



GENECHIP[®] CUSTOMEXPRESS[™] ARRAY DESIGN GUIDE

GeneChip® CustomExpress™ Array Design Guide

Welcome to the GeneChip® CustomExpress™ Array Design Guide. The CustomExpress Array Program enables you to design GeneChip arrays tailored to your specific needs. GeneChip CustomExpress Arrays are produced using the same top-quality design and manufacturing technologies as Affymetrix® GeneChip catalog arrays to ensure consistent results across your gene expression studies. The Design Guide has been developed to assist you in the design process, from planning your array design to creating the design files Affymetrix will use to select probes and create your array.

This design guide is divided into six sections:

- Section 1: provides a brief overview of the entire design process, from sending the initial set of data to completing the array design.
- Section 2: discusses the standards for a CustomExpress design, as well as some important concepts you should consider when putting together a design.
- Section 3: provides more details regarding how to submit a design and the detailed formatting of the design files.
- Section 4: explains the Design Proposal you will receive once probe selection is complete, and the necessary action(s) required from you in order to complete the design.
- Section 5: lists Affymetrix contact information that can assist you in custom design process.
- Section 6: is a glossary to help you quickly locate key words that are used throughout this document.

Table of Contents

1.	CUSTOMEXPRESS™ DESIGN PROCESS OVERVIEW	1
1.1	PRELIMINARY DESIGN REVIEW	2
1.2	DESIGN REQUEST	2
1.3	ARRAY DESIGN	2
1.4	POST ARRAY DESIGN	3
2.	EXPRESSION DESIGN STANDARDS AND CONSIDERATIONS.....	5
2.1	GENECHIP® CUSTOMEXPRESS™ ARRAYS.....	5
2.1.1	<i>GeneChip CustomExpress Premier Arrays</i>	5
2.1.2	<i>GeneChip CustomExpress Advantage Arrays</i>	5
2.2	CUSTOMEXPRESS DESIGN STANDARDS.....	6
2.3	MINIMUM NUMBER OF ARRAYS PER FORMAT.....	6
2.4	NUMBER OF PROBE SETS ON AN ARRAY PER FORMAT	6
2.5	SEQUENCE SELECTION FOR EUKARYOTIC EXPRESSION ARRAYS.....	8
2.5.1	<i>Content Assembly Approaches</i>	8
2.5.1.1	Genome Annotation-Based Design	8
2.5.1.2	Exemplar Approach	8
2.5.1.3	Consensus Approach.....	8
2.5.2	<i>Sequence Orientation</i>	8
2.5.2.1	Sequence Annotations.....	9
2.5.2.2	Consensus Splice Sites	9
2.5.2.3	Protein Homology Hits.....	9
2.5.2.4	Polyadenylation Sites & Signals	9
2.5.3	<i>Polyadenylation</i>	9
2.5.4	<i>Sequence Quality</i>	10
2.5.5	<i>Sequence Prioritization</i>	11
2.6	PROBE SELECTION	11
2.6.1	<i>Probe Selection Process</i>	11
2.6.2	<i>Probe Set Types</i>	12
2.6.2.1	Gene Family Probe Sets (_a).....	12
2.6.2.2	Unique Probe Sets.....	12
2.6.2.3	Identical Probe Sets (_s).....	12
2.6.2.4	Mixed Probe Sets (_x).....	12
2.6.2.5	Examples of Probe Set Types.....	14
2.7	PRUNING FOR PROBE SELECTION	15
2.7.1	<i>Hard Pruning</i>	15
2.7.2	<i>Soft Pruning</i>	15
2.8	ARRAY CONTROLS.....	15
2.8.1	<i>Standard Controls</i>	15
2.8.2	<i>Species-Specific Controls</i>	16
2.8.3	<i>Other Controls</i>	17
2.8.3.1	Contaminant Controls	17
2.8.3.2	Scaling and Normalization Controls	17
2.9	CUSTOM DESIGN OUTPUT FILES	18
3.	SUBMITTING A DESIGN REQUEST	19
3.1	PURCHASE ORDER.....	19
3.2	DESIGN INFORMATION	19
3.2.1	<i>Design Request Form</i>	19
3.2.1.1	Requestor Information.....	19
3.2.1.2	Expression Probe Array Information.....	19
3.2.1.2.1	Array Name	19
3.2.1.2.2	Array Description	20
3.2.1.2.3	Feature Size	20

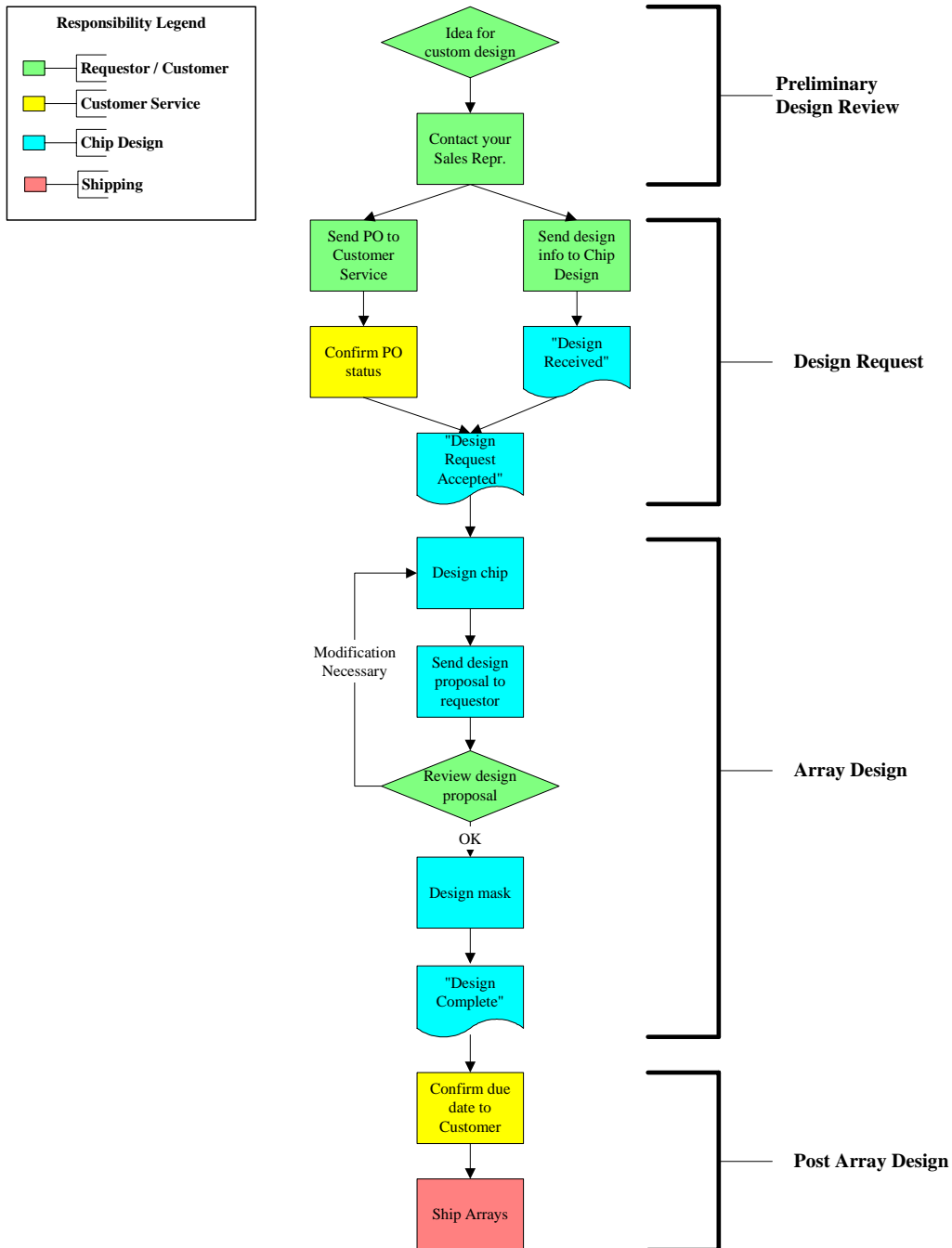
3.2.1.2.4	Array Format.....	20
3.2.1.3	Design Probe Sets	20
3.2.1.3.1	Sequence File Name	20
3.2.1.3.2	Instruction File Name	20
3.2.1.3.3	Probe Set Request File Name.....	20
3.2.1.3.4	Target Type.....	20
3.2.1.3.5	Probe Selection Region	20
3.2.1.3.6	Number of Probe Pairs per Sequence.....	21
3.2.1.4	Controls.....	21
3.2.1.4.1	Species on Array.....	21
3.2.1.4.2	Controls (for Pruning & Tiling)	21
3.2.1.5	Pruning.....	22
3.2.1.5.1	When Selecting Probes	22
3.2.1.5.2	FASTA Format Hard Pruning Sequence File Name	22
3.2.1.5.3	FASTA Format Soft Pruning Sequence File Name.....	22
3.2.1.5.4	Use Affymetrix Standard Pruning Library.....	22
3.2.1.6	Special Instructions	22
3.2.2	<i>Probe Set Request File</i>	23
3.2.3	<i>Sequence File</i>	23
3.2.4	<i>Instruction File</i>	24
3.2.4.1	Description of Columns	24
3.2.4.1.1	name.....	24
3.2.4.1.2	database.....	25
3.2.4.1.3	alias	25
3.2.4.1.4	geneCluster	26
3.2.4.1.5	transcriptExpressed	26
3.2.4.1.6	start	26
3.2.4.1.7	end	26
3.2.4.1.8	startSeq	27
3.2.4.1.9	endSeq	27
3.2.4.1.10	hyperlink.....	27
3.2.4.1.11	hyperlinkDatabase.....	27
3.2.4.1.12	description.....	28
3.2.4.1.13	probes.....	28
3.2.4.1.14	minProbes	28
3.2.4.1.15	copies.....	28
3.2.4.1.16	classification columns.....	28
3.2.4.2	Examples.....	29
3.2.5	<i>Pruning Sequence File</i>	31
4.	UNDERSTANDING THE DESIGN PROPOSAL.....	33
4.1	PROBE SET REDUNDANCY REDUCTION IN DESIGN PROPOSAL	33
4.1.1	<i>Extended Probe Selection Region</i>	33
4.1.2	<i>Probe Set Score</i>	33
4.1.3	<i>Applied Probe Set Score</i>	34
4.1.4	<i>Redundancy Reduction Process</i>	35
4.1.5	<i>Standard Redundancy Reduction Criteria</i>	35
4.2	DESIGN PROPOSAL.....	36
4.2.1	<i>Name</i>	36
4.2.2	<i>Alias</i>	36
4.2.3	<i>ProbeSetCount</i>	36
4.2.4	<i>ProbeSetName</i>	36
4.2.5	<i>Type</i>	36
4.2.6	<i>ProbeSetScore</i>	37
4.2.7	<i>GapMultiplier</i>	37
4.2.8	<i>CrossHybridizationMultiplier</i>	37
4.2.9	<i>ProbeCount</i>	37
4.2.10	<i>AvgRawProbeScore</i>	37
4.2.11	<i>RawStandardDeviation</i>	37

4.2.12	<i>RepresentedBy</i>	38
4.2.13	<i>RepresentedByAppliedScore</i>	38
4.2.14	<i>NumberCrossHybProbes</i>	38
4.2.15	<i>SifLength</i>	38
4.2.16	<i>NumberNonoverlappingProbes</i>	38
4.2.17	<i>NumberIndependentProbes</i>	38
4.2.18	<i>TilingOrder</i>	38
4.2.19	<i>Remove</i>	38
4.2.20	<i>ChangeRemoveValue</i>	39
4.3	OTHER FILES.....	39
4.3.1	<i>Proposal Represented File (REP File)</i>	39
	Column Heading.....	39
	Definition.....	39
4.3.2	<i>Cross Hybridization File (XHY File)</i>	40
	Column Heading.....	40
	Definition.....	40
4.4	ACTIONS REQUIRED.....	41
5.	CONTACT INFORMATION	43
5.1	SALES & FIELD APPLICATION SUPPORT.....	43
5.2	CUSTOMER SERVICE.....	44
5.3	CHIP DESIGN.....	44
6.	GLOSSARY	45

1. CUSTOMEXPRESS™ DESIGN PROCESS OVERVIEW

This section provides you with a step-by-step overview of the GeneChip® CustomExpress™ Array design process. The design process consists of four major steps:

- Preliminary Design Review
- Design Request
- Array Design
- Post Array Design



1.1 Preliminary Design Review

The preliminary design review begins with your idea for a custom design. When ready, please contact your Affymetrix Account Manager. The Account Manager reviews your request with both the Field Application Specialist (FAS) and the Custom Array Design Team.

If your design request is straightforward, your Account Manager will then work with you to prepare a Purchase Order. If your design idea requires modification to our design standards (see Section 2.2), our technical group is available to help. If necessary, a conference call can be arranged with the Custom Array Design Team.

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

1.2 Design Request

Your next step in the GeneChip CustomExpress array design process is to send a Purchase Order to our Customer Service Department, and to send the associated design information to the Chip Design Group (see Section 3 for details).

Affymetrix Customer Service reviews 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 the design information is received by the Chip Design Group you will be sent a “Design Received” message¹, regardless of the Purchase Order status. When both the completed Purchase Order and the design information have been received, the Chip Design Group sends you a “Design Request Accepted” message. This message signifies that array design will begin on your GeneChip CustomExpress array.

1.3 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 you need to send us additional sequences or new instructions after our discussion(s), please send the new, **complete** Sequence and/or Instruction Files to the Chip Design Group. This ensures that we are always using the correct information.

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-RATDISC1a510777: Design Request Accepted”.

Designs may contain “commercial content,” which is content selected from Affymetrix catalog and/or Made-to-Order Arrays, or unique content, whereby you submit sequences and Affymetrix designs probes for you. It is also possible to create a design that contains both commercial content and unique content.

