgaaccctaaaccctgttccctttccctgttcc**cA**ggat
ctacttcttacttactactaccttacttacctcttatttc**aC**ttacatctaggtccttatcctactacccaaaccctaat
cctttaccctaaaccctaaaccctgaaccctaaaccctgaaccctaaaccctaaaccctaaaaccctaaaccct
aaaccctataccctgaaccctaaaccctaaaccctttttccctaaaccctgaaccctaaaccctaaaccctaaacc
taaaccctgttccctaaaggctaaaccctgaaccctgaaccctttacttttcttactcttatgttcttactattcctatctt
agtcttactttcttcttcttaggtccttacttttcatttgtttttatcatgttcaaggtgttacttcttacagactacgtcttt
cagctcatactactagtcttcagttcttacttctagaaccttctacttactttcttcatcttacatcctcctattcaagtct
ttcagcttatcttcttactcttacatacctttctcttcctacttccagcatcactactcactcatcattccttggtcttactt
ctttcttcttcaggtcatacttttatcccctattatcatcttcttagtcttcaatttctcttctaagttcttcatctatcacttt
tcatcatcattcatacttacttaacttactttcatctacttacttctactacttcttcaccaaatccgcacttagcttcat
atttctttgtttgatcttactttacttagctcttacatacttattacttctcttatcttacttctttattctcaaaatccggact
tagcttaacttttctctctttattcttattcttaatcatatattcttactc**aT**atagttcttgacttaacttttttattcttactt
acttagtgttata

>Snp1

accctgttccctttccctgttcc**cG**ggatctacttcttacttactacta

>Snp2

ctaccttacttacctcttatttc**aT**ttacatctaggtccttatcctact

>Snp3

ttcttaatcatatattcttactc**aC**atagttcttgacttaacttttttat

NOTE

**The bottom three fragments contain redundant tiling for three known polymorphic sites.**

## Instruction File

The Instruction File provides a tabular summary of the start and end position for each contiguous fragment tiled on the array. This information is crucial for accurate array design, as well as being a component of the quality control process. This file must be in a tab-delimited text file format, and can easily be created from Excel by choosing to save the file in the "Text (tab delimited) (*.txt)" format. There should be at least one entry for each sequence in the Sequence File.

**Example 1: Instruction File containing multiple discontinuous fragments from same gene.**

| name | alias | start | end | startSeq | endSeq | design |
|---|---|---|---|---|---|---|
| AK097958p53FLa | P53exon1 | 1 | 70 | ACGTATGA | AGCATGTA | 1 |
| AK097958p53FLb | P53exon3-4 | 298 | 821 | AGTCGTAT | ATCGTAGT | 1 |
| AK097958p53FLb | P53exon5 | 809 | 1924 | ACGTAGTC | GCTGCTGA | 1 |

NOTE

**If sequence or instruction files are modified at any time during the design process, please resubmit the modified sequence and/or instruction files to us. This ensures that we use the correct information and minimizes errors due to manual editing or miscommunication.**

**Example 2: Instruction File containing redundant tiling for multiple variants**

| name | start | end | startSeq | endSeq |
|---|---|---|---|---|
| reference_seq | | 1 | 980 TAAA | TATA |
| Snp1 | 1 | 49 | ACCC ACTA | |
| Snp2 | 1 | 49 | CTAC TACT | |
| Snp3 | 1 | 49 | TTCT TTAT | |

## DESCRIPTION OF COLUMNS

### Name

The "name" is the **unique** designation for each **gene or sequence** represented on the array. Please provide a "name" for each sequence described in the Design Sequence File. Any combination of up to 20 alphanumeric characters may be used provided it does not end with an underscore followed by a single alphanumeric character. Please refer to *Sequence File* on page 24 for sequence naming rules.

The Accession Number in GenBank® or the unique sequence ID from a public domain or proprietary database is most commonly used. Utilization of standard accession numbers will facilitate linkage to annotations during data analysis.

### Alias

The "alias" is the unique designation used for the fragment name. When tiling multiple non-contiguous fragments for the same gene or sequence, you may use an "alias" to differentiate between them. The software will report the alias, if no alias is provided, then the contents of the "name" field will also be used as the alias name. It may contain up to 20 characters and follows the same rules as those for sequence names (see *Sequence File* on page 24).

To exclude certain regions within a fragment, such as highly repetitive elements and/or homologous sequences, you can produce multiple entries with the same "alias" value and different "start" and "end"

regions.

| name | alias | start | end |
|---|---|---|---|
| AK097958p53FL | p53exon1 | 1 | 70 |
| AK097958p53FL | p53exon2-5 | 298 | 821 |
| AK097958p53FL | p53exon2-5 | 900 | 1924 |

## Start

The Start designates the first base in the fragment. This refers to the base at the beginning of the probe, and not the first position of interrogation. The first base sequenced is position 13 in the sequence file.

