AFFYMETRIX ARABIDOPSIS GENOME ARRAY DESIGN, ANNOTATION, AND ANALYSIS

Original summary generated in 2003
Updated in 2007

Brandon Le
Javier Wagmaister
Anhthu Bui

TABLE OF CONTENTS

I. ARRAY INFORMATION & DESIGN

This section contains information about the design of the GeneChip array and general interpretation of features on the array.

A. Common Definition & Terminology

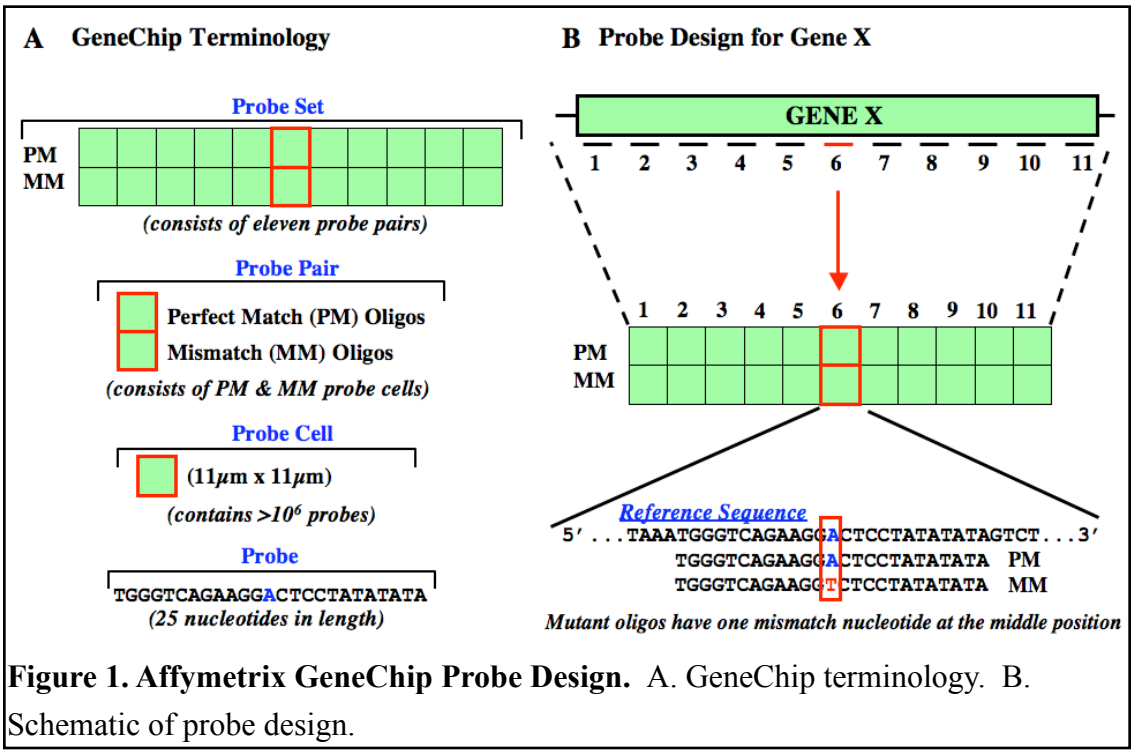

Figure 1. Affymetrix GeneChip Probe Design. A. GeneChip terminology. B. Schematic of probe design.

Probe: A single stranded DNA oligonucleotide designed to match a specific mRNA sequence. GeneChip probe arrays use oligonucleotide probes that are up to 25 bases long (**Figure 1**). The probes are synthesized directly on the surface of the array using photolithography and combinatorial chemistry.

Probe Cell: A single square-shaped feature on an array containing one type of probe. The size can vary depending on the array type, but the AtGenome1 (8K) and ATH1 (22K) array contains 24 and 18 µm probe cells, respectively. Each probe cell contains millions of probe molecules representing a unique gene-specific 25-mer oligo (**Figure 1**).

Perfect Match (PM): Probes that are designed to be exactly the same as the reference sequence (**Figure 1**).

Mismatch (MM): Probes that are designed to be exactly the same as the reference sequence except for a homomeric mismatch at the central position. Mismatch probes serve as a control for cross-hybridization (**Figure 1**).