If probe selection is necessary for your custom design, we will send you a “Design Proposal” describing the results of probe selection. If your design contains only commercial content, we will send you a Design Proposal to confirm your list of probe sets. Please review the proposal, and let us know how to proceed with your design (see Section 4 for details). Once you accept the design proposal, we then design the masks.

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, quality control tested, and shipped to you. No changes to your design can be made at this time.

We send these series of communications so you are aware of the status of your array as it moves through the design process.

1.4 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.

2. EXPRESSION DESIGN STANDARDS AND CONSIDERATIONS

2.1 GeneChip® CustomExpress™ Arrays

2.1.1 GeneChip® CustomExpress™ Premier Arrays

GeneChip® CustomExpress™ Premier Arrays use the same probe set design and manufacturing technology (including the same number of masks and synthesis steps) as the catalog arrays and are offered in the same formats as our catalog products. GeneChip CustomExpress Premier Arrays are typically utilized for:

- high-content designs, such as whole genomes of organisms not represented on our catalog arrays
- combination designs, such as multiple bacterial organisms
- lower content arrays when thousands of arrays are planned utilizing a single design

Content for your custom array can come from your proprietary sequences, from your previous custom designs, and/or from Affymetrix catalog and Made-to-Order products.

Your GeneChip CustomExpress Premier Arrays will be shipped to you approximately six weeks from the time you receive the “Design Complete” message from Affymetrix.

2.1.2 GeneChip® CustomExpress™ Advantage Arrays

The CustomExpress Advantage Array Program enables you to quickly create custom, low- to medium- content GeneChip arrays with a rapid turnaround time at an affordable price. Whether following up on candidate genes from a screening study or beginning a focused experiment, CustomExpress Advantage arrays are a powerful component of any expression research program. Two different design types are available for CustomExpress arrays: Commercial Content and Unique Content designs.

Commercial Content designs enable you to select content from any of the GeneChip catalog or Made-to-Order arrays. Unique Content designs are created using your unique or proprietary content. These arrays follow the same design and ordering process used for GeneChip CustomExpress Premier Arrays.

GeneChip CustomExpress Advantage Arrays use the same probe set design and manufacturing technology as catalog and GeneChip CustomExpress Premier Arrays, providing a single high-quality system from start to finish for any gene expression study.

Your GeneChip CustomExpress Advantage Arrays will be shipped to you approximately four weeks from the time you receive the “Design Complete” message from Affymetrix.

2.2 CustomExpress Design Standards

Below are the seven default design parameters for your custom array. Deviations from these criteria should be reviewed with your FAS and the Custom Array Design Team.

- Sequence Information: eukaryotic or prokaryotic organisms; multi-species designs are permitted if the different species are pruned together; up to two different species if the different species are pruned independently
- Target Strandedness: sense or antisense (see Glossary)
- Feature Size: 18, 20, 24, or 50 micron
- Number of Probe Pairs per Sequence: 11 to 60 for eukaryotic designs; 13 to 60 for arrays using a cDNA assay (many prokaryotic designs)
- Probe Length: 25-mer
- Probe Set Format: distributed probe sets
- Design: single array or multi-array set²

All designs are Quality Control tested using the same test protocols as catalog arrays.

2.3 Minimum Number of Arrays per Format

The array format selected for your custom design dictates the minimum number of arrays that must be purchased with **every order** placed for your CustomExpress Arrays. The chart below defines the minimum number of arrays for each of the corresponding array formats.

CustomExpress Premier Arrays

Format	Synthesis Area	Minimum Order
49	12.8 mm	80 ± 10
64	10.8 mm	110 ± 10
100	8.1 mm	90 ± 5
169	5.3 mm	155 ± 10
400	2.5 mm	385 ± 10

CustomExpress Advantage Arrays

Format	Synthesis Area	Minimum Order
49-7875	7.9 mm	80 ± 10
49-5241	5.2 mm	80 ± 10
100-3660	3.7 mm	90 ± 5
100-2187	2.2 mm	90 ± 5

2.4 Number of Probe Sets on an Array per Format

The following matrix **approximates the number** of probes and probe sets that can fit on an array based on the feature size and array format (this is in addition to the Affymetrix controls we tile on the array). The “# probe sets” calculation is based on 11 probe pairs used for our standard catalog arrays:

² A multi-array set design means that you will submit enough sequences to be tiled onto two or more arrays. However, you will only be submitting one set of design information, and the probe selection will be done as for a single design. The arrays will be manufactured as **separate** designs. Each array will be charged as a separate design.

CustomExpress™ Premier Arrays

Feature Size	Array Format	# Probes	# Probe Sets
18 micron	49	501,170	22,780
	64	357,070	16,230
	100	195,986	8,908
	169	81,250	3,693
	400	16,194	736
20 micron	49	404,114	18,368
	64	288,670	13,121
	100	157,468	7,157
	169	64,750	2,943
	400	12,526	569
24 micron	49	280,094	12,731
	64	199,570	9,071
	100	108,626	4,937
	169	43,806	1,991
	400	7,966	362
50 micron	49	61,586	2,799
	64	42,866	1,948
	100	22,190	1,008
	169	7,654	347

CustomExpress™ Advantage Arrays

Feature Size	Array Format	# Probes	# Probe Sets
18 micron	49-7875	186,376	8,471
	49-5241	80,818	3,673
	100-3660	38,392	1,745
	100-2187	11,628	528
20 micron	49-7875	150,076	6,821
	49-5241	65,188	2,963
	100-3660	30,166	1,371
	100-2187	8,928	405
24 micron	49-7875	103,864	4,721
	49-5241	44,244	2,011
	100-3660	20,088	913
	100-2187	5,418	246
50 micron	49-7875	10,728	975
	49-5241	7,992	363

If you would like to use more than 11 probe pairs, you can calculate the number of sequences that will fit on a given array format using the following formula:
(# of probes for a particular feature size per array format) / (2 × (# of probe pairs))

We recommend that you consider a prioritization strategy for your sequences in the event that there are more probe sets generated than can be tiled in the format specified.

2.5 Sequence Selection for Eukaryotic Expression Arrays

While there are several different approaches to assembling content for an array design, any approach should consider sequence orientation and quality, the 3' bias of the standard labeling assay, and biological particulars, such as trans-splicing, alternative splicing, and alternative polyadenylation. The volume and quality of sequence information, as well as the specific experimental needs, will often dictate which design approach to use and the level of rigor that should be applied to sequence selection.

2.5.1 Content Assembly Approaches

There are three general approaches to assembling a design. Any one method, or combination of methods, used will depend on the available sequence data and the bioinformatics resources of the group doing the design.

2.5.1.1 Genome Annotation-Based Design

For organisms with finished genomes or sufficient draft coverage, it may be appropriate to select sequences based on gene annotations or alignments (and clustering of alignments) of cDNAs to the genome. In such a case, the implied mRNA transcript from the genome annotations or alignments would be submitted for probe selection.

2.5.1.2 Exemplar Approach

The exemplar approach selects a representative cDNA sequence for each gene. This method usually involves some initial clustering into gene groups and the subsequent selection of a representative from each gene group.

2.5.1.3 Consensus Approach

The consensus approach involves clustering cDNA sequences, generating cDNA assemblies, and calling a consensus sequence for each assembly.

2.5.2 Sequence Orientation

Accurate determination of sequence orientation is critical. When the orientation of a sequence cannot be determined, it should either be excluded or have probe sets selected for both strands. A number of metrics can be used to determine sequence orientation. We recommend using more than one. Possible metrics for orientation include:

2.5.2.1 Sequence Annotations

Sequence annotations, such as CDS annotations on mRNAs or EST read directions, are extremely valuable; however, caution should be applied when using these metrics. For instance, lane tracking problems can result in incorrect EST read directions (and incorrect clone annotations). CDS annotations can also be incorrect, as they may simply be the longest open reading frame for the mRNA sequences rather than the correct open reading frame. There is no rule for EST orientation. By convention, 3' ESTs are antisense and 5' ESTs are sense oriented. Application of this convention in assessing EST orientation should be examined for systematic reverse complementing of 3' EST reads.

2.5.2.2 Consensus Splice Sites

For consensus splice sites revealed by cDNA alignments to DNA or DNA annotations, the flanking intronic sequence at a splice site junction can be used to infer the orientation of a cDNA sequence aligned to the genome or a DNA transcript annotation. For instance, a cDNA sequence aligned to genomic sequence may have “gt...ag” pairs for all the introns, where the “gt” is at the 5' end of the intron and “ag” at the 3' end. This would infer that the cDNA sequence is in the sense direction. Conversely, “ct...ac” pairs would imply an antisense oriented transcript. Orientation calls based on consensus splice sites tend to be of high confidence; however, its usefulness is limited by the availability of genomic sequence and the requirement that the cDNA or gene annotation be spliced.

2.5.2.3 Protein Homology Hits

Strong alignments to protein sequence, even protein sequence from another organism, can provide strong evidence for orientation. One limit to this approach is that in organisms with moderate to long 3' UTRs, the 3' EST reads often do not extend into the coding region.

2.5.2.4 Polyadenylation Sites & Signals

Identification of polyadenylation sites and signals can provide strong evidence for orientation. False priming may lead to incorrect orientation calls; however, looking for the polyadenylation signal in addition to the polyadenylation site can mitigate some of this risk. Annotated polyadenylation sites and signals, as well as computationally detected sites and signals, can be used. Often poly-A tails are trimmed, therefore access to untrimmed reads can improve detection rates.

2.5.3 Polyadenylation

The standard labeling assay for GeneChip eukaryotic expression arrays involves a 3' bias. As a result, sequences from which probe sets are to be selected should be within 600 bases of the polyadenylation site. Regions further away are less likely to be detected by the standard labeling assay. There are several metrics that can be

used to evaluate whether or not a sequence captures the 3' end of the transcript. These include:

- **3' EST reads:** Given that ESTs can be mislabeled and not represent the true 3' end, multiple 3' EST reads or confirmation from other metrics should be used. When possible, multiple 3' ESTs or confirmation from other metrics should be used.
- **Polyadenylation sites and signals:** The presence of a polyadenylation site and signal provides strong evidence that a sequence captures the 3' end. (See additional comments above.)
- **Putative full-length cDNA sequences:** If a sequence is supported by a putative full-length cDNA sequence, it is likely that the sequence includes the 3' end. In general, a putative full-length cDNA sequence has both an annotated CDS and 3' UTR. A number of cDNA sequences are submitted with the UTRs trimmed off, so this is an important check to make before relying on this metric.

Alternative polyadenylation can complicate the determination of 3' ends. If the alternative polyadenylation sites are relatively close, then submission of the region upstream of the most 5' polyadenylation site is preferred. If the sites are relatively distant (> 600 bases apart), then multiple probe selection regions can be specified within the Instruction File so that probe sets are picked against the different alternative polyadenylation variants.

2.5.4 Sequence Quality

The sequences submitted for probe selection should be of high confidence. Individual bases or regions of uncertainty should be masked out with "N"s. This uncertainty could be based in biology, such as coding SNPs, or it could reflect the low-quality portion of a read or an area of ambiguity within a cDNA assembly. The risk of an undetected coding SNP affecting the performance of the chip is substantially mitigated by the presence of multiple 25-mer probes per probe set and the expression analysis algorithms. Nevertheless, known or computed variation and ambiguities should be masked out.

Sequence quality is an especially important issue when dealing with EST sequences. When possible, trace files or base call scores should be used to remove low-quality sequences prior to use in a design. If these are unavailable, other possible solutions are to use low/high-quality annotations, genomic alignments to infer high- and low-quality regions, or to trim EST sequences to an expected size. Appropriately, trimming off low-quality sequences will improve cDNA clustering (if the design involves clustering cDNAs) and will reduce the chance that probe sets are selected in low-quality regions of a sequence.

Many genes in eukaryotic genomes undergo alternative splicing. Often, alternatively spliced transcripts have identical 3' ends. In such cases, a single common probe set is generated, which can identify transcripts from the gene, but cannot differentiate between the splice variants. If alternative splicing is occurring

in the 3' end of the transcript, then both variants should be submitted. When possible, probe selection will generate a unique probe set for each of the variants and a common probe set that recognizes all of the variants. Depending on the intended use of the design, retention of all probe sets, only the unique probe sets, or only the common probe sets may be desired.

For eukaryotic organisms with long 3' UTRs, most of the effective probe sets will be within the 3' UTR due to the 3' bias of the labeling assay. For organisms with short 3' UTRs, there is the option of using just the 3' UTR sequence, just the coding sequence, or both. Designs based on the CDS region will be less likely to differentiate between genes in the same gene family and may be too distant from the 3' end to be detected with the standard labeling assay. For 3' UTR-based designs, there will be more specificity in the probe sets for distinct genes. One need not submit just 3' UTR or just CDS portions of sequences. Regions of the sequence spanning both these feature types can be submitted.

2.5.5 Sequence Prioritization

Since the number of probes on any given GeneChip array format is finite, prioritizing the probe sets is usually required. This prioritization depends on the intended application of the design; however, some general rules should be noted. Usually, probe sets against well annotated and supported sequences are preferred over more speculative and poorly annotated sequences. The confidence of orientation and the 3' transcript end can be used to prioritize those probe sets that are more likely to perform well over those that may be against the wrong strand or too 5' to be labeled. Another area for prioritization focuses on depth versus breadth of gene coverage. For instance, selecting only a single probe set per gene, rather than multiple probe sets per gene covering alternative splicing and polyadenylation, will allow for more genes to be represented on the chip at the risk of losing sensitivity.

2.6 Probe Selection

2.6.1 Probe Selection Process

Probe selection and array design lie at the heart of the reliability, sensitivity, specificity, and versatility of GeneChip probe arrays. The probe selection algorithm for expression designs is intended to find the optimal set of probes for a given sequence while minimizing the number of probes used.