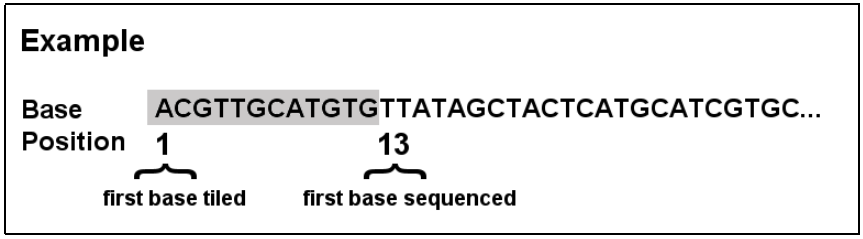


**Figure 3.1**

## End

The "end" designates the last base in the sequence fragment. This refers to the absolute position of the end of the probe, which is 12 bases after the last base of interest. For example, if your last base of interest is at base 500, then you should specify the "end" value to be base 512.


**Figure 3.2**

If multiple probe selection regions from the same sequence are specified, the "end" (last possible probe) from the first region and the "start" (first possible probe) from the next region should not overlap by more than 24 bases.

**Example:**

| name | alias | start | end |
|---|---|---|---|
| AK097958p53FL | p53exon1 | 1 | 70 |
| AK097958p53FL | p53exon2-5 | 298 | 821 |
| AK097958p53FL | p53exon2-5 | 900 | 1924 |

## StartSeq

The "StartSeq" is used for quality control of the sequence file. It includes the first base at the "start," and the next few bases up to a maximum of eight. This information allows us to cross-check your sequences and the fidelity of the sequence file.

### EndSeq

The "EndSeq" is used also for quality control of the sequence file. It includes the last base at the "end," and the previous few bases up to a maximum of eight. This information allows us to cross-check your sequences and the fidelity of the sequence file.

### Design

A design number should be designated when submitting sequence for multiple designs. Separate designs are recommended for sequencing highly homologous regions. If the "design" column is not provided, then the assumption is that all the sequences should be tiled on the same array.

## EXAMPLES[1]

1. An example of an instruction file for a simple design is shown below. The first base of interrogation starts at base 13 for sequence AK097958-a, and ends at base 988. However, there are no interrogations between bases 501 and 612. For sequence AK097958-b, there are actually two distinct fragments. p53exon2 starts interrogation at base 512 and ends at base 800. p53exon5 starts interrogation at base 1512 and ends at base 2988.

| name | alias | start | end | startSeq | endSeq |
|---|---|---|---|---|---|
| AK097958-a | p53exon1 | 1 | 512 | ACCG | CCGT |
| AK097958-a | p53exon1 | 601 | 1000 | CTTA | TGAT |
| AK097958-b | p53exon2 | 500 | 812 | GGGA | TAAA |
| AK097958-b | p53exon5 | 1500 | 3000 | TACC | GCTA |

[1] *Examples are provided for illustration purposes only. Only the relevant columns are shown and they may not line up as nicely in a real instruction file.*

2. If sequences A and B are very similar to each other, you might want to specify tiling to different designs by including the design number. If you do not provide this number, we will send you a recommended revised instruction file should our checking program find two sequences sharing high similarity.

| name | start | end | startSeq | endSeq | design |
|---|---|---|---|---|---|
| A | 1 | 70 | ACCG | CCGT | 1 |
| B | 100 | 200 | GGGA | TAAA | 2 |

## Annotation Files

GeneChip® Sequence Analysis Software (GSEQ) software allows you to display the genomic position and the PCR start/stop positions in the Table and Sequence Views of the Resequencing window. We recommend that you create your genomic position files at the time of array design. You must provide the genomic position and PCR start/stop position data in a space-delimited text file using the following formats:

### Genomic Position File

The genomic position file (Figure 3.3) for each array type must have a unique name, and must contain the following information for each fragment:

- Fragment Name
- Chromosome in which the fragment resides
- Chromosome position for the first base in the fragment

