Axiom myDesign Custom Array design guide for human genotyping applications

Overview

In the past, custom genotyping arrays were expensive, required large sample commitments, and took a lot of time to design and deliver. Investigators are now demanding more custom flexibility.

Applied Biosystems™ Axiom™ myDesign™ Custom Genotyping Arrays overcome these challenges by offering a fast, cost-effective way to modify an existing array or create an entirely new design for as few as 480 samples.

Adding, replacing, and removing markers on existing arrays is a simple process. This flexibility allows you to optimize your study to achieve more power, whatever your population, trait, or application. It also enables you to update your array as new genomic information becomes available.

Our bioinformatics specialists work directly with you during our streamlined design process, as shown in Figure 1. They guide you through the process, verify your SNP sequences, and provide design scores, reports, and advice to help ensure your final array meets your study objectives. Your new custom arrays will typically be delivered in less than 6 weeks after the final design review.

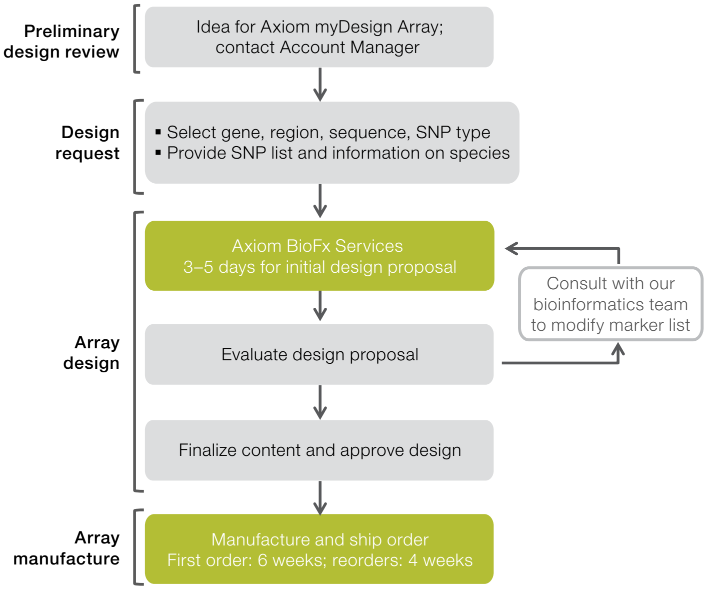

Preliminary design review
Idea for Axiom myDesign Array; contact Account Manager
Design request
▪ Select gene, region, sequence, SNP type
▪ Provide SNP list and information on species
Array design
Axiom BioFx Services
3–5 days for initial design proposal
Consult with our bioinformatics team to modify marker list
Evaluate design proposal
Finalize content and approve design
Array manufacture
Manufacture and ship order
First order: 6 weeks; reorders: 4 weeks

Figure 1. The Axiom myDesign Custom Array design process.

Sources of genomic markers for Axiom myDesign Arrays

Variants for a myDesign Array may be selected from several sources including, but not limited to, dbSNP, the Applied Biosystems™ Axiom™ Genomic Database, or markers identified by sequencing.

The Axiom Genomic Database contains 9.4 million wet lab–tested markers that have been validated via genotyping 270 samples used in the HapMap project to ascertain cluster resolution characteristics (Figure 2). The markers in the Axiom Database directly or through pairwise tagging ($r^2 \geq 0.8$) cover 25.8 million of the 39.5 million 1000 Genomes Project (1KG) variants in at least one of the 14 populations studied by the 1000 Genomes Project. These 9.4 million markers have been collected from external discovery sources, including pilot phases of the 1000 Genomes Project, dbSNP, HapMap, and various other screening projects conducted by Thermo Fisher Scientific. Figure 2 is a Venn diagram detailing the content of the Axiom Genomic Database compared to the 1000 Genomes Project, March 2012 release. For markers not found in the Axiom Genomic Database, novel probe design and performance prediction algorithms using multiparametric analysis of SNP and indel probe properties will be used.