One of the key elements of selection and design is the new probe quality model. This model studies the thermodynamic properties of each probe to predict its hybridization potential. Hybridization, under particular pH, salt, and temperature conditions, can be optimized by taking into account melting temperatures and by using empirical rules that correlate with desired hybridization behaviors. Based on data from thousands of experiments that monitor the hybridization of target sequences in complex mixtures, Affymetrix uses computer models to predict the intensity and concentration-dependence of probe hybridization. Thus, data quality can be maximized while probe number is minimized.

Another core element of array design is the perfect match/mismatch probe strategy. For each probe designed to be perfectly complementary to a target sequence, a mismatch probe is generated that is identical except for a single base mismatch in its center. These probe pairs, called the perfect match (PM) and the mismatch (MM) probes, allow the quantification and subtraction of signals caused by non-specific cross-hybridization. The difference in hybridization signals between the PM and MM probe pairs, as well as their intensity ratios, serve as indicators of specific target abundance.

The probe selection process begins by generating a pool of all possible probes from the sequences submitted. Each of these possible probes is then evaluated against various criteria, such as the number of synthesis steps, ambiguities, etc. to determine if it is viable. All viable probes are then evaluated against the probe quality model and assigned a probe score. They are also evaluated for potential cross-hybridization to other sequences. When trying to select the best set of probes from all viable probes for a given sequence, the raw probe quality score, the cross-hybridization potential, and the distance between different probes are all taken into consideration to make the optimal probe set.

2.6.2 Probe Set Types

There are different types of probe sets that can result from the probe selection process. Most probe sets have an extension of an underscore and a letter to designate the probe set type, except for unique probe sets.

2.6.2.1 Gene Family Probe Sets (_a)

Probes in a gene family probe set (_a set) all cross-hybridize to the same set of sequences that belong to the same gene family (i.e., having same name in the “geneCluster” column). This probe set type is only created if the “geneCluster” column is included in the Instruction File and contains information.

2.6.2.2 Unique Probe Sets

Probes in a unique probe set do not cross-hybridize to any other sequences in the design (including any additional pruning sequences provided).

2.6.2.3 Identical Probe Sets (_s)

Probes in an identical probe set (_s set) all cross-hybridize to the same set of sequences that are used for the design (including any additional pruning sequences, if provided). These sequences are not defined as from the same gene family for one the following reasons: the values in the “geneCluster” column are different, or the gene family information is not provided.

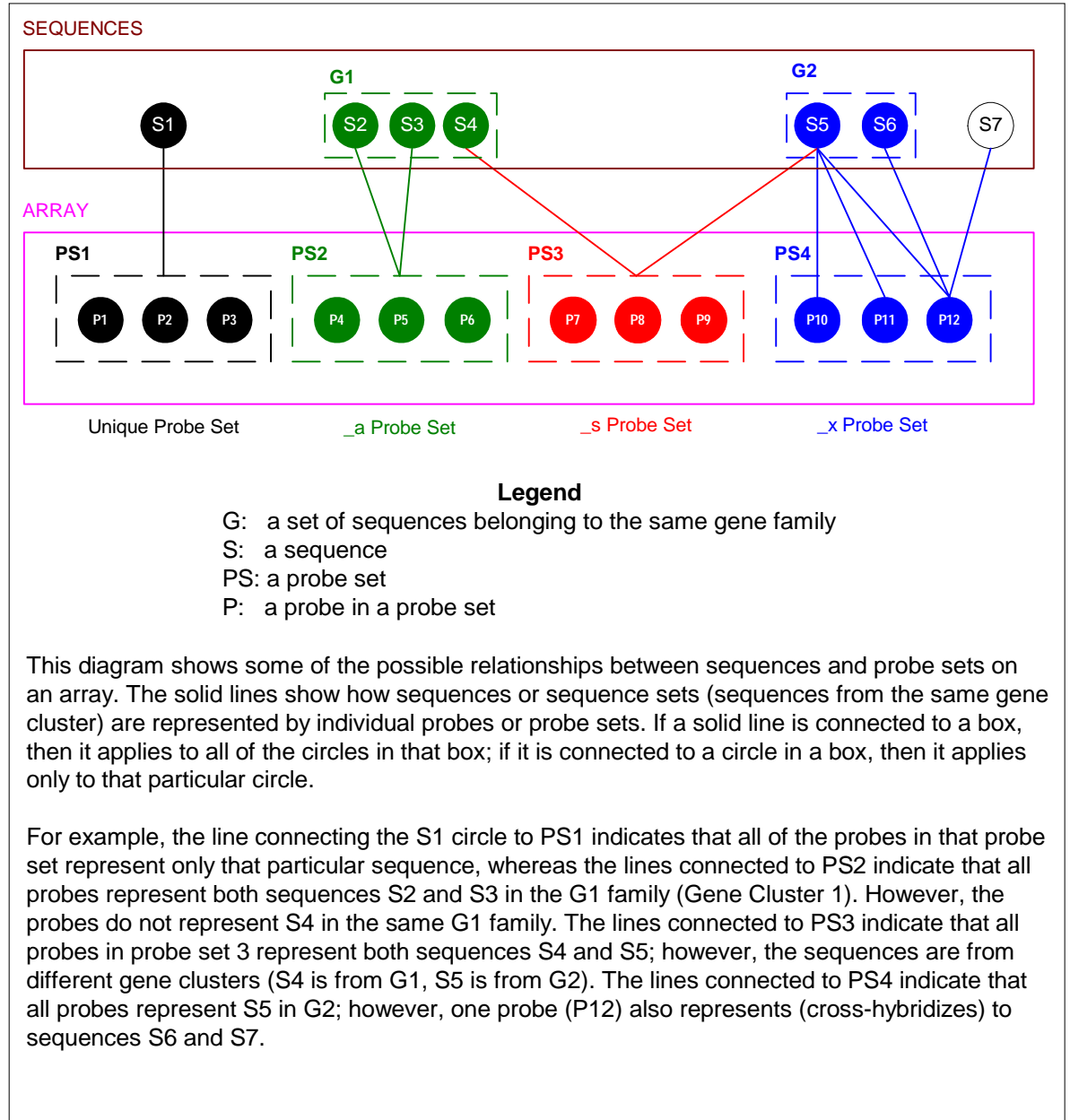
2.6.2.4 Mixed Probe Sets (_x)

Probes in a mixed probe set (_x set) contain at least one probe that cross-hybridizes with other sequence(s) used for the design. Cross-hybridizing

probes have a cross-hybridization penalty applied to their raw probe scores, thus, favoring unique probes of the same quality over cross-hybridizing probes.

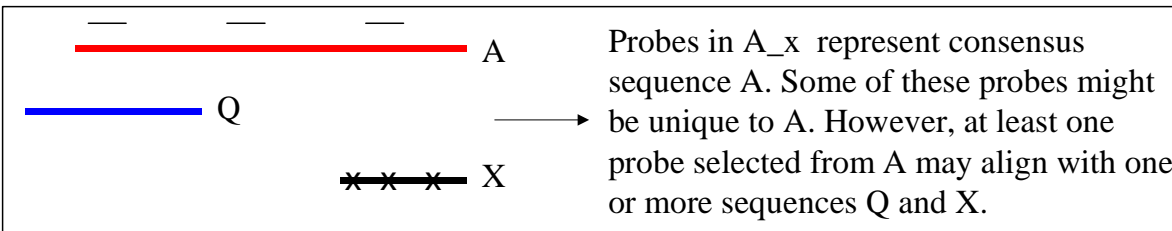
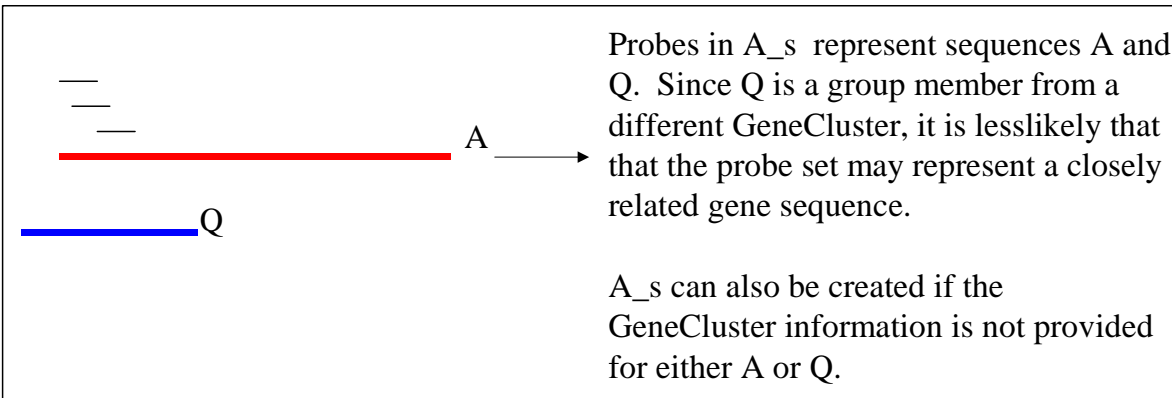
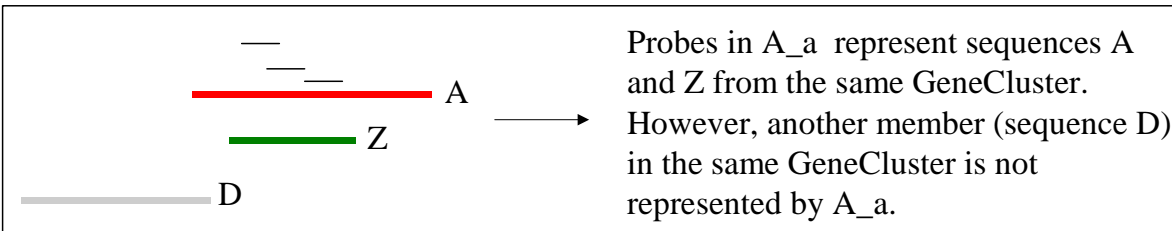
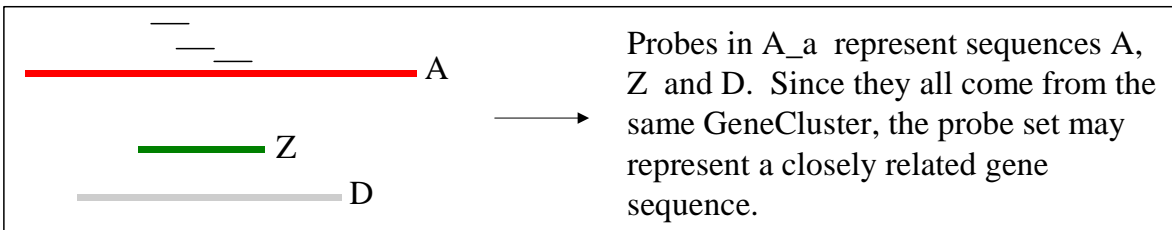
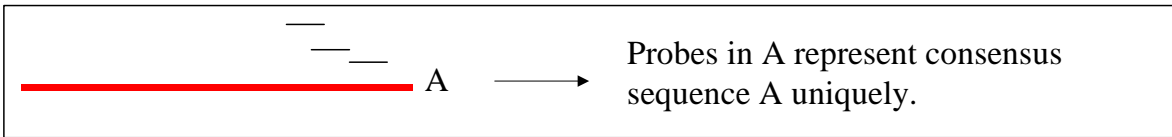
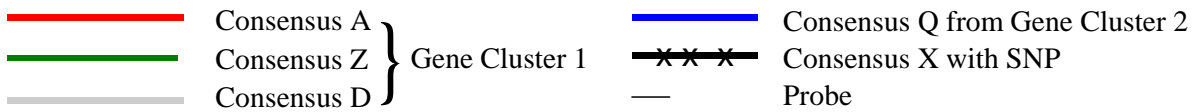
A mixed set is attempted if neither gene, unique nor identical probe sets, can be produced or if none of the probe sets produced meet the default probe set score threshold ($0.36 \times \#$ probe pairs requested).

The following diagram is a graphical representation of these different probe set types:



2.6.2.5 Examples of Probe Set Types

Here are some examples of how different probe set types can be created.



2.7 Pruning for Probe Selection

Pruning is a sequence comparison method. The standard practice for probe selection is to prune against specific bacterial and species-specific controls, in addition to any custom sequences provided for the design. Pruning increases the quality of the unique probe sets selected for the design and reduces the risk of cross-hybridization with other sequences. There are two types of pruning performed in probe selection — hard pruning and soft pruning.

2.7.1 Hard Pruning

Sequences used for hard pruning are generally highly repetitive elements, such as alu-like elements, or abundantly expressed RNA, like rRNA. Probes that cross-hybridize to hard pruning sequences are not included in a probe set.

2.7.2 Soft Pruning

After hard pruning is complete, the design sequences are pruned against other sequences for cross-hybridization. These sequences include the standard bacterial and species-specific Affymetrix controls, all of the design sequences, as well as any additional pruning file(s) provided for the design. Probes that cross-hybridize to these soft pruning sequences will be noted, and if a probe set is later made from these cross-hybridizing probes, the corresponding soft pruning sequences will be logged as cross-hybridization members (see Section 4.3 for details).

2.8 Array Controls

Array controls are an important part of the array design. Tiling control probe sets on an array allows you to verify that hybridization and samples are working correctly. Depending on how they are used, they can act as controls to monitor scaling and normalization process during probe array analyses. They also allow for troubleshooting, should there be problems with the arrays.