Probe Pair: Consists of two probe cells, a PM and the corresponding MM probe cells. On the array, a probe pair is arranged with a PM cell directly above the MM cell (**Figure 1)**.

Probe Set: A set of probes designed to detect one transcript. A probe set usually consists of 11-20 probe pairs. The AtGenome1 (8K) and ATH1 (22K) array contains probe sets with 16 and 11 probe pairs, respectively (**Figure 1**).

B. Probe Set Nomenclature

Each probe set is assigned an Affymetrix serial number ID (five and six digits for the AtGenome1 and ATH1 array, respectively) (see below). IDs containing "AFFX" represents control features on the array. Each ID is followed by suffixes that denotes the type of sequence each probe set represents. Detailed description of each suffixes is presented in **Table 1** and **Figure 2**.

AFFYMETRIX SERIAL NUMBER ID_SUFFIXES

Table 1. Description of probe set suffixes

SUFFIXES	ARRAY*	DESCRIPTION
_at	**8K & 22K**	A unique probe set designed to be complementary to the cRNA target. ALL probe sets on the array have an _at designation.
_st ***(obsolete in the 22K)***	**8K**	Probe sets
_s_at	**8K**	(Similarity) Probe sets that corresponds to a small number of unique genes that share identical sequence. Probes were chosen from regions that are common to these genes. Group members can be singleton or group of sequences. The sequences that make up these probe sets are BAC sequences with corresponding mRNA sequences or overlapping BAC sequences. The sequences are **IDENTICAL.**
_g_at ***(obsolete in the 22K)***	**8K**	(Group) Probes chosen in region of overlap. Sequences are represented as singletons on the same probe array. These probe sets have sequence with overlapping concensus.
_f_at ***(obsolete in the 22K)***	**8K**	(Family) Probe sets that corresponds to sequences for which it was not possible to pick a full set of 16 unique and/or shared-similarity-constrained probes. Some probes are similar but not necessarily identical to other gene sequences.

SUFFIXES	ARRAY*	DESCRIPTION
_i_at ***(obsolete in the 22K)***	**8K**	(Incomplete) Sequences for which there are fewer than 15 unique probes.
_r_at ***(obsolete in the 22K)***	**8K**	(Rules Dropped) Sequences for which it was not possible to pick a full set of unique probes using Affymetrix probe rules. Probes were picked by dropping some of these rules.
_a_at	**22K**	A probe set that recognizes alternative transcripts from the same gene (**Figure 3**).
_s_at	**22K**	A probe set with all probes common among multiple transcripts within a gene family (these probe sets can detect members of a gene family - **Figure 3**).
_x_at	**22K**	A probe sets with some probes that are identical, or highly similar, to unrelated sequences. These probes may cross-hybridize in an unpredictable manner with sequences other than the main target. **Data generated from these probe sets should be interpreted with caution, due to the likelihood that some of the signal is from transcripts other than the one being intentionally measured.**

***** 8K refers to the first generation *Arabidopsis* array (AtGenome1) representing ~8,000 genes. 22K refers to the second generation *Arabidopsis* array (ATH1-121501) representing ~25,000 genes (commonly referred as the "whole genome" array".