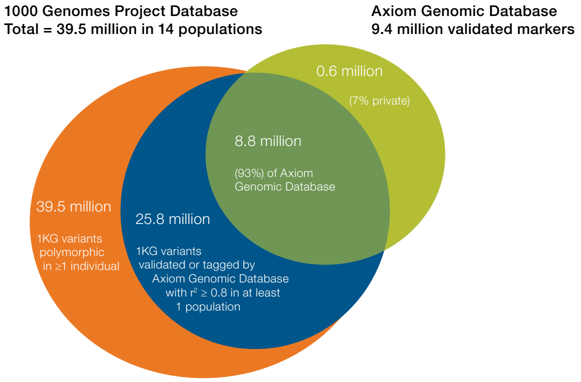

1000 Genomes Project Database
Total = 39.5 million in 14 populations
Axiom Genomic Database
9.4 million validated markers
0.6 million
(7% private)
8.8 million
(93%) of Axiom Genomic Database
39.5 million
1KG variants polymorphic in ≥1 individual
25.8 million
1KG variants validated or tagged by Axiom Genomic Database with r² ≥ 0.8 in at least 1 population

Figure 2. Graphic representation of the Axiom Genomic Database.

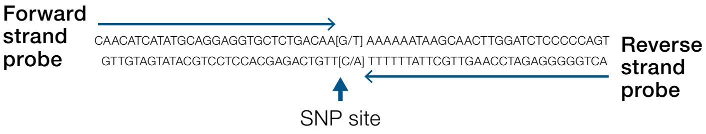
Determining the number of markers on your array

The number of markers (SNPs and indels) on your array is determined by the number of features on the array and the number of features required per marker. Each Axiom microarray contains up to 1.39 million individual features, which refer to locations on the array with a distinct 30-mer probe sequence. This 30-mer probe sequence is also referred to as a probe. Specifically:

$$\frac{\text{features}}{\text{marker}} = \frac{\text{strands}}{\text{marker}} \times \frac{\text{probes}}{\text{strand}} \times \frac{\text{features}}{\text{probe}}$$

Number of strands per marker

Markers are interrogated by probe sequences that are complementary to either the forward and reverse strand sequence flanking the marker site, as shown in the example below.


Forward strand probe
CAACATCATATGCAGGAGGTGCTCTGACAA[G/T] AAAAAATAAGCAACTTGGATCTCCCCCAGT
GTTGTAGTATACGTCCTCCACGAGACTGTT[C/A] TTTTTTATTCGTTGAACCTAGAGGGGGTCA
Reverse strand probe
SNP site

Markers selected from the Axiom Genomic Database are interrogated with probes (1 or 2, as discussed in “Number of probes per strand”, below) for just 1 strand, usually the strand that gave the best empirical performance. Markers not found in the Axiom Genomic Database are typically interrogated by both forward and reverse strand probe sequences to increase the probability of obtaining accurate genotypes. Markers not previously validated for Axiom microarrays therefore require roughly twice as much space on the array as markers selected from the Axiom Genomic Database.

Number of probes per strand

SNPs with alleles A/C, A/G, C/T, and G/T require just 1 probe per strand, and the 30-mer sequence of the probe terminates on the 5´ side of SNP site (shown in the above figure). A/T and C/G SNPs require 2 distinct probe sequences in order to discriminate the alleles. The 30-mer sequence of each distinct probe terminates with one of the bases of the A/T or C/G SNP, thus requiring twice as many features. The set of probes (1 or 2) that are complementary to a given strand and used to interrogate the same marker are referred to as probe sets. Probe sets for A/T and C/G SNPs are sometimes referred to as allele-specific. Indels frequently require only a single probe per strand, but some require 2 allele-specific probes. About 16% of all markers require allele-specific probe sets.

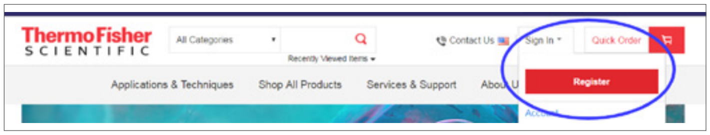
Number of features per probe

The tandard practice for placing a probe on an array is to replicate the probe at 2 feature locations on the array. This is referred to as 2-rep tiling. Relatively few probes require more than 2 replicates for good resolution. In addition, we have empirically identified markers that can be interrogated with probes that only require 1 feature (1-rep tiling) for good genotyping performance. These markers were prioritized in the creation of “imputation-based GWAS grids”, which provide high coverage of genetic variation in a population with a minimum number of features. Such grids can be added as modules to a custom design.