2.8.1 Standard Controls

A set of standard controls are included on all expression array designs. These standard controls include:

a) Quality Control (QC) and Alignment Controls:

- Spaced normalization grid
- Grid alignment corner checkerboards
- Text on the array
- Edge controls
- Borders
- QC oligo controls (absent on CustomExpress designs)
- Center cross

b) Target Preparation and Hybridization Controls:

- Hybridization Controls: *bioB*, *bioC*, *bioD* from *E. coli*, and *cre* from P1 bacteriophage
- Poly-A Controls (for sample preparation): *dap*, *lys*, *phe*, *thr*, *trp* from *B. subtilis*
- Species-specific controls for sample quality

For the Hybridization Controls, three *E. coli* genes and one phage gene (*cre*) are used (available from Affymetrix, part #900299 or 900362). Probe pairs representing the 5', Middle, and 3' regions of these genes are present on the array. Labeled antisense RNA transcripts from control genes, along with target cRNA, are hybridized to the arrays. These hybridization controls can be used to assess overall sensitivity and provide a rough calibration curve to estimate mRNA abundance.

For the poly-A Controls, five *B. subtilis* genes are used. Probe pairs representing the 5', Middle, and 3' regions of these sequences are present on the array. The control gene constructs contain a poly-A sequence at their 3' end. Sense strand cRNAs synthesized from the control genes can be added to samples prior to the reverse transcription step in the target-labeling protocol to monitor target synthesis and labeling efficiencies. Labeled antisense RNA transcripts from control genes, along with target cRNA, are hybridized to the arrays, serving as sample process controls in addition to indicators of assay sensitivity.

For species-specific controls, only species available through commercial designs are available as standard controls. If your design is for a species for which we do not have a commercial design, you will need to provide us with your own set of species-specific controls. A list of the available species and their standard controls can be found on the Affymetrix web site http://www.affymetrix.com/support/technical/technotes/expression_controls_tech_note.affx.

2.8.2 Species-Specific Controls

For species-specific controls we generally select the following types of sequences:

- rRNA sequences, which serve as controls to assess the ribosomal RNA contamination of target samples. These same rRNA sequences are also normally used for hard pruning. Some investigators may have an interest in the detection of labeled rRNA to monitor historical levels of rRNA in their target preparations to get another overall checkpoint in the assay. We prefer not to set limits on the range in detected rRNA. It has not been our experience that randomly labeled eukaryotic rRNA produced during sample preparation in our 3' biased assay interferes with data quality. While we may want to monitor rRNA contamination in the *E. coli* sense assay, where enrichment of non-ribosomal RNA is required, the antisense prokaryotic assay is not generally influenced by potential rRNA contamination. Probe sets for rRNA

detection in the antisense prokaryotic assay provide a function similar to the checkpoint function in the eukaryotic assay.

- Constitutively expressed genes, which can be used for scaling and/or normalization of arrays, for assessing the overall labeling efficiency, and for assessing the quality of biotin-labeled target sample for our 3' biased eukaryotic assay (not cDNA assay). Common genes used in other mRNA expression assays (e.g., glyceraldehydes 3-phosphate dehydrogenase (GAPDH) and beta-actin) are typically selected. For the 3' biased assay, transcript lengths greater than 1,000 bases are recommended as a way to look at the trend in labeling efficiency over the range of typical cRNA transcribed (i.e., ~1,000) bases. Each of these should be broken down into the 3', middle, and 5' regions for probe selection.

2.8.3 Other Controls

2.8.3.1 Contaminant Controls

To assess when genomic DNA is contaminating the total RNA preparation used to make labeled target for the expression assay, we recommend that you tile intergenic, or non-transcribed, portions of genomic DNA.

2.8.3.2 Scaling and Normalization Controls

Scaling is the first critical step in analyzing expression profiles from two or more microarray experiments. When building sets of arrays, or when building a subset array of a larger probe array format, special attention should be paid to scaling and normalization strategies for expression data.

To quantify changes in expression, it is important to minimize systematic variations from accountable and predictable sources that cause differences in absolute expression measurements (i.e., variations in labeling efficiency, quantification of labeled target, hybridization, washing, staining, and chip performance). Scaling and normalization strategies have been developed to account for these naturally occurring experimental variations.

Scaling and normalization strategies in the Affymetrix expression assay are similar to those used for other expression assays. For example, in Northern Blots Analyses, quantifying expression between experimental states is often measured after normalizing to a common constitutively expressed gene, such as actin. Since actin is thought to have a very similar level of expression in different experimental conditions, scientists can proportionally factor, i.e., “normalize”, expression levels of other genes relative to the actin level of expression in each experiment. In a similar fashion, artificially introduced spike-in transcripts can be introduced to simulate a constitutively expressed gene. Scaling and normalization to probe sets representing these spike-ins can also be very informative.

For more information about scaling and normalization strategies, please visit the Affymetrix web site:

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

2.9 Custom Design Output Files

For each custom expression design, you will receive a set of Library Files corresponding to your custom arrays. The Library Files are necessary for expression analysis using the Affymetrix® brand data analysis software, Microarray Suite (MAS). An installation program that loads the Library Files on your desktop or Laboratory Information Management System (LIMS) is provided for each design.

The MAS 5.x algorithm allows the setting of parameters based on statistically derived *p*-values to determine the Absent/Present calls and the Increase/Decrease calls. (Please refer to Affymetrix Technical Note, *New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays* at http://www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf.) You can increase your statistical degrees of freedom for each probe set by increasing the number of probe pairs. Please consult with your statistician or FAS for more information.

If you would like to receive the probe sequences for your design, please send your request to the Chip Design Group at: chip_design@affymetrix.com.

3. SUBMITTING A DESIGN REQUEST

The Chip Design Group will not start your design until both the Purchase Order and design information have been received.

3.1 Purchase Order

A Purchase Order (PO) is a commitment to buy our products and services. A PO for a CustomExpress™ Array consists of the following line items:

1. design fee³
2. order for the first lot(s) of arrays (refer to Section 2.3 for minimum purchase requirements).

The PO should be faxed to the Affymetrix Customer Service Group (see Section 5 for contact info).

3.2 Design Information

3.2.1 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:

http://www.affymetrix.com/support/technical/other/custom_designform.doc

In the Design Request Form, you will need to provide the following information:

3.2.1.1 Requestor Information

The Requestor Information fields provide us with 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 lines regarding Affymetrix contacts are optional; however, this information is helpful as it allows us to notify your Account Manager and FAS of the design status as well.

3.2.1.2 Expression Probe Array Information

The Expression Probe Array Information fields provide general information regarding your design.

3.2.1.2.1 Array Name

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

The final array name for a custom design is created in this fashion:
<customized name><1-letter strandedness><part number>F

The customized name is provided by you in this “Array Name” field. The array strandedness uses “a” for antisense designs and “s” for sense designs.

³ Please contact your account manager if you have any questions.

The part number is the part number we assign to your arrays. The final character “F” is only used for CustomExpress Advantage Arrays.

For example, if you have an antisense custom design with “Rat1” as the Array Name in this field, and we assign part number 510777 to your design, then the final array name will be: Rat1a510777.

If your design is an antisense CustomExpress Advantage design with “Rat1” as the Array Name, and we assign part number 510999 to your design, then the final array name will be: Rat1a510999F.

3.2.1.2.2 **Array Description**

Please provide a description for your design. It can be up to 26 characters.

3.2.1.2.3 **Feature Size**

Please select a feature size. Options include: 18, 20, 24, and 50 micron.

3.2.1.2.4 **Array Format**

Please select the array format of your choice. See sections 2.3 and 2.4 for more details.

3.2.1.3 **Design Probe Sets**

The Design Probe Sets section allows you to specify the information used to design your CustomExpress Arrays.

3.2.1.3.1 **Sequence File Name**

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

3.2.1.3.2 **Instruction File Name**

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

3.2.1.3.3 **Probe Set Request File Name**

If you would like to include commercial content as part of your design, you should list desired probe sets in a Probe Set Request File. Multiple file names may be specified by providing one file name per line.

3.2.1.3.4 **Target Type**

Please select the target type for your design. Options include:

- Eukaryotic Antisense, RNA
- Prokaryotic Antisense, cDNA
- Prokaryotic Sense, RNA

The default selection is Eukaryotic Antisense, RNA.

3.2.1.3.5 **Probe Selection Region**

Please select one of the following options for defining the probe selection region of your sequences.

- **600 bases from 3' end:** This default option sets the probe selection region to be the last 600 bases from the 3' end (30 bases from the end of the sequence).
- **Full Length Sequence:** This option should be selected if you want probe selection to be performed across the entire sequence submitted. This is normally recommended for Prokaryotic cDNA designs.

- **Per Instruction File:** This option should be selected if you have specified the start/end regions in your Instruction File. This allows you to set the probe selection region on a sequence by sequence basis.
- **Other:** If you want to specify another probe selection region for all of the sequences submitted, you can select this option. For example, if you want to select probes from 1,000 bases from the end of your sequences, you should select this option, and specify 1,000 from End, and 0 from End. The “600 bases from 3’ end” option would correspond to 630 from End, and 31 from End. The “Full Length Sequence” would correspond to 0 from Start and 0 from End.

3.2.1.3.6 Number of Probe Pairs per Sequence

There are two options available for designating the number of probe pairs per sequence:

1. Set the same number of probe pairs for ALL sequences. This is recommended.
2. Set a different number of probe pairs per sequence (see Sections 3.2.4.1.13 and 3.2.4.1.14 for details).

If we cannot produce a probe set with the Desired Number of probe pairs, please specify the Minimum Number that is acceptable to you for your design. This number should be an integer greater than 0 and less than or equal to the number of desired probe pairs per sequence. For an 11 probe pair per probe set design, we recommend a value of 8 for this field.

3.2.1.4 Controls

The Controls section allows you to specify the species and controls to be used with your design.

3.2.1.4.1 Species on Array

Please provide us with the species for your design. More than one species may be checked if it is a multi-species design. The value(s) in this field will help us determine species-specific controls to tile on your array.

If “Other” is checked in the Species field, please provide the name(s) of the species, as well as your strategy for controls.

3.2.1.4.2 Controls (for Pruning & Tiling)

Select the appropriate controls to be used for pruning and tiling on your arrays. In general, we recommend selecting the standard controls corresponding to the species of your design.

In addition to the standard controls, normalization controls are available for human, mouse, and rat. These normalization controls are tiled on Affymetrix® catalog arrays. If you plan to correlate your custom array data to the corresponding Affymetrix catalog arrays, we strongly recommend tiling the normalization set as well.

If your design is a Prokaryote, you may also select our Prokaryotic Spike Set, which includes Yeast and HXB2 sequences.

3.2.1.5 Pruning

The Pruning section allows you to specify how pruning should be performed, as well as additional sequences to be used for hard and soft pruning.

3.2.1.5.1 When Selecting Probes

When selecting probes, these are the options available for how pruning can be done:

- **Prune all sequences against all other sequences:** This is the recommended default option. All sequences for a given design (including the soft pruning and hard pruning sequences) will be pruned against each other when selecting probes.
- **Prune sequences only against other sequences from the same instruction file:** This option is used if the design contains design sequences for multiple species, and probe selection should be done independently for any single species. For example, you have sequences for mouse and rat on the same design, but you would like to have the mouse sequences pruning only against other mouse sequences during probe selection, and rat sequences against only other rat sequences. If this is the case, you should put the sequences in separate Sequence Files so that only sequences that should be pruned together are in the same file. A separate Instruction File should be provided for each corresponding Sequence File.
- **Other:** Please specify clearly how pruning needs to be performed.

3.2.1.5.2 FASTA Format Hard Pruning Sequence File Name

The Hard Pruning Sequence File contains additional sequences you want us to hard prune against your design (for more details on hard pruning, please see section 2.7.1). Multiple file names may be specified by providing one file name per line.

For performance reasons, please try to limit **each** sequence to about 5000bp.

3.2.1.5.3 FASTA Format Soft Pruning Sequence File Name

The Soft Pruning Sequence File contains additional sequences you want us to soft prune against your design (for more details on soft pruning, please see section 2.7.2). Multiple file names may be specified by providing one file name per line.

For performance reasons, please try to limit **each** sequence to about 5000bp.

3.2.1.5.4 Use Affymetrix Standard Pruning Library

This field allows you to request to have your design sequences pruned against an existing set of sequences used by Affymetrix for the most recent catalog design for a given species. For example, if you have a small set of human sequences for your design, you may want to check the “Human” option, and we will prune your sequences against the most recent catalog design for human genome array. This reduces the risk of cross-hybridization of probes selected for your design.

3.2.1.6 Special Instructions

If you have any other design-specific information that cannot be captured in the Instruction File or other fields in the Design Request Form, they can be mentioned here.

3.2.2 Probe Set Request File

To include existing probe sets from our commercial arrays or from your custom array(s) as part of a new design, please provide the desired list of probe sets in a Probe Set Request File.

The Probe Set Request File must be a tab-delimited text file containing the following columns with the following headers:

- **ArrayName:** The name of the commercial array that contains your probe set. This is referring to the short array name, e.g., HG-U95Av2.
- **ProbeSetName:** The name of the probe set.
- **Rename:** This column is only required if a probe set name is duplicated in a design. A new name must be assigned to the duplicated probe set using the value in this column. If two probe sets are duplicated, only the second one needs a value in the “rename” column. The value in the Rename column must have the same last three characters (“_at”) and the same probe set category (i.e., _s, etc.) as the original probe set name. In addition, integers are not allowed as the new name.

Below is a template of how the Probe Set Request File should look (it might not line up as nicely in actual file). You can also download an example of a Probe Set Request File at:

www.affymetrix.com/products/arrays/custom_express/FAQ/SubsetRequestTemplate.txt

ArrayName	ProbeSetName	Rename
HG-U95Av2	12345_at	
HG-U95Av2	12456_at	
HG-U95Av2	12456_at	12456_copy_at

3.2.3 Sequence File

The Sequence File should contain all the sequences from which probes should be selected. If a sequence is to be used for pruning purposes only (i.e., no probes will be selected), then that sequence should be put into the Pruning Sequence File instead. For performance reasons, please try to limit **each** sequence to about 5000bp.