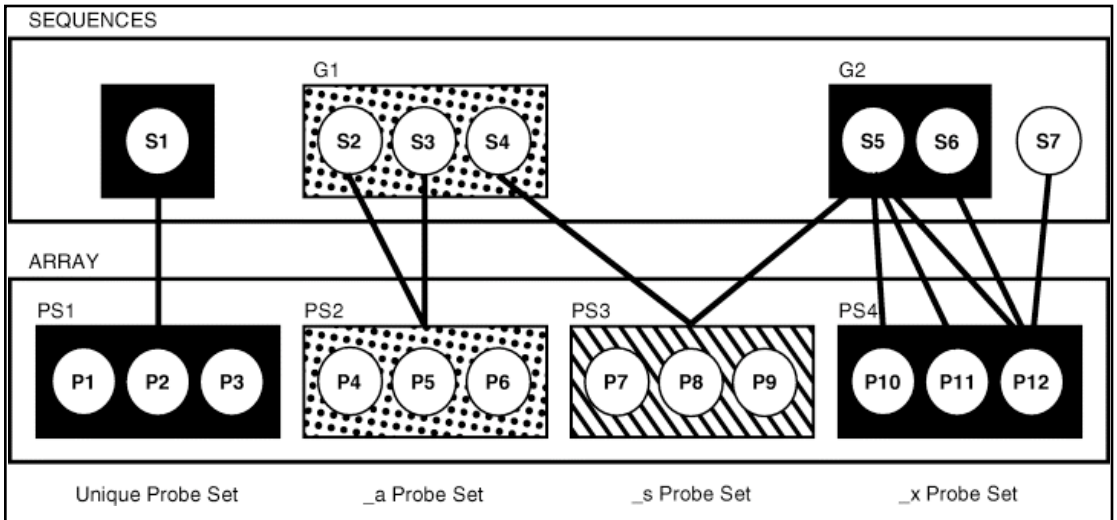

Figure 2. Possible Relationships Between Sequences And Probe Sets on an Array. G1, G2, etc., represent families of related gene sequences; S1, S2, S3, etc., represent individual gene sequences; PS1, PS2, PS3, etc., represent probe sets; P1, P2, P3, etc., represent probes. Solid lines show how sequence sets or sequences are related to probe sets or individual probes. If a solid line is connected to a box, the relationship applies to all probes in that box. If it is connected to an individual circle, the relationship holds only for that probe. (Figure and figure legend taken from Redman et al., Plant J 38(3) (2004)).

II. ANNOTATION OF THE ATGENOME1 ARRAY (2001)

A. Array Information

The Affymetrix *Arabidopsis* Genome GeneChip (AtGenome1) array is the first generation *Arabidopsis* array designed by Affymetrix in collaboration with Novartis Agriculture Discovery Institute, Inc (NADII). There are ~ 8200 genes represented on the Affymetrix *Arabidopsis* GeneChip. Eighty percent of the features are predicted coding sequences from genomic bacterial artificial chromosome (BAC) while 20% of the features are represent high quality cDNA sequences. Furthermore, there are > 100 EST clusters sharing homology with the predicted coding sequences from BAC clones. Each gene (probe set) is represented by sixteen probe pairs, with each probe consisting of 25-mer oligos localized to a 24µm feature size.

B. Array Features

We fully characterized the array by first examining the total number of features and genes represented on the array. Then, we examined the types of probe sets represented on the array based on the suffixes used to distinguish different probe types.

B1. How many features are represented on the array?
We counted the total number of probe sets on the array. Control probe sets have the "AFFX" designation in the header.

Table 2. Total number of features represented on the array

Description	# Probe Sets
Arabidopsis Genes	8,247
Proprietary Features	578
Controls	50
Total	**8,875***

***** Total number of probe sets based on the paper by Zhu and Wang, Plant Physiol. 124, (2000). The actual number of probe sets we have information for is 8,297.

B2. What are the representation of different suffixes on the array?
Probe IDs were extracted from the annotation file provided by Affymetrix. Using Microsoft Excel and the COUNTIF function, we determined the distribution of different sufffixes on the array. This analysis only considers the 8,247 non-proprietary features on the array excluding controls (**Table 2**).

Table 3. Distribution of suffixes on the array

Suffix	# Probe Sets
_at	6,242
_s_at	1,423
_i_at	208
_g_at	196
_f_at	101
_r_at	77
Total	**8,247**

C. Array Annotation Strategy

During the initial release of this array, very limited information was provided by Affymetrix and TMRI. We had to contact many people to obtain enough information to characterize and annotate the array (**Figure 3**).

1. We contacted Stacey Harmer, a post-doc in Steve Kay's laboratory, after reading her paper using this array (Harmer et al., Science 290, (2000)). We were able to obtain information regarding the sequences of the genes represented on the array.

2.We were directed to the Torrey Mesa Research Institute (TMRI) website (http://www.tmri.org/gene_exp/index.html), a division of Novartis that supplied Affymetrix with the sequences used to design the oligos on the GeneChip.

3. Sequences were downloaded from the site and we manually annotated the genes, assigning functional categories based on BLAST results, published literatures, internet searches (www.googles.com), molecular and biochemistry books, etc. (similar to our approach with EST sequence analysis).

4. Ceres provided a small file containing annotations for some of the genes on the GeneChip.

5. Using all the resources we had, we manually annotated all 8200 genes assigning each gene into a functional category based on the EU *Arabidopsis* Sequencing Project (**Figure 3**).