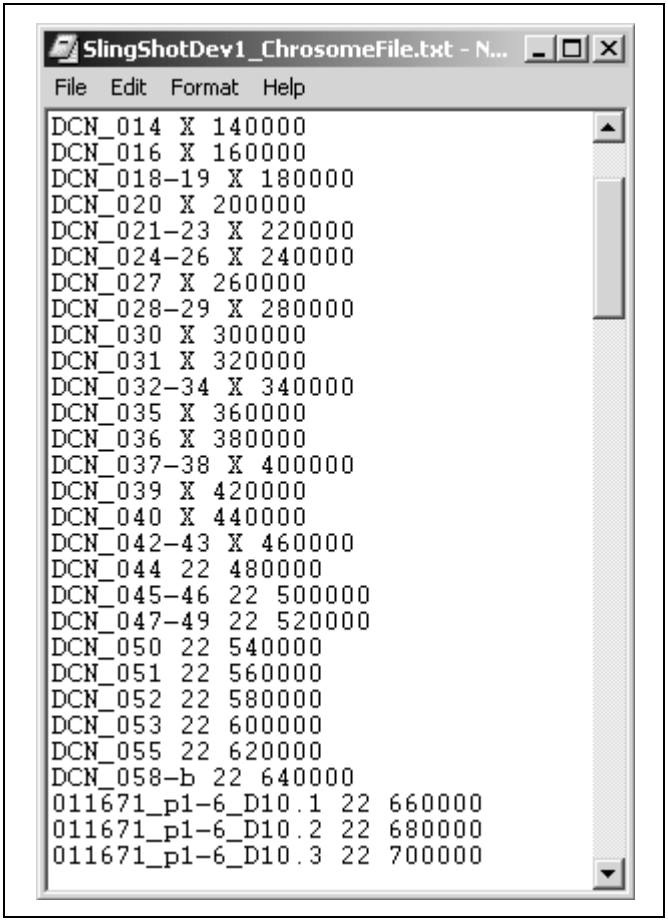


**Figure 3.3**
Genomic Position File

### PCR Start/Stop Position File

The PCR start/stop position file (Figure 3.4) for each array type must have a unique name, and must contain the following information for each fragment:

- Tiled Fragment name for the PCR sequence
- Chromosome position for the start position of the PCR sequence
- Chromosome position for the stop position of the PCR sequence

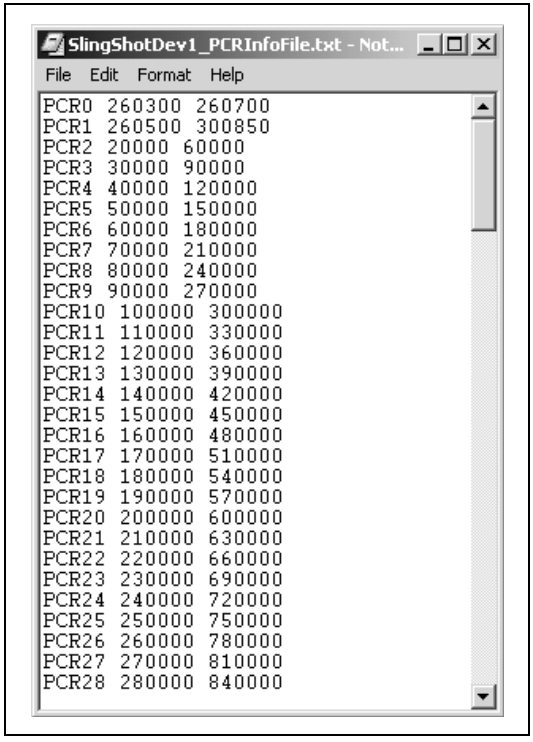


**Figure 3.4**
Genomic Position File

NOTE **You must provide genomic position data to display PCR start/stop data.**

Chapter 4

# Resequencing Design Checklist

Chapter 4

# Resequencing Design Checklist

Before you submit your design, please review the following checklist to make sure you have considered all aspects of a design.

**Table 4.1**
Resequencing Design Checklist

| No. | Description | Status | | | |
|---|---|---|---|---|---|
| 1. | Selected sequence in FASTA format | ☐ | Done | ☐ | N/A |
| 2. | Repetitive elements removed through the Instruction File | ☐ | Done | ☐ | N/A |
| 3. | Ambiguous bases edited | ☐ | Done | ☐ | N/A |
| 4. | Sequence homology checked | ☐ | Done | ☐ | N/A |
| 5. | Number of bases will fit on desired format | ☐ | Done | ☐ | N/A |
| 6. | Flanking regions inserted (12 bases up and down stream of first base interrogated) | ☐ | Done | ☐ | N/A |
| 7. | Primers designed | ☐ | Done | ☐ | N/A |
| 8. | Primers tested | ☐ | Done | ☐ | N/A |
| 9. | Design request form filled out | ☐ | Done | ☐ | N/A |
| 10. | Sequence and Instruction Files named appropriately | ☐ | Done | ☐ | N/A |
| 11. | Design submitted to Chip Design | ☐ | Done | ☐ | N/A |
| 12. | PO submitted to Customer Service | ☐ | Done | ☐ | N/A |