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 must correspond to the value in the “name” column in the Instruction File (see Section 3.2.4 for more information on the Instruction File).
- All names defined in the Instruction File must exist in the Sequence File, and vice versa.

- The sequence name in the design and Pruning Sequence Files may be up to 20 characters that are alphanumeric or special characters: “+,” “@,” “\$,” “%,” “^,” “&,” “(,” “),” “:,” “-,” “_,” “=,” “#,” “~.”
- The sequence name in the design and Pruning 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 design and Pruning Sequence Files may not end in “_at” or “_st”, as these are identifiers already used by Affymetrix.
- A comment may follow the sequence name; however, the information here will not be utilized by the Chip Design Group. Any annotations you would like to include in the Library Files need to be included in the Instruction File.
- No blank lines should exist between sequences.

Please note that we assume all sequences submitted by you are from the sense strand (sense strand represents the forward direction of the gene sequence from 5' to 3'). If, for whatever reason, the sequence you have is from the antisense strand, you should reverse-complement that particular sequence before including it in the Sequence File. For an ambiguously oriented sequence, we would recommend that you include both the sense and antisense strands of the sequence in the Sequence File, and have probes selected for both strands. This will ensure that we are always selecting probes for the correct target.

Example: contents of file “gene.txt” in FASTA format

```
>AA618977      any comments here are ignored
gtttgtctttggtaaagtacctttgcatcatgattcttgagatgtagattattctaggtcccgaatggcttag
atcttcattgattcatagcaagttgtgcatagataagtggtgtgtaaacattgtagaatcattgaggttgat
aaataattcgtatagctgataactggttatagcttgattctattgttgataaaaaaa
>AA618981
agatcttacggactatctgatgaagatctgactgagagaggttacagtttacgacgacagcggaaactgag
atagccgagacatcaaggagaaactgtgttatgttccttgattcgaacaggagatgggtacggcagctt
cgagttcggcgttgagaagattatgagcttcctgatggtaagtgattactattgtaacgagcggtaatt
ctattatgaagtgacgtagatattcgtaaagatctgtaccaacacagtagttgc
>ABCD
agatcttacggactatctgatgaagatctgactgagagaggttacagtttacgacgacagcggaaactgag
atagccgagacatcaaggagaaactgtgttatgttg
```

3.2.4 Instruction File

3.2.4.1 Description of Columns

An Instruction File is a text file containing the following tab-delimited columns:

3.2.4.1.1 name

Mandatory column. The Accession Number in GenBank or the ID referring to a sequence in the public domain databases, or if not applicable, then a proprietary ID up to 20 characters that does not end with an underscore followed by a single alphanumeric character (i.e., _i, _3, etc. may not be used).

A value for the name column must be present for all entries in the Instruction File. Every value in the name column must have a

corresponding entry within the Design Sequence File. All rules that apply to sequence name will, therefore, apply to the value in the name column (see Section 3.2.3 for details).

If you want to give a name that is different than the sequence name to the probe sets that represent your submitted sequences for use in data analysis, you may do so by providing an alias (see “alias” field description below). If no alias is provided, then the contents of the “name” field will also be used as the probe set name.

3.2.4.1.2 database

A mandatory column that defines the biological database from which the sequences were derived. The value for this column can be up to 50 characters in length. If the sequences are NOT from a public database, please use “Proprietary” as the value for this column. Currently, Affymetrix only supports linking to the following public databases:

“database” value	Name in MAS	URL
GenBank	Entrez	http://www3.ncbi.nlm.nih.gov:80/htbin-post/Entrez/query?form=6&db=n&Dopt=g&uid=%s
Stanford Public Database ⁴	SGD	http://genome-www.stanford.edu/cgi-bin/dbrun/SacchDB?find=Locus+%22%s%22
BDGP	BDGP	http://www.fruitfly.org/cgi-bin/EST/community_query/cloneReport.pl?id_type=0&id_value=%s&db=estlabtrack
Flybase	Flybase	http://flybase.bio.indiana.edu/.bin/fbidq.html?%s
NCBI Ecoli Genome	NCBI Ecoli Genome	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/altvik?gi=115&db=g&from=%s
Pseudomonas	Pseudomonas	http://www.pseudomonas.com/AnnotationByPA.asp?PA=%s
Entrez Pseudomonas Stable RNAs	Entrez Pseudomonas Stable RNAs	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/rnatab?gi=163&db=%s
Entrez Pseudomonas Genome	Entrez Pseudomonas Genome	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/altvik?gi=163&db=g&from=%s
TubercuList	TubercuList	http://genolist.pasteur.fr/TubercuList/genome.cgi?gene_detail+%s
WormBase	WormBase	http://www.wormbase.org/db/seq/sequence?name=%s
TREMBL	TREMBL	http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[EMBL:%s]+-newId
TIGR-ath1	TIGR-ath1	http://www.tigr.org/tigr-scripts/e2k1/euk_display.dbi?db=ath1&locus=%s

If you have other databases to which you would like to link, you might be able to do so by executing SQL statements. Please check with the Chip Design Group or your FAS for more details.

3.2.4.1.3 alias

The “alias” is the designation used for the probe set name. It can contain up to 20 characters and follows the same rules as those for sequence names.

If multiple entries in the Instruction File contain the same value in the alias column, then the only values that can be different are start, end, startSeq, and endSeq.

If multiple entries in the Instruction File contain the same value in the alias column, then the Start and End columns also need to be specified. The

⁴ This is the Saccharomyces Genome database.

regions specified by the Start/End values must not overlap by more than 24 bases and must not be a subset of each other.

Please refer to Section 3.2.4.2 for examples of different interactions between “name” and “alias”.

3.2.4.1.4 **geneCluster**

Please provide the name of the gene the sequence represents. If multiple sequences are classified to be from the same gene cluster, we try to make a probe set that will best represent ALL of the sequences within the same cluster (i.e., `_a` set). However, an `_a` set simply means that all the cross-hybridizing members are from the same gene cluster, not necessarily that all of the gene members are covered by that `_a` set.

If this column is not provided, we will assume that every sequence is from a different gene cluster. If this column is provided but some of the sequences do not belong to any gene clusters, this field can be left blank.

3.2.4.1.5 **transcriptExpressed**

This column may contain any value between 0 and 1 to state the level of expression for a particular sequence. Higher values correspond to more expressed transcripts. For example, a sequence like human alu will probably get a value of “1,” whereas a very rare gene would be assigned a value closer to 0.

Another possibility is to use this column to state the relative size of the gene cluster. For example, if a sequence belongs to a really large cluster (with over 100 members), you might want to assign a value close to or equal to “1.”

The value in this column affects how the mixed (`_x`) sets are made. If two probes (P1 and P2) with equal probe scores cross-hybridize to two different sequences (P1 to sequence A and P2 to sequence B), where A has a value of 1 (coming from a large gene cluster) and B has a value of 0.3, then P1 will be penalized more than P2, and P2 will be selected to make up the `_x` set.

3.2.4.1.6 **start**

Within the sequence, the numeric (1-based) position of the start of the region of interest. This column is required if “end” is specified. Normally, probes are selected from the last 600 bases from the 3’ end (the 3’ end is the 30th base from the end of the sequence). If this is not desirable, then the “start” and “end” columns should be used. For example, a sequence with 1,000 bases will have a probe selection region from base 371 to base 970. A sequence with 630 bases will have a probe selection region from base 1 to base 600. In all cases, the last 30 bases are dropped to avoid regions that are typically poor quality near the poly-A tail.

3.2.4.1.7 **end**

Within the sequence, the numeric (1-based) position of the end of the region of interest. This column is required if “start” is specified.

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.

3.2.4.1.8 startSeq

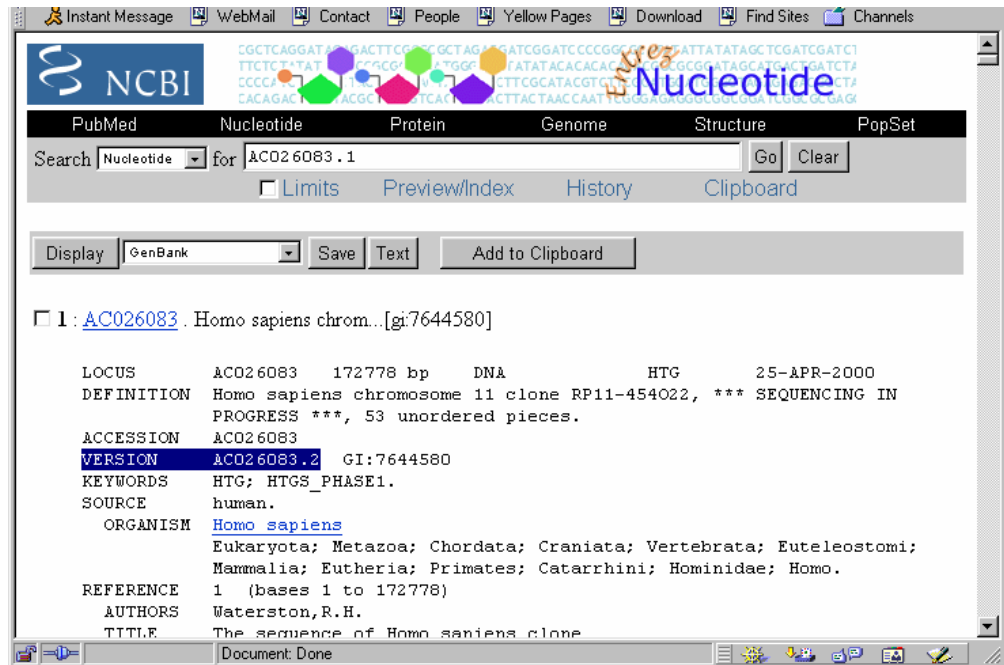
A sequence used for checking purposes. It includes the base at “start” and the next few bases up to a total of eight bases. This column is required if “start” is specified. The information here allows us to cross check your sequences, making sure that we process the sequences in the same method you expect.

3.2.4.1.9 endSeq

A sequence used for checking purposes. It includes the base at “end” and the previous few bases up to a total of eight bases. This column is required if “end” is specified.

3.2.4.1.10 hyperlink

Provide the identifier in the public databases from which the sequences were derived, preferably the GenBank version number, when possible (see picture below for an example of the version number). If this information is not provided, the values in “name” are used.



The screenshot shows the NCBI Nucleotide search interface. The search bar contains "AC026083.1" and the results list shows one entry: "1 : AC026083 . Homo sapiens chrom...[gi:7644580]". The entry details are as follows:

LOCUS	AC026083	172778 bp	DNA	HTG	25-APR-2000
DEFINITION	Homo sapiens chromosome 11 clone RP11-454022, *** SEQUENCING IN PROGRESS ***, 53 unordered pieces.				
ACCESSION	AC026083				
VERSION	AC026083.2 GI:7644580				
KEYWORDS	HTG; HTGS_PHASE1.				
SOURCE	human.				
ORGANISM	Homo sapiens				
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 172778)				
AUTHORS	Waterston, R.H.				
TITLE	The sequence of Homo sapiens clone				

3.2.4.1.11 hyperlinkDatabase

The name of the public database for the hyperlink. If “hyperlink” and “hyperlinkDatabase” are provided, you will be able to click on the hyperlink within the Affymetrix® Microarray Suite software and view the sequence information through your web browser (if the public database supports links). If this is not provided, the values in “database” will be used. Please see “database” for a list of public databases to which we currently link.

For example, if you would like to link to a related public entry in GenBank for one of your proprietary sequence, you can do so by using “hyperlink” and “hyperlinkDatabase” to specify the public information, and still provide your sequences with its proprietary identifier, and specify the database to be Proprietary.

- 3.2.4.1.12 description**
An optional description of the sequence up to 4,000 characters. However, any characters after the 255th character will be truncated when the information is viewed in MAS. This serves as the annotation for your sequences in the Microarray Suite software.
- 3.2.4.1.13 probes**
The number of probe pairs per sequence. This column should only be used if you want to select different numbers of probes for different sequences. Every value in this column must be greater than or equal to the Minimum Number of Acceptable Probes specified in the Design Request Form.
- 3.2.4.1.14 minProbes**
The minimum number of probe pairs that is acceptable per sequence. This value should be used if the “probes” column is used. It should be greater than 0 and less than or equal to the value in the corresponding “probes” column.
- 3.2.4.1.15 copies**
The number of times the same probe set representing the same sequence is to be tiled on the array. If not provided, all probe sets are tiled only once. However, if it is desirable for some of the probe sets representing specific sequences to be tiled more than once, then this column should be used.

Example (Instruction File):

name	alias	copies
AB0001	seqAB	3

These replicate probe sets will be designated on the array as “seqAB_at” (the original set) and “seqAB_copy1_at” and “seqAB_copy2_at” (two copies of the original set).

- 3.2.4.1.16 classification columns**
Finally, you may include up to 10 classification columns of additional data for your sequences. These columns can have any headings other than those mentioned above and are limited to 256 characters each. These columns are for descriptive purposes only. For example, some popular classification columns include: species, type of sequence, locus link, public database info, gene name, product, sequence orientation, etc. These data will only be available in the Affymetrix Data Mining Tool software.

Additional characteristics of the Instruction File:

- You should include only columns you use. For example, if you originally include a column in the Instruction File and later find that no values exist at all for this column, it should be removed from your file.
- All alphanumeric and character fields are case insensitive.
- Columns should not be duplicated.
- If there is a value for an optional column (e.g., alias) in one row, then there must be values for that optional column in ALL rows, except for the following columns:
 - description
 - classification columns

The values you provide in the following columns will directly affect the array design:

- alias
- geneCluster
- transcriptExpressed
- start
- end
- copies
- probes
- minProbes

The values you provide for the other columns will have no effect on the array design. However, providing information in these columns gives you more information on your array, which is useful when analyzing your data.