Submitting a SNP list

Necessary information for SNP submission is defined below and should be emailed to **Bioinformatics Services**.

To facilitate the exchange of files containing SNP submission information (discussed below), we can provide access to a directory on our **secure file exchange server**. However, to provide this access we require that you have previously registered on **thermofisher.com** with your email address. All files exchanged for the purposes of array design are kept confidential.



A. Requestor information

Complete requestor information in the first tab of the accompanying Microsoft™ Excel™ file (SNP template for Axiom myDesign Custom Arrays) as detailed below. All fields marked with an asterisk (*) are mandatory.

- *Customer name
- *Company name
- *Company address
- *City
- State/province
- *Postal code
- *Country
- *Email: The email address of customer/PI who will be contacted by our bioinformatics team if we encounter difficulties in the SNP selection process
- *Phone number

Axiom Arrays are sometimes run and analyzed at an offsite location specified by the customer. These services require library files specifically created for each myDesign Array. Please provide contact information if you would like us to provide these files to an offsite service provider or lab. All fields marked with an asterisk (*) are mandatory.

- *Contact name
- *Company name
- *Company address
- *City
- State/province
- *Postal code
- *Country
- *Email: The email address of customer/PI who will be contacted by our bioinformatics team if we encounter difficulties in the SNP selection process
- *Phone number

B. Array information

The second tab in the Excel file is reserved for providing information on the Axiom myDesign Array as follows. All fields marked with an asterisk (*) are mandatory.

- Array name (maximum of 8 alphanumeric characters); e.g., humanSNP
- Array description (maximum of 24 characters); e.g., HumanDisease SNPproject
- Population of interest; e.g., CHB, CEU, YRI, etc.
- Number of SNPs being submitted; e.g., 650,000
- NetAffx™ IDs of authorized project users (maximum of 5 NetAffx IDs per project)
- Will you allow us to mention the array in our marketing literature without revealing specific information about your institution or company: Yes/No

C. Marker specification

Direct tiling: The specified markers will be included on the array. Markers that have not been validated on the Axiom platform will be represented by novel probe sets, if requested. Note that novel markers generally require twice as much space as validated markers on the array.

Best tag: For markers from the 1000 Genomes Project, we have calculated the pairwise r^2 between each marker and a validated "tag" marker in the Axiom Genomic Database. Using this method, target markers that are not in the database will be represented by a validated tag marker if r^2 exceeds the specified threshold in the relevant population. Otherwise, novel probe sets will be used, if requested.

Greedy tag selection: A single, validated marker can often cover multiple target markers. Using this method, we will algorithmically select an efficient set of validated markers to cover as many target markers as possible. This method can be applied to specific sets of markers or regions, or the entire genome. The population and minor allele frequency range of interest must be specified.

Physical density: Using this method, we will algorithmically select validated markers in the target region or regions of interest, with the aim of providing a uniform density of markers along the genome. The population, MAF range, and desired density must be specified.

Target markers may be specified in several ways. Only biallelic markers are allowed; the Axiom platform does not currently permit triallelic or multiallelic markers.

Range of positions in hg19/GRCh37 coordinates: Please specify the chromosome and start/stop positions for each interval to include. Also specify whether to include only Axiom-validated markers or all markers in the 1000 Genomes Project, along with the population and MAF range of interest. Additional selection rules may be provided, such as "coding, nonsynonymous markers only."

HUGO gene symbols: Along with the gene symbols, please specify the population, MAF range, and marker types of interest, as for ranges of positions, above.

- Please note that genes should be specified using official, HUGO gene symbols. Synonyms or nonstandard gene names can result in errors or require clarification. If in doubt, a range of coordinates may be a better choice.
- When selecting markers associated with genes, we consider all markers falling within any transcript for the gene. If this is not appropriate (e.g., if additional flanking regions are desired), please specify the rule to use or the specific coordinates in hg19.