6. Probe sets representing predicted or hypothetical proteins that couldn't be assigned into a category were placed in the Unknown Function category.

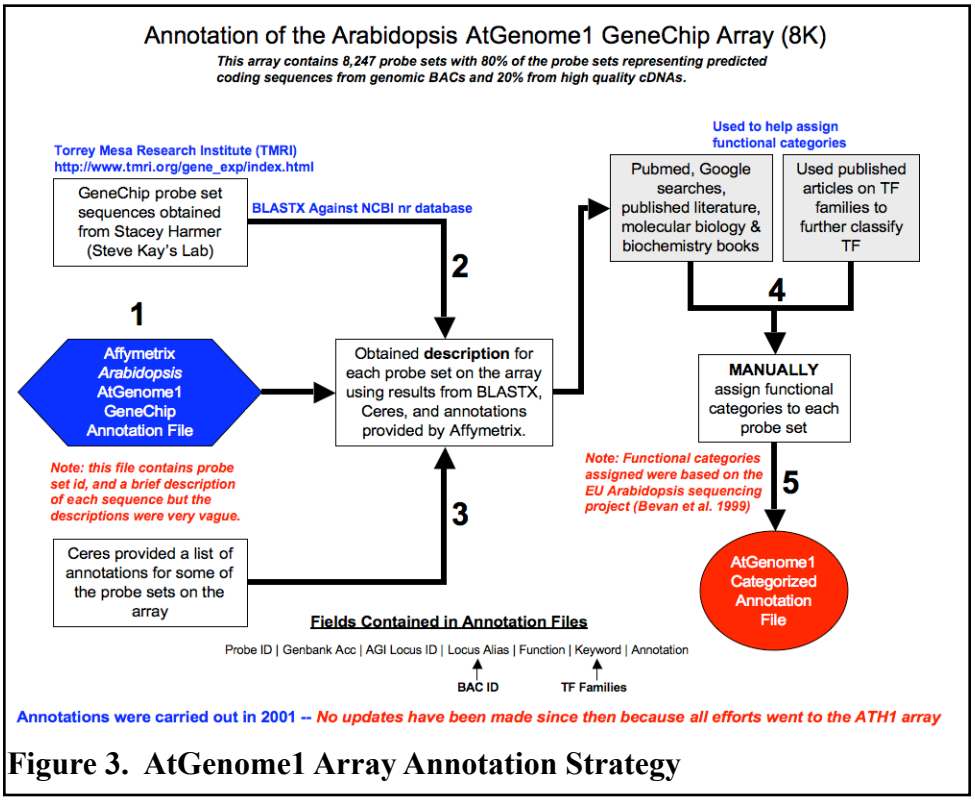

Figure 3. AtGenome1 Array Annotation Strategy

D. Array Annotation Summary

Using the strategy presented in **Figure 3**, we manually annotated sequences on the array and assigned functional categories to each sequence based on BLAST results, references, web searches, etc. For genes that encode transcription factors, we assigned the keyword transcription factor with a description of the transcription factor family i.e. myb, homeobox, AP2 domain containing protein, etc. The results are summarized in **Tables 4 & 5** and **Figures 4 & 5**.