3.2.4.2 Examples⁵

1. An example of a simple design is shown below. Probes are selected from the last 600 bases from the 3' end (where the 3' end is the 30th base from the end of the sequence) and will name the probe sets using the same name as the sequences:

name	database
A	Proprietary
B	Proprietary

2. To specify a probe set name different from the sequence name, you can add an "alias" column to your Instruction File:

name	database	alias
A	Proprietary	C
B	Proprietary	D

The above instruction will give you exactly the same probe sets as instructions used in Example 1 (in terms of probes). However, the probe set names will be different. The probe set for sequence A will be called C(C_at), and D for sequence B.

3. To select probes from a specific region of interest, you should add the "start," "end," "startSeq," and "endSeq" columns:

name	database	start	end	startSeq	endSeq
A	Proprietary	1	70	ACCG	CCGT
B	Proprietary	100	200	GGGA	TAAA

Probes for sequence A will be selected from the first 70 bases of the sequence. Probes for sequence B will be selected from the 100th base to the 200th base of the sequence.

⁵The examples are provided for illustrative purposes only. Only the relevant columns are shown and they might not line up similarly in a real Instruction File.

- To select probes from disjoint regions within a sequence, you can specify the Instruction File as follows:

name	database	alias	start	end	startSeq	endSeq
A	Proprietary	C	1	70	ACCG	CCGT
A	Proprietary	C	80	200	GGGA	TAAA

Bases 71 through 79 will be masked out with “N”s. Probes will then be effectively selected from the concatenated region of bases 1 through 70 and 80 through 200. The probe set generated will be called C to distinguish it from the original sequence, since it is not the entire sequence that we are using to select probes. This can be a good way to avoid selecting probes from certain intronic regions within your sequence.

- To have more than one probe set created for the same sequence, you would need to specify a different alias for each probe set you are interested in tiling:

name	database	alias	start	end	startSeq	endSeq
A	Proprietary	C	1	450	ACCG	CCGT
A	Proprietary	D	550	1000	GGGA	TAAA

Probe set C will be selected from bases 1 through 450 of sequence A. Probe set D will be selected from bases 550 through 1,000 of sequence A. This is typically used when you suspect there are multiple genes within the sequence (such as having multiple poly-A sites).

- To produce a probe set among the common region of different sequences belonging to the same gene family, you may take advantage of the “geneCluster” column.

name	database	geneCluster
A	Proprietary	geneA
B	Proprietary	geneA
C	Proprietary	
D	Proprietary	geneX
E	Proprietary	geneX
F	Proprietary	geneY

Sequences A & B are from the same gene family geneA, and as much as possible, we will try to select a probe set (_a set) that covers the common area between the two. This also applies to sequences D & E. For sequence C & F, no _a sets will be attempted since they are treated as single members from different gene families.

3.2.5 Pruning Sequence File

By default, we prune against all sequences submitted for probe selections, as well as certain Affymetrix controls (*bioB*, *bioC*, etc). If the design is for human, murine, rat, yeast, or arabidopsis, we also prune your sequences against constitutively expressed genes and hard pruning rRNA sequences.

If there are other sequences that you would like us to prune against while selecting probes, they should be included in the Pruning Sequence File (we will not select probes for these additional sequences). The format of this file is the same as the Sequence File (see Section 3.2.3).

If you only want us to prune against the sense strand, then only the sense strand sequences should be included. If you want us to prune against both strands, please make sure that you reverse-complement these sequences, and add them to the Pruning Sequence File. In addition, you may want to provide us with a Sequence File containing alu-like elements, or abundantly expressed RNA, such as rRNA for hard pruning, especially if Affymetrix does not have a catalog design for your organism of interest.

If you wish, you may also specify to prune against Affymetrix' standard library sequences used for the catalog GeneChip® arrays. This can be specified in the Design Request Form by selecting the corresponding species from the Affymetrix standard library field.

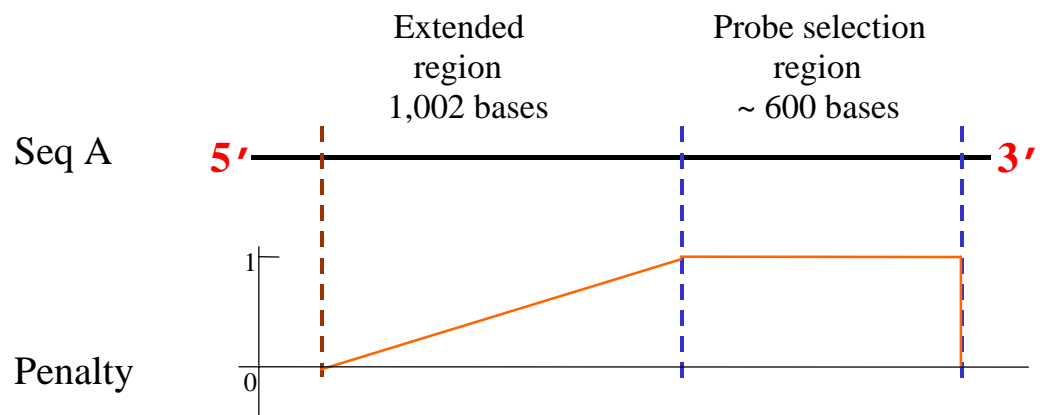
4. UNDERSTANDING THE DESIGN PROPOSAL

The Design Proposal is sent to you when Affymetrix finishes selecting probes for your design. The Design Proposal lists all of the sequences you have requested for your design, followed by columns indicating how these sequences are represented.

4.1 Probe Set Redundancy Reduction in Design Proposal

The Design Proposal we send will have redundant probe sets removed. We define redundancy as the ability of a probe set from one sequence to represent other sequences. The main considerations in determining redundancy involve the evaluation of how well one probe set might perform when mapped to another sequence, as well as how many probes from this one probe set fall within the extended probe selection region of another sequence.

4.1.1 Extended Probe Selection Region



The extended probe selection region is defined to be X bases from the 5' end of the probe selection region defined for a given sequence. A multiplicative penalty is assigned for each extended base when calculating the applied probe set score, going from 1 to 0. In other words, the probe furthest from the 3' end of the probe selection region is penalized the most. The current standard is to set the extended regions to 1,002 bases long, which is roughly based on the 3' to 5' ratio we have seen with control probe sets.

4.1.2 Probe Set Score

The probe set score has three components: the probe quality metric, the cross-hybridization penalty, and the gap penalty.

The probe quality metric was developed on sets of Latin Square experiments. The quality metric is defined as the slope of the line that relates natural logarithms of signal intensities and of target concentrations for each probe. The predicted slopes are highly correlated with observed slopes.

For more detail on the Latin Square experiments, refer to the Affymetrix technical note “*New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays*” at:

www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf.

The cross-hybridization penalty is a multiplier applied only to the cross-hybridizing probes in the mixed (_x) sets, thus favoring the selection of unique probes over cross-hybridizing probes of the same raw quality.

The gap penalty is a multiplier applied to probes if they are too close to each other. The current default is 15bp. If probes are within 15bp of each other, then a penalty is applied to the probes.

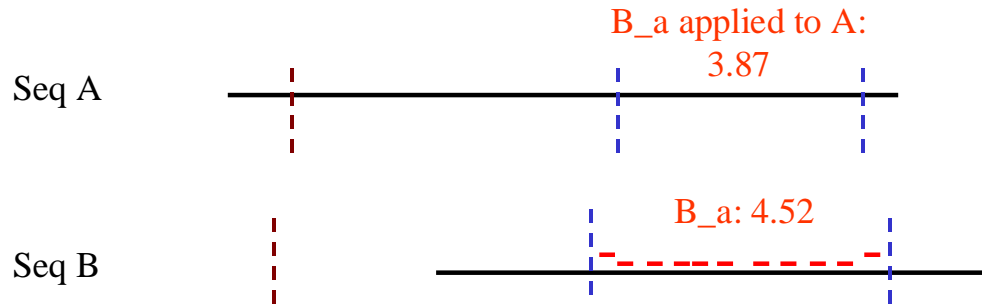
The Probe Set Score is a prediction of how well a probe set will respond in an experiment. It should be used as a guide, not as an exclusive measure of probe set performance. In some cases, probe sets with high probe set scores may not hybridize well, and the opposite can also be true. If there is a sequence with only low scoring probe sets, and the sequence is of high importance to you, it might still be worthwhile to tile a probe set for that sequence.

4.1.3 Applied Probe Set Score

When we use one probe set to represent another sequence other than itself, the alignment of the probes from the original sequence can align differently to the new sequence it represents. Therefore, we re-align the probes to the new sequence, and calculate an *applied* probe set score (probes from sequence B *applied* to sequence A). This applied score gives a prediction of the quality of one probe set representing another sequence.

In the following example, B_a has a probe set score of 4.52, but B_a mapped to sequence A has an applied score of only 3.87. This lower applied score is caused by two factors:

- 1) The last probe in the B_a probe set is to the 3' of the probe selection region for sequence A. Therefore, it receives a penalty of 0, and is discarded when calculating the applied probe set score.
- 2) The first probe in the B_a probe set falls within the extended probe selection region of sequence A. Therefore, it receives a small penalty, making its probe score slightly lower.



4.1.4 Redundancy Reduction Process

When we generate a Design Proposal using results from probe selection, the application goes through an iterative process to remove as many redundant probe sets as possible.

A probe set can replace/cover another probe set if ALL of the following criteria are met:

- 1) probe set from sequence A has X% of probes falling within the probe selection and extended probe selection regions of sequence B that are greater than or equal to the percentage threshold defined;
- 2) probe set from sequence A has an applied score to sequence B greater than or equal to the probe set score threshold; and
- 3) the probe set type from sequence A is a valid replacement for the probe sets from B. For example, if you define identical sets can only replace other identical sets, and you have A_s and Bset (unique set), even though A_s meets criteria #1 and #2 when applied to sequence B, it still will not replace Bset because an identical set is not defined to be able to replace a unique set.

The iterative process starts with the target probe set score threshold (specified in the Design Request Form), processes all the probe sets, and then marks the appropriate sets to be removed if they meet the criteria. These processed probe sets are then set aside. The iterative process then lowers the probe set score threshold by a step size, processes the remaining probe sets, and repeats the iteration again until the minimum probe set score threshold is reached. Any remaining probe sets that are below the minimum probe set score threshold are removed.

4.1.5 Standard Redundancy Reduction Criteria

Here are the default set up values we use to generate your Design Proposal. If you would like to change any of them, please let us know and we will be happy to do so.

- Target Applied Probe Set Score Threshold: <# desired probes × probe score>
- Minimum Applied Score Threshold: the minimum of <1/2 of the target score threshold> or <# minimum probes × probe score>
- % of probes in Extended Probe Selection Region: <from the Design Request Form, (min. # pp) / (# pp desired), round to the nearest 100th decimal places.>

- `_a` set can replace other: `_a`, `unique`, `_s`, `_x` sets
- `unique` set can replace other: `_s`, `_x` sets from the same probe selection region (same alias value)
- `_s` set can replace other: `_s`, `_x` sets
- `_x` set can replace other: `_x` sets

4.2 Design Proposal

A Design Proposal contains the following columns:

4.2.1 Name

The sequence name. This is the same as provided in the Instruction and Sequence Files.

4.2.2 Alias

The instruction alias. This is the same as provided in the Instruction File.

4.2.3 ProbeSetCount

The number of probe sets generated for this sequence based on the Instruction File. If no probe sets can be created, the value is “0.” If one probe set is created, the value is “1.” If two probe sets are created, the value is “2” and there will be two entries for this alias in the proposal describing these two probe sets. This probe set count reflects the total number of probe sets that can be generated for a given sequence, not the number of probe sets that are actually tiled on an array.

4.2.4 ProbeSetName

The name of the probe set consists of the alias plus `_a`, `_s`, or `_x`, when applicable, to describe the type of probe set (see Section 2.6.2 for more details on different probe set types):

1. No suffix (same as the alias) - The probe set is unique.
2. `_a` means gene - All the probes in this probe set can identify a set of sequences from the same gene family, but not necessarily all of the members within that gene family.
3. `_s` means identical - All the probes in this probe set can identify a set of sequences from different gene families.
4. `_x` means mixed - The probe set can consist of unique and cross-hybridized probes that might or might not cross-hybridize with the same set of sequences.

4.2.5 Type

The type of probe set:

- `<blank>` - if no probe set could be generated
- `0` - gene family `_a`
- `1` - `unique`
- `2` - `identical_s`
- `3` - `mixed_x`

4.2.6 ProbeSetScore

This column gives the probe set score for a probe set. See Section 4.1.2 for details.

4.2.7 GapMultiplier

This is the penalty applied to the raw probe quality score used to calculate the probe set score. Dividing the Probe Set Score by this Gap Multiplier gives you the raw probe set score illustrating how the probes might perform if we do not take the spacing between them into consideration.

4.2.8 CrossHybridizationMultiplier

The CrossHybridizationMultiplier is used for only identical (*_s*) and mixed (*_x*) probe sets during probe selection. This should **not** be confused with the cross-hybridizing penalty applied to the *_x* probe sets as part of the ProbeSetScore calculation.

During probe selection, there are additional cross-hybridization penalties assigned to different probes. For example, when evaluating two potential identical *_s* sets for the same sequence, we want to favor the set with probes cross-hybridizing to the fewest number of sequences. Therefore, probes cross-hybridizing with two sequences will be penalized more than probes that cross-hybridize with just one sequence, and the probes that cross-hybridize with only one sequence will be selected to make up the probe set (given all other qualities being equal).