Specifying individual markers: Lists of individual markers may be specified in several ways, listed below in order of preference. Please be sure to specify only biallelic markers and to avoid duplication of any identifiers (e.g., rsIDs) or sequences. If using multiple formats (e.g., rsIDs for some markers; chromosome, position, and alleles for others), please provide only one format per file, and avoid duplicating markers in multiple files. File formats are shown in section E.

- Affy SNP IDs from our validated database: These can be obtained from annotation files for existing arrays or in the course of exploring possible content with our array bioinformatics specialists.
- hg19/GRCh37 chromosome, position, and alleles, and dbSNP rsID if available:
  VCF format (**http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41**) is preferred and is essential for any non-SNP marker (e.g., indels). Please note:
  - – If an rsID is provided that does not fully agree with the position or alleles given, we will flag this to your attention and proceed based on the coordinates and alleles given. This includes cases in which 2 alleles are specified, but the dbSNP record includes multiple alleles.
  - – In a small number of cases, we do not agree with dbSNP's positioning of the marker relative to the hg19 reference genome. In these cases, we will design probe sets using the position and alleles you specify. We will flag these cases to let you know that we are operating from your coordinates rather than the rsID.
  - – It is possible to design probes to interrogate 2 alternate alleles, rather than the reference and a single alternate. However, in order to reduce the possibility of errors, we will flag these entries to your attention and request clarification.
  - – It may be necessary to provide context sequence (see below).
- dbSNP rsIDs only, preferably from the most recent dbSNP build: Please note:
  - – rsIDs do not need to be accompanied by coordinates and alleles. However, if rsIDs are given without coordinates and alleles, we may be unable to design probes for some of them due to ambiguity in coordinates, presence of more than 2 alleles in dbSNP, or disagreement with dbSNP's positioning (see above). These cases will be flagged to your attention for clarification.
  - – It may be necessary to provide context sequence (see below).

Context sequence: If necessary, probes can be designed that don't match the hg19 reference genome (e.g., if it's highly desirable to match the dbSNP consensus flanking sequence rather than the reference genomic sequence). This requires chromosome, position, alleles, and optionally rsID (ideally in VCF format), as well as "context sequence" on the forward strand for 35 bp upstream and downstream of the variant.

- Context sequence for a SNP submission: The context SNP sequence is a 71-mer nucleotide sequence from the "+" or "forward" strand centered on the target SNP and written in the 5´ to 3´ direction. The alleles of the target SNP are indicated in brackets at the 36th position. Note that sequence is not required if the human reference genome (hg19/GRCh37) is to be used; in that case, the position and alleles are sufficient. Please note:
  - – The context sequence and alleles should be given for the forward strand even if, for example, dbSNP places the SNP on the reverse strand. Alleles will be reported in terms of the forward strand in the annotation files provided with the completed array design.
  - – The SNP alleles should be in alphabetical order (e.g., [A/C], not [C/A]).
  - – We do not disambiguate ambiguity codes, so any bases other than "A", "C", "G", and "T" will be treated as nonbases, and we will not tile probes that overlap those ambiguous positions. If a context sequence contains ambiguities on only 1 side of an SNP site, a probe set can be included for the other side (strand).
  - – If sequence for only 1 flank (forward or reverse) is available and the SNP is either an [A/T] or [C/G] SNP, then the sequence must include at least the first base on the far side (opposite flank) of the SNP.

Example of a context sequence for a SNP submission:

GGGGTCATAGTCGTTCCTCCAGGGCTCACAGACTT[**A/C**]GACTCAATACGTTTGGCGCAAACTCGGACCAGTTT

- Context sequence for an indel submission: In most cases the Axiom assay can also interrogate insertion and deletion (indel) polymorphisms. The context sequence for an indel is 35 bases of context on either side of the marker, plus the alleles of any length from the "+" or "forward" strand centered on the target indel and written in the 5´ to 3´ direction. The alleles of the target indel are indicated in brackets at the 36th position. Please note:
  - – Indels should be submitted in the same format as SNPs (i.e., with the alleles of the target indel indicated in brackets at the 36th position). Deletion alleles are indicated with a dash, which should be the first allele listed (e.g., [-/ACA], not [ACA/-]). Please note that indel alleles are written differently in VCF format and in the context sequence. If both types of information are provided, please check that each is specified correctly. Some indels will be automatically rejected if probes cannot be designed. There is no restriction on the length of the alleles. Examples of indel submissions are shown below.
  - – If sequence for only 1 flank (forward or reverse) is available, the sequence must include at least 5 bases on the far side of the indel.
  - – As for SNPs, ambiguous bases will prevent probe design. In order to permit the design of a probe set for an indel complementary to 1 side of the variant site, the 35 bases on that side and the 5 bases adjacent to the variant site on the other side must all be unambiguous.