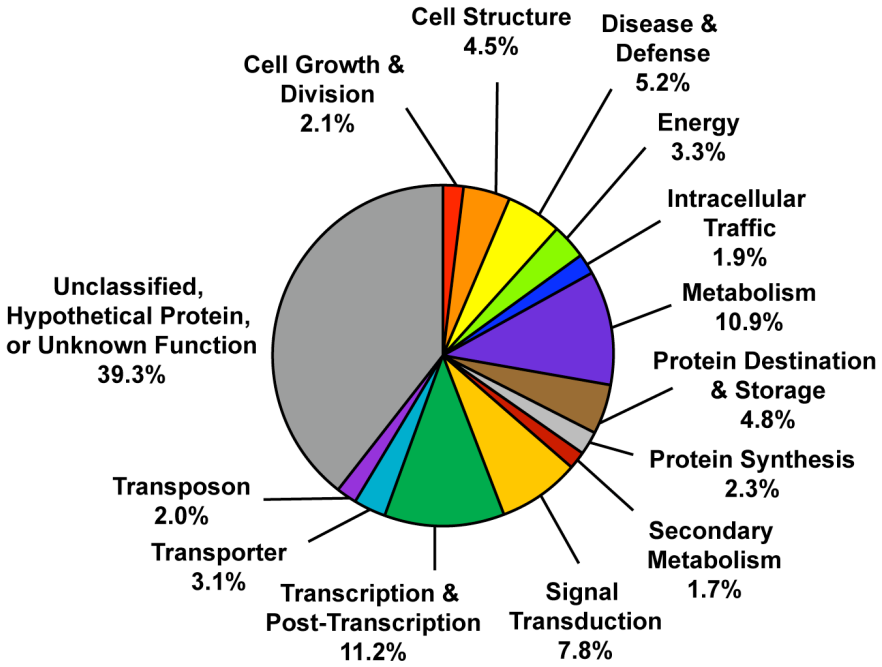

Figure 4. Distribution of 8,247 Features on the AtGenome1 Array.

Table 4. Functional category distribution of all probe sets

Functional Category	# Probe Sets	% Total
Cell Growth & Division	**172**	**2.1%**
Cell Structure	**370**	**4.5%**
Disease & Defense	**427**	**5.2%**
Energy	**269**	**3.3%**
Intracellular Traffic	**161**	**1.9%**
Metabolism	**896**	**10.9%**
Protein Destination & Storage	**399**	**4.8%**
Protein Synthesis	**189**	**2.3%**
Secondary Metabolism	**139**	**1.7%**
Signal Transduction	**643**	**7.8%**
Transcription & Post-Transcription	**926**	**11.2%**
Transporter	**255**	**3.1%**
Transposon	**161**	**2.0%**
Unknown Function	**3240**	**39.3%**
Total	**8,247**	

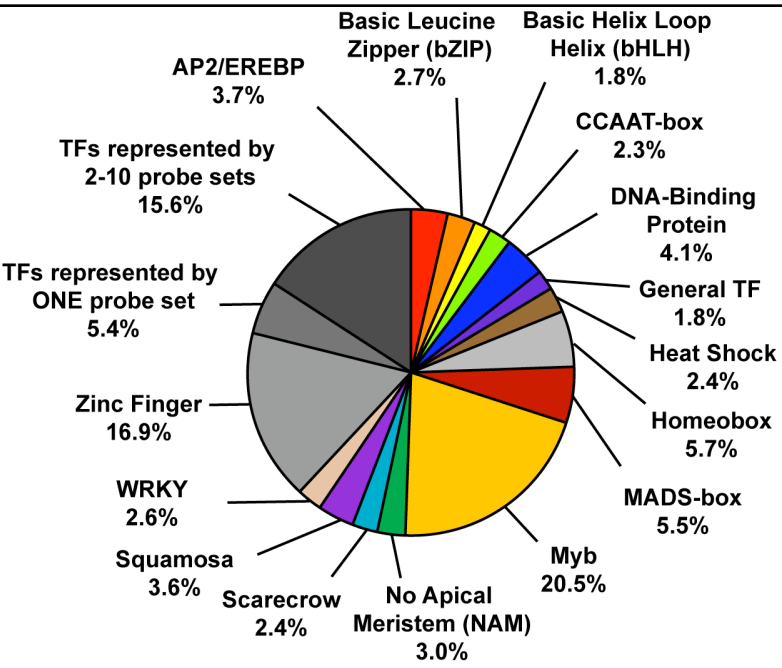

Figure 5. Distribution of Transcription Factor Families on the AtGenome1 Array.

Table 5. Distribution of transcription factor families on the array

Transcription Factor Families	# Probe Sets	% Total
AP2 Domain Containing Protein	**26**	**3.7%**
Basic Leucine Zipper (bZIP)	**19**	**2.7%**
Basic Helix-Loop-Helix (bHLH)	**13**	**1.8%**
CCAAT-box	**16**	**2.3%**
DNA-Binding Protein	**29**	**4.1%**
General Transcription Factor	**13**	**1.8%**
Heat Shock	**17**	**2.4%**
Homeobox	**40**	**5.7%**
MADS-box	**39**	**5.5%**
Myb	**144**	**20.5%**
No Apical Meristem (NAM)	**21**	**3.0%**
Scarecrow	**17**	**2.4%**
Squamosa	**25**	**3.6%**
WRKY	**18**	**2.6%**
Zinc Finger	**119**	**16.9%**
TFs represented by ONE probe set	**38**	**5.4%**
TFs represented by 2-10 probe sets	**110**	**15.6%**
Total	**704**	