However, once we complete probe selection and move onto probe set selection (the design proposal), we would not want these additional cross-hybridization penalties to interfere with the probe set score, since these do not really affect the intrinsic quality of how well the probes will hybridize to the target. Therefore, we combine all the different cross-hybridization penalties applied at the time of probe selection to give you a sense of what effect cross-hybridization might have on this particular probe set. By taking the product of the Probe Set Score and this Cross Hybridization Multiplier, you get the fully penalized score that the probe selection software used in making decisions between potential probe sets within the same sequence.

4.2.9 ProbeCount

The number of probes in this probe set.

4.2.10 AvgRawProbeScore

Average of the raw probe scores (without any penalties) in this probe set.

4.2.11 RawStandardDeviation

The standard deviation of the raw probe scores in a probe set.

4.2.12 RepresentedBy

The probe set that is the exemplar for a set of redundant probe sets. When two probe sets are redundant, the probe set with a higher score gets picked as the exemplar. If the probe sets have identical scores, then the one listed first in the Instruction File is picked.

4.2.13 RepresentedByAppliedScore

The score of exemplar probe set (“RepresentedBy” probe set) as applied to the sequence being represented.

4.2.14 NumberCrossHybProbes

The number of cross-hybridizing probes in this probe set.

“0” if the probe set is unique; it is equal to the ProbeCount if the probe set is gene_a or identical_s; it should be anywhere between “1” and the ProbeCount if the probe set is mixed_x.

4.2.15 SifLength

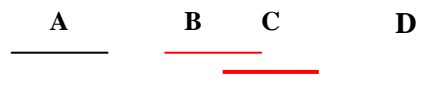
The length of the actual probe target region (SIF region) for this probe set (between the first base of the first probe and the last base of the last probe).

4.2.16 NumberNonoverlappingProbes

The number of probes that do not overlap with each other (≥ 25 bases apart from all other probes).

4.2.17 NumberIndependentProbes

The number of independent probes. The number of probes that are ≥ 25 bases apart from each other. For example,



There are two non-overlapping probes, A and D, but there are three independent probes, A, B or C, and D.

4.2.18 TilingOrder

The sequential order in which the probe sets will be tiled on the chip. It does not influence the physical positioning of probes on the chip, it simply governs the order in which to input these probes into the program until there are too many probes to tile. This order also corresponds to the order in the Instruction File and helps with picking exemplars for identical probe sets in the case where they have identical scores.

4.2.19 Remove

True if this probe set is not to be tiled on the chip, else false. Initially, some redundant probe sets are marked as “TRUE” by default.

4.2.20 ChangeRemoveValue

This column is initially blank. It allows you to give us feedback on whether you would like to remove a probe set that is currently showing “FALSE” in the Remove column, or to add a probe set that is currently marked as “TRUE” in the Remove column.

If you mark a probe set to be removed (TRUE), then when we run Proposal again, it will be removed, and the “Remove” column will become “TRUE” as well.

If you mark a probe set to be added (FALSE), then this probe set will definitely be kept on the next Design Proposal we send you, regardless of whether there might exist other probe sets that could potentially represent this set as well. In addition, we will try to use this probe set as the exemplar probe set, and try to replace other probe sets with this one.

For example, if probe set A_a can replace B_a and C_a, and B_a can only replace probe set C_a. Originally, the Design Proposal shows that both B_a and C_a are removed and are represented by A_a. You decide that you really want to make sure that B_a is tiled on the array, so you mark B_a as “FALSE.” When we send you the new Proposal based on your request, you should see that B_a is kept, C_a is removed and represented by B_a, and A_a is also kept, since nothing else can represent A_a.

4.3 Other Files

At the time of the Design Proposal, there are two other files available at your request, which give you more detailed information on the probe set cross-hybridization. These files are generally quite large, and are useful only if you want to import them into databases and conduct queries.

4.3.1 Proposal Represented File (REP File)

This file shows all of the potential redundant probe set representations, for all iterations. The records in this file have many-to-many relationships.

Column Heading	Definition
OrderProcessed	Integer. The sequential order for the entries in this file as processed.
AppliedScoreThreshold	Double. The applied score threshold used during the “represented by” iteration to determine if the instruction is covered or not.
ProbeSetName	String. The name of the probe set.
Type	Integer. The probe set type, blank = empty, 0 = gene, 1 = unique, 2 = identical, 3 = mixed.
RepresentedBy	String. This is the name of the probe set that is representing this probe set. This is set by the Proposal application.

Column Heading	Definition
RepresentedByAppliedScore	Float. This is the applied score that shows how well the representing probe set covers this instruction.
ChangeRemoveValue	Boolean. This value is set by the customer to determine if a probe set is to be removed (TRUE) or kept and preferred as an exemplar for representing other probe sets (FALSE).
RepresentationAllowed	Boolean. This value is "TRUE" if representation is allowed by the application, or "FALSE" if disallowed. An entry with a "TRUE" value in this column will indicate that the corresponding probe set is removed in the Design Proposal.

4.3.2 Cross Hybridization File (XHY File)

This file shows the cross-hybridization associated with each probe contained within a probe set. It includes ALL of the probe sets produced in probe selection, including those that are later removed due to redundancies in the Design Proposal.

Column Heading	Definition
ProbeSetName	String. The name of the probe set.
Type	Integer. The probe set type.
ExonPosition	Integer. The exon position (mismatch position) of the probe in the sequence.
SequenceName	String. The sequence name with which this probe is cross-hybridizing.
SequencePosition	Integer. The one based position in the cross-hybridizing sequence where the cross-hybridization can be found.
Identical	Boolean. "Y" if the cross-hybridization is identical, else "N." A probe is considered to cross-hybridize with a sequence when it has two non-overlapping eight base matches, a run of at least 12 matching bases, and when the insertion and/or deletion for the two eight base matches between the probe and the sequence is ≤ 5 . A probe will also be considered to cross-hybridize with a sequence if there are only two mismatching bases between the probe and the sequence. An identical match is when all 25 bases are the same.
Score	Integer. A rough count of the number of matching bases in the cross-hybridization. This will normally range from 16 to 25.
Insertion	Integer. The number of bases inserted into the cross-hybridizing sequence.

Column Heading	Definition
Deletion	Integer. The number of bases deleted from the cross-hybridizing sequence.
GeneCluster	String. The gene cluster associated with this cross-hybridization. This comes from the cross-hybridizing sequence.
TranscriptExpressed	Float. The transcript expressed value associated with this cross-hybridization. This comes from the cross-hybridizing sequence.
Distance	Integer. The distance from the probe selection region associated with this cross-hybridization. Zero if in the probe selection region, negative if 3' of the probe selection region, N/A if no probe selection region can be found for the cross-hybridizing sequence.
InProbeSelectionRegion	Boolean. "Y" if the cross-hybridization is in the probe selection region of the cross-hybridizing sequence, else "N."

4.4 ACTIONS REQUIRED

Once you receive the Design Proposal, please review it and decide how you would like to proceed. If you like the Design Proposal the way it is, you can advise us and we will proceed with your mask design. If you would like to add or remove probe sets from your Proposal, please indicate your preference in the "ChangeRemoveValue" column of the Design Proposal. If you would like to change any of the criteria used in generating the Design Proposal, you can communicate that to us as well at this point.

5. CONTACT INFORMATION

5.1 Sales & Field Application Support

Contact Sales for GeneChip® brand product information. All general technical questions not related to a specific design and all support issues should be addressed to Field Application Support.

Address:	USA Affymetrix, Inc. 3380 Central Expressway Santa Clara, CA 95051 USA	Europe Affymetrix UK Ltd Voyager, Mercury Park Wycombe Lane, Wooburn Green High Wycombe HP10 0HH United Kingdom
Phone:	1-888-362-2447 (1-888-DNA-CHIP)	+44 (0) 1628 552550
Fax:	1-408-731-5441	+44 (0) 1628 552585
E-mail:	sales@affymetrix.com support@affymetrix.com	saleseurope@affymetrix.com supporteurope@affymetrix.com
Address:	Japan Affymetrix Japan, K.K. Mita NN Bldg., 16 F 4-1-23 Shiba, Minato-ku, Tokyo 108-0014 Japan	
Phone:	+81-(0)3-5730-8200	
Fax:	+81-(0)3-5730-8201	
E-mail:	salesjapan@affymetrix.com supportjapan@affymetrix.com	

5.2 Customer Service

For prompt attention to your purchase orders, please address all purchase orders to Customer Service. If you send your order to a specific salesperson, it may be delayed, as most of our salespeople travel extensively.

Address:	USA Affymetrix, Inc. 3380 Central Expressway Santa Clara, CA 95051 USA Attn: Customer Service	Europe Affymetrix UK Ltd Voyager, Mercury Park Wycombe Lane, Wooburn Green High Wycombe HP10 0HH United Kingdom
Phone:	1-800-499-7309	+44 (0) 1628 552550
Fax:	1-408-481-9442	+44 (0) 1628 552559
E-mail	customerservice@affymetrix.com	cse@affymetrix.com

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

5.3 Chip Design

All design information should be sent to the Chip Design team.

Address:	Affymetrix, Inc. 3380 Central Expressway Santa Clara, CA 95051 Attn: Chip Design
Phone:	1-408-731-5669 (Voice Mail)
Fax:	1-408-731-5380
E-mail:	chip_design@affymetrix.com

6. GLOSSARY

_at	Probe set name extension that signifies an antisense target-detecting probe set.
_st	Probe set name extension that signifies a sense target-detecting probe set.
3' ESTs	Expressed Sequence Tag, often representing the most downstream portion (3 prime end) of a transcript most often discovered by “end sequencing” of cDNA clones.
3' (labeling) bias	Referring to the standard target preparation process used by Affymetrix for the T7-Oligo(dT) primed synthesis of cDNA and amplified by <i>in vitro</i> transcription. The end result is a labeled cRNA that is biased towards the 3' end of the transcript.
alias	The value in the “alias” column specified in the Instruction File is the designation used to form the probe set name. This “alias” value is carried into the Design Proposal.
ambiguity	Uncertain base identity in a DNA sequence.
annotated and supported sequences	Gene sequence found by gene prediction or by actual cDNA sequencing whose accuracy is supported by different means (e.g., alignment of predicted gene with additional EST evidence).
annotated sequence	Type of sequence (e.g., CDS, EST, mRNA, or predicted).
antisense design	A design that generates probes hybridizing to antisense targets.
antisense strand	Reverse complement of sense strand (see “sense strand”).
applied probe set score	Given that a probe set may represent two or more overlapping gene sequences, an applied score provides a prediction of the quality of one probe set representing another sequence.
array controls	Types of array controls include Standard Controls, Target Preparation and Hybridization Controls, Species-Specific Controls, Contaminant Controls, and Scaling and Normalization Controls.
array description	Referring to the Design Request Form, a description for your design. It can be up to 255 characters.

array format	Referring to the size of the probe array. Affymetrix has various sizes of probe arrays.
array name	The official name of the array. For catalog designs, this refers to the short array name, such as “HG-U133A.” For custom designs, this refers to the official array name assigned to your design, which is in the format of <customized name><1-letter strandedness><part number>, with “F” at the end if it is a CustomExpress™ Advantage Array design.
AvgRawProbeScore	Referring to the Design Proposal, the average of the raw probe scores (without any penalties) in this probe set.
CDS	Abbreviation for the amino acid coding sequence of the gene.
center cross control	Affymetrix control probes that hybridize to oligo B2 and illuminate a “cross” pattern in the central area of the array.
ChangeRemoveValue	Referring to the Design Proposal, this categorical information in the file is initially blank. It allows you to give us feedback on whether you would like to remove a probe set that is currently showing “FALSE” in the Remove column, or to add a probe set that is currently marked as “TRUE” in the Remove column.
Chip Design	The Affymetrix group responsible for the implementation of a conceived probe array design, e-mail chip_design@affymetrix.com .
classification columns	An additional category of up to 10 columns of information in the Instruction File to provide extra useful information about the sequence that can be linked specifically by Affymetrix® software in the expression output files. These columns are linked to probe set names in Affymetrix software (e.g., Data Mining Tool).
clustering	Process of grouping and aligning similar sequences often to create a virtual gene sequence by consensus base calling algorithms.
commercial content	Probe array content selected by the customer from probe sets currently on Affymetrix catalog or “Made-to-Order” expression products.
consensus	Refers to the end result or “consensus” sequence from base calling algorithms of aligned, clustered sequences.

corner grid alignment controls	Affymetrix control probes that hybridize to oligo B2 in a small checkerboard pattern in the corners of the array used for the alignment of the “grid” used to determine the boundaries of each probe cell.
cross-hybridization penalty	Penalty applied to <code>_x</code> and <code>_s</code> probes to enable the creation of probe sets that potentially cross-hybridized to the fewest number of sequences possible.
CrossHybridizationMultiplier	Referring to the Design Proposal, the CrossHybridizationMultiplier is used for only identical (<code>_s</code>) and mixed (<code>_x</code>) probe sets during probe selection. This should not be confused with the cross-hybridizing penalty applied to the <code>_x</code> probe sets as part of the ProbeSetScore calculation.
CustomExpress™ Advantage Arrays	Cost-effective probe arrays, with faster manufacturing turn-around time, developed from pre-existing commercial and/or custom probe set content or from a set of sequences provided by the customer.
Customer Service	The Affymetrix group responsible for processing purchase orders and assisting you with your account information.
database	Referring to the Instruction File, the public or private database from which sequences were derived.
description	Referring to the Instruction File, description provided for the sequence which will be used as annotations for the probe set(s) produced from that sequence.
Design Proposal	Output file from the Chip Design group that contains probe selection results and metrics set to enable the selection of optimal probe sets for the CustomExpress Array.
Design Request Form	Part of the required design request package to be sent to Affymetrix in order to begin your array design process. This document is in Word format and provides us with contact information, array logistics, and special instructions.
end	Referring to the Instruction File, within the sequence, the numeric (1-based) position of the end of the region of interest. This column is required if “start” is specified.

endSeq	Referring to the Instruction File, a sequence used for checking purposes. It includes the base at “end” and the previous few bases up to a total of eight bases. This column is required if “end” is specified.
exemplar	A representative cDNA sequence for each gene. The exemplar approach method usually involves some initial clustering into gene groups and the subsequent selection of a representative from each gene group.
FAS	Abbreviation for Affymetrix support representative known as the Field Application Specialist. All general technical questions not related to a specific design and all support issues should be addressed to your Field Application Specialist.
FASTA format	A standard format used to create a file containing sequences. FASTA format begins with a single-line description, followed by one or more lines of sequence. The description line is distinguished from the sequence data by a greater-than symbol (“>”) that occupies the first character-space and is usually followed by the sequence identifier or name.
Feature	Also known as probe cell on the probe array defined by having identical probes synthesized jointly in the one unit area (e.g., there are over 500,000 features on the HG-U133A Array).
gap multiplier	Referring to the Design Proposal, this is the penalty applied to the raw probe quality score used to calculate the probe set score.
gap penalty	A multiplier applied to probes if they are too close to each other and may not perform independently.
geneCluster	The name provided by you that represents the gene the sequence belongs to. If multiple sequences are classified to be from the same gene cluster, we try to make a probe set that best represents ALL of the sequences within the same cluster (i.e., _a set). However, an _a set will be created as long as a probe set can be found to represent SOME of the members of the same gene cluster exclusively.
hyperlink	Referring to the Instruction File, used to link to public databases from the Affymetrix® software.