Probe set distinguishes reference from inserted A (note this could equally represent the deletion of an A relative to the reference genome; this ambiguity is why the reference and alternate alleles must also be specified).

GGGGTCATAGTCGTTCCTCCAGGGCTCACAGACTT[**-/A**]GACTCAATACGTTTGGCGCAAACTCGGACCAGTTT

Probe set distinguishes multinucleotide polymorphism alleles.

GGGGTCATAGTCGTTCCTCCAGGGCTCACAGACTT[**A/CAG**]GACTCAATACGTTTGGCGCAAACTCGGACCAGTTT

For some indels, such as those inside of repetitive elements, it is not possible to design probes. Such indels will be automatically rejected during the design process.

GGGGTCATAGTCGTTCCTCCAGGGCTCACCAGCAG[**-/CAG**]CAGCAGTACGTTTGGCGCAAACTCGGACCAGTTT

D. Considerations for choosing markers to maximize conversion rate and space on the array

Several factors influence the fraction of markers that will perform well on an Axiom Array ("convert").

- Markers that have been wet-lab validated on the Axiom platform have the highest rate of success (>98% high performance) on any new Axiom Array.

- Novel markers typically require twice as much space on the array as Axiom-validated markers and typically have an 80–85% “conversion” rate. It may be advisable to deprioritize novel markers in favor of validated, “tagging” markers in strong linkage disequilibrium (LD) to the target marker, when available. Information on available tags will be provided upon request.
- A/T and G/C SNPs require twice as much space on the array as other SNPs; deprioritizing these will save space on the array for additional markers.
- When selecting novel markers for inclusion on an Axiom Array:
  - SNPs observed on more than 1 NGS platform or in multiple sequencing projects should be prioritized.
  - SNPs observed at greater sequencing depth should be prioritized.
  - If quality scores are available from an NGS platform, SNPs observed at a higher quality should be prioritized.
  - SNPs that have been validated on another genotyping platform are more likely to convert and should be prioritized.
  - SNP sites without other polymorphisms within 30 bp should be prioritized, and we recommend that the target SNP should not have any other SNP, indel, or translocation within 20 bp.
    - Even if a known polymorphism exists in the flanking sequence, please submit unambiguous bases for the sequence corresponding to the reference state/genome.
  - SNP sites within repetitive element regions should be deprioritized.
  - SNP sites with flanking sequence that has lower similarity to other sites in the genome should be prioritized.

E. Marker context information:

No.	Column header	Contents
1	Affy SNP ID	Internal identifier—please provide if known; otherwise, ignore.
2	Chromosome	1–22, X, Y, MT Specify mitochondrial build. Typically one of: NC_012920 (GRCh37) NC_001807 (hg19)
3	Position	Either dbSNP-style or VCF coordinates (these differ for some indels and MNPs). Please use one consistent system for all markers.
4	rsID	If available; most recent dbSNP build preferred.
5	Reference Allele	In terms of forward-strand sequence. For example, if the marker is from dbSNP, please ensure that alleles for reverse-strand markers are reverse-complemented to match the forward strand. If not using VCF notation, please use “-” to indicate a deletion.
6	Alternate Allele	In terms of forward-strand sequence. If not using VCF format, please use “-” to indicate a deletion.
7	“Flank” or Context Sequence	Optional; required only if it is necessary to design probes that do not match the reference genome (hg19/GRCh37). See text for details.
8	Additional Information	Any number of additional columns may be included for convenience (e.g., HGMD accession numbers, phenotypic annotations, source or contributor information, etc.).

Find out more at **thermofisher.com/microarrays**

ThermoFisher
SCIENTIFIC

For Research Use Only. Not for use in diagnostic procedures.