Instruction File	Input file provided by you which provides information to tell us what, how, and where each probe set should represent the sequence. Instruction File also contains annotations associated with each sequence and subsequent probe set.
Latin Square experiments	An experimental design used to monitor the detectability of a transcript accurately over a range of concentrations. It also allows the statistical analysis of patterns and variability in repeated measurements in a systematic fashion.
Library Files	Files used by Affymetrix® software to analyze specific probe arrays based upon the specific content and design of that probe array.
Made-to-Order Array Program	The GeneChip® Made-to-Order Array Program gives you the flexibility to utilize arrays from selected new designs and previous-generation GeneChip expression arrays not available as catalog products. Content from any Made-to-Order Array Design can be utilized for your custom array.
mask	Refers to the light-directing filter in the photolithography process used to manufacture the probe array.
mismatch (MM)	A 25-mer oligonucleotide designed to be complementary to a reference sequence except for a single, homomeric base change at the thirteenth position. Mismatch probes serve as specificity controls when compared to their corresponding Perfect Match probe.
multi-species design	A design containing sequences from more than one species. This information is provided in the Design Request Form.
non-specific cross-hybridization	Unintended annealing of imperfectly matched, or non-target, sequence to the probe sequence.
Ns	Ambiguous bases. Any characters in the Sequence File that are not A, C, G, or T will be replaced by N. Probes that contain ambiguity are not selected.
NumberCrossHybProbes	Referring to the Design Proposal, the number of cross-hybridizing probes in this probe set: “0” if the probe set is unique; it is equal to the ProbeCount if the probe set is gene_a or identical_s; it should be anywhere between “1” and the ProbeCount if the probe set is mixed_x.




NumberIndependentProbes	Referring to the Design Proposal, the number of probes that are greater than 25 bases apart from each other.
NumberNonoverlappingProbes	Referring to the Design Proposal, the number of probes that do not overlap with each other (≥ 25 apart from all other probes).
perfect match (PM)	A single stranded DNA, 25-mer oligonucleotide that is complementary to the target. For antisense-target detecting arrays, the Perfect Match oligo is identical to the designated forward strand of the gene sequence it represents. For sense-target detecting arrays, the Perfect Match oligo is complementary to the designated forward strand of the gene sequence it represents.
probe	Any oligo sequence synthesized on the array used to detect a complementary target in solution. If referring to the Instruction File, this is the number of probe pairs per probe set.
probe array format	Refers to the size of the probe array. Affymetrix has standard size probe arrays depending on the linear dimension (e.g., Standard probe array is also known as the 49 format and contains a probe synthesis area with a one-sided dimension of 12800 μm).
probe length	The length in bases of any oligonucleotide synthesized on the probe array. The standard length is 25-mer.
probe pairs	Refers to the fundamental detection unit of any Affymetrix probe set consisting of a perfect match and corresponding mismatch oligo.
probe quality metric	Same as raw probe quality score.
probe selection	A process by which optimally hybridizing probes are identified.
probe selection region	The region of the expressed sequence from which probes are selected. Typically, this is the last 600 bases from the 3' end for eukaryotic designs.
probe set	A set of probe pairs selected to represent expressed sequences on an array.
probe set redundancy	Used for removal of extra probe sets. Redundancy is defined as the ability of a probe set from one sequence to represent other sequence(s).

Probe Set Request File	A tab-delimited text file that contains the list of existing probe sets from previous commercial or custom designs that the customer requests for the new design.
probe set score	A prediction of how well a probe set will respond in an experiment. It is derived from three components: the probe quality metric, the cross-hybridization penalty, and the gap penalty.
probe set type	Probe sets are identified by one of four types: unique (no extension), identical gene cluster (_a), identical non-gene cluster (_s), or mixed (_x). Refer to Section 2.6.2.
ProbeCount	Referring to the Design Proposal, the number of probes in this probe set.
ProbeSetCount	Referring to the Design Proposal, the number of probe sets that generated this sequence based on the Instruction File.
probe set name	Derived from the “alias” value provided in the Instruction File. It is formed in this fashion: <alias><suffix, such as “_s” etc.><target type, such as “_at” etc.>
	In the Probe Set Request File, the probe set name must be provided, as well as the name of the array from which it is coming.
proprietary	Sequences that are not in the public domain.
prune	Process by which uniquely hybridizing probes are identified.
Pruning Affymetrix Standard Library	Standard set of FASTA-formatted Sequence Files that Affymetrix uses to determine probe sequence specificity. Available for some organisms; please inquire.
Pruning Sequence File	A file that contains sequences used for pruning. It needs to be in FASTA format, just like the Sequence File.
QC oligo controls	Used by Affymetrix during manufacturing for quality control.
raw probe quality score	A metric that was developed based on Latin Square experiments. It is defined as the slope of the line that relates to natural logarithms of intensities and of target concentrations for each probe.

RawStandardDeviation	Referring to the Design Proposal, the standard deviation of the raw probe scores in a probe set.
Remove	Referring to the Design Proposal, “TRUE” if this probe set is not to be tiled on the chip, else “FALSE”. Initially, some redundant probe sets are marked as “TRUE” by default.
Rename	Referring to the Probe Set Request File, this column is only required if a probe set name is duplicated in a design.
RepresentedBy	Referring to the Design Proposal, the probe set that is the exemplar for a set of redundant probe sets.
RepresentedByAppliedScore	Referring to the Design Proposal, the score of exemplar probe set (“RepresentedBy” probe set) as applied to the sequence being represented.
scaling	Used in analyzing expression profiles from two or more microarray experiments. Scaling is used to quantify changes in expression, thereby minimizing systematic variations from accountable and predictable sources that cause differences in absolute expression measurements (i.e., variations in labeling efficiency, quantification of labeled target, hybridization, washing, staining, and chip performance).
sense design	A design that generates probes hybridizing to sense strand targets.
sense strand	Represents forward strand of the gene sequence 5' to 3'.
Sequence File	Input chip design file containing FASTA-formatted sequences selected for representation on the probe array.
sequence name	The name of a sequence in the Sequence File. This same sequence name must also appear in the Instruction File under the “name” column. It is carried over to the Design Proposal.
sequence orientation	Referring to the sense (forward) or antisense (reverse) strand in the double-stranded gene sequence.
sequence selection	The process of compiling a set of sequences by filtering or modification such that optimal sequences for the gene expression assay are obtained.

SIF sequence	After probe selection, the region of the transcript between the most 5' and 3' probes. Also referred to as the Target Sequence in the NetAffx™ viewer.
SifLength	Referring to the Design Proposal, the length of the actual probe target region (SIF region) for this probe set (between the first base of the first probe and the last base of the last probe).
space normalization controls	Refers to control probes that could be used for spatial reference and for monitoring uniform hybridization.
start	Referring to the Instruction File, within the sequence, the numeric (1-based) position of the start of the region of interest. This column is required if “end” is specified.
startSeq	Referring to the Instruction File, a sequence used for checking purposes. It includes the base at “start” and the next few bases up to a total of eight bases. This column is required if “start” is specified.
target type	Referring to the orientation of the labeled nucleic acid that will be hybridized to the array, Eukaryotic Antisense, Prokaryotic Antisense, and Prokaryotic Sense. The default selection is Eukaryotic Antisense.
tile	To be synthesized on the probe array.
transcriptExpressed	Referring to the Instruction File, abundance level of the transcript from low to high denoted with a value between “0” and “1,” respectively.
trimmed	Removal of poorer quality sequence from ends of sequences prior to probe selection.
Type	Referring to the Design Proposal, flags representing the type of probe set: unique, identical, gene family, mixed. For probe set types other than unique, special suffixes are added to the “alias” value to form part of the probe set name: “_a” for gene, “_s” for identical, and “_x” for mixed.
unique content	Set of sequences provided by customer for which new probes will be selected.
unique probe	Probes that uniquely query a given target sequence.
UTR	Untranslated region.

Trademarks

Affymetrix®, GeneChip®, , , , HuSNP®, GenFlex®, EASI™, MicroDB™, Flying Objective™, NetAffx™, CustomExpress™, CustomSeq™, ‘Tools To Take You As Far As Your Vision™’, and ‘The Way Ahead™’ are trademarks owned or used by Affymetrix, Inc.

LIMITED LICENSE

EXCEPT AS EXPRESSLY SET FORTH HEREIN, NO RIGHT TO COPY, MODIFY, DISTRIBUTE, MAKE DERIVATIVE WORKS OF, PUBLICLY DISPLAY, MAKE, HAVE MADE, OFFER TO SELL, SELL, USE, OR IMPORT PROBE ARRAYS OR ANY OTHER PRODUCT IS CONVEYED OR IMPLIED WITH THE PROBE ARRAYS, INSTRUMENTS, SOFTWARE, REAGENTS, OR ANY OTHER ITEMS PROVIDED HEREUNDER. EXCEPT FOR CERTAIN ARRAYS AND REAGENTS DESIGNATED AS “ANALYTE SPECIFIC REAGENTS” (SEE APPLICABLE PACKAGE INSERT) WHICH ARE LICENSED FOR USE AS ANALYTE SPECIFIC REAGENTS OR RESEARCH USE, ALL PRODUCTS (INCLUDING THE PROBE ARRAYS, INSTRUMENTS, SOFTWARE, AND REAGENTS) DELIVERED HEREUNDER ARE LICENSED TO BUYER FOR RESEARCH USE ONLY. THIS LIMITED LICENSE PERMITS ONLY THE USE BY BUYER OF THE PARTICULAR PRODUCT(S), IN ACCORDANCE WITH THE WRITTEN INSTRUCTIONS PROVIDED THEREWITH, THAT BUYER PURCHASES FROM AFFYMETRIX (“AFX”) OR ITS AUTHORIZED REPRESENTATIVE. THE PURCHASE OF ANY PRODUCT(S) DOES NOT BY ITSELF CONVEY OR IMPLY THE RIGHT TO USE SUCH PRODUCT(S) IN COMBINATION WITH ANY OTHER PRODUCT(S). IN PARTICULAR, (I) NO RIGHT TO MAKE, HAVE MADE, OR DISTRIBUTE OTHER PROBE ARRAYS IS CONVEYED OR IMPLIED BY THE PROBE ARRAYS, (II) NO RIGHT TO MAKE, HAVE MADE, IMPORT, DISTRIBUTE, OR USE PROBE ARRAYS IS CONVEYED OR IMPLIED BY THE INSTRUMENTS OR SOFTWARE, AND (III) NO RIGHT TO USE PROBE ARRAYS IN COMBINATION WITH INSTRUMENTS OR SOFTWARE IS CONVEYED UNLESS ALL COMPONENT PARTS HAVE BEEN PURCHASED FROM AFX OR ITS AUTHORIZED REPRESENTATIVE. FURTHERMORE, PROBE ARRAYS DELIVERED HEREUNDER ARE LICENSED FOR ONE (1) TIME USE ONLY AND MAY NOT BE REUSED. THE PRODUCTS DO NOT HAVE FDA APPROVAL. NO PATENT LICENSE IS CONVEYED TO BUYER TO USE, AND BUYER AGREES NOT TO USE, THE PRODUCTS IN ANY SETTING REQUIRING FDA OR SIMILAR REGULATORY APPROVAL, OR EXPLOIT THE PRODUCTS IN ANY MANNER NOT EXPRESSLY AUTHORIZED IN WRITING BY AFX IN ADVANCE.

PATENTS

Products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,744,305; 6,261,776; 6,291,183; 5,700,637; 5,945,334; 6,346,413, and 6,399,365; and EP 619 321; 373 203 and other U.S. or foreign patents.

COPYRIGHT

©2002 - 2003 Affymetrix, Inc. All rights reserved.

AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 U.S.A.
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd



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

AFFYMETRIX JAPAN K.K.

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

www.affymetrix.com

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

Part No. 700506 Rev. 4
©2002-2003 Affymetrix, Inc. All rights reserved. Affymetrix®, GeneChip®, , , , , HuSNP®, EASIT™, MicroDB™, GenFlex®, Flying Objective™, CustomExpress™, CustomSeq™, NetAffx™, Tools To Take You As Far As Your Vision™, and The Way Ahead™ are trademarks owned or used by Affymetrix, Inc. Products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,744,305; 6,261,776; 6,291,183; 5,700,637; 5,945,334; 6,346,413; and 6,399,365; and EP 619 321; 373 203 and other U.S. or foreign patents.