III. ANNOTATION OF THE ATH1 ARRAY (2003)

A. Array Information

The Affymetrix *Arabidopsis* ATH1 Genome GeneChip (ATH1-121501) array is the second generation *Arabidopsis* array designed by Affymetrix in collaboration with The Institute for Genomic Research (TIGR). There are 22,746 probe sets representing 23,734 genes in the *Arabidopsis* genome (Redman et al., Plant J. 38(3) (2004)). 26,200 genes were available in the TIGR-ATH1 database as of December 15, 2001. To represent as many gene sequences on the array as possible, non-unique probe sets (_s_at) were used to represent two or more highly similar genes. Preference was given to genes where there was evidence of expression, supported by database matches and robust gene models. Each gene (probe set) is represented by eleven probe pairs, with each probe consisting of 25-mer oligos localized to an 18µm feature size. For more information about the array design, please read **Redman et al., Plant J. 38(3) (2004)**.

B. Array Features

Similar to the analysis carried out in **Section II**, we wanted to fully understand the nature of the ATH1 array by characterizing the features and representation of probe sets on the array.

B1. How many features are represented on the array?
We counted the total number of probe sets on the array. Control probe sets have the "AFFX" designation in the header.

Table 6. Total number of features represented on the array

Description	# Probe Sets
Arabidopsis Genes	22,746
Controls	64
Total	**22,810**

B2. What are the representation of different suffixes on the array?
Probe IDs were extracted from the annotation file provided by Affymetrix. Using Microsoft Excel and the COUNTIF function, we determined the distribution of different sufffixes on the array. This analysis only considers the 22,746 probe sets representing *Arabidopsis* genes and does not include control features.

Table 7. Distribution of suffixes on the array

Suffix	# Probe Sets
_at	21,685
_s_at	935
_x_at	126
Total	**22,746**

C. Array Annotation Strategy

Unlike the AtGenome1 array, sequences on the ATH1 array were freely available and accessible. We used a similar strategy as the AtGenome1 array to annotate the ATH1 array. The steps taken to annotate the ATH1 array are summarized in **Figure 6** and are elaborated in more details below:

1. We first downloaded the target sequences of each probe set on the array from the Affymetrix site (http://www.affmetrix.com//support/technical/byproduct.affx?product=arab). These target sequences represent the last 600 bp of the coding sequence. Note: 3' untranslated regions (UTRs) were not included in the probe design.

2. Sequences were blasted against NCBI non-redundant database using the blastX program.

3. Using results from the blast run and descriptions provided by Affymetrix for each probe set, manually annotate each features using classification based on the EU *Arabidopsis* Sequencing Project. To expedite the annotation process, we used the functional category assignment of 6,714 probe sets from the AtGenome1 array that have comparable match to probe sets on the ATH1 array.

4. We marked probe sets representing ~5,000 full-length cDNAs provided by Ceres to TIGR for gene annotation analysis.

5. We also included information of probe sets representing genes with known mutant phenotypes extrapolated from David Meinke's 2003 Plant Physiology paper (Meinke et al., Plant Physiol. 131, 2003).

6. Genes encoding transcription factors and signal transduction proteins were further classified into transcription factor families and signal transduction categories (e.g. kinases, phosphatases, etc.) based on published information, molecular biology books, and the web.

7. There are three categories for unclassified proteins: unclassified - hypothetical proteins with no cDNA support, unclassified - hypothetical proteins with cDNA support, and unclassified - proteins with unknown function.

a. **Unclassified - hypothetical proteins with no cDNA support** category includes sequences with description pertaining to hypothetical proteins or predicted proteins without any EST evidence.
b. **Unclassified - hypothetical proteins with cDNA support** category includes sequences with description indicating expressed proteins or predicted proteins supported by ESTs or Ceres's full-length cDNA.
c. **Unclassified - proteins with unknown function** category includes sequences representing proteins with known domains (i.e. WD40, DUF) but cannot be assign a category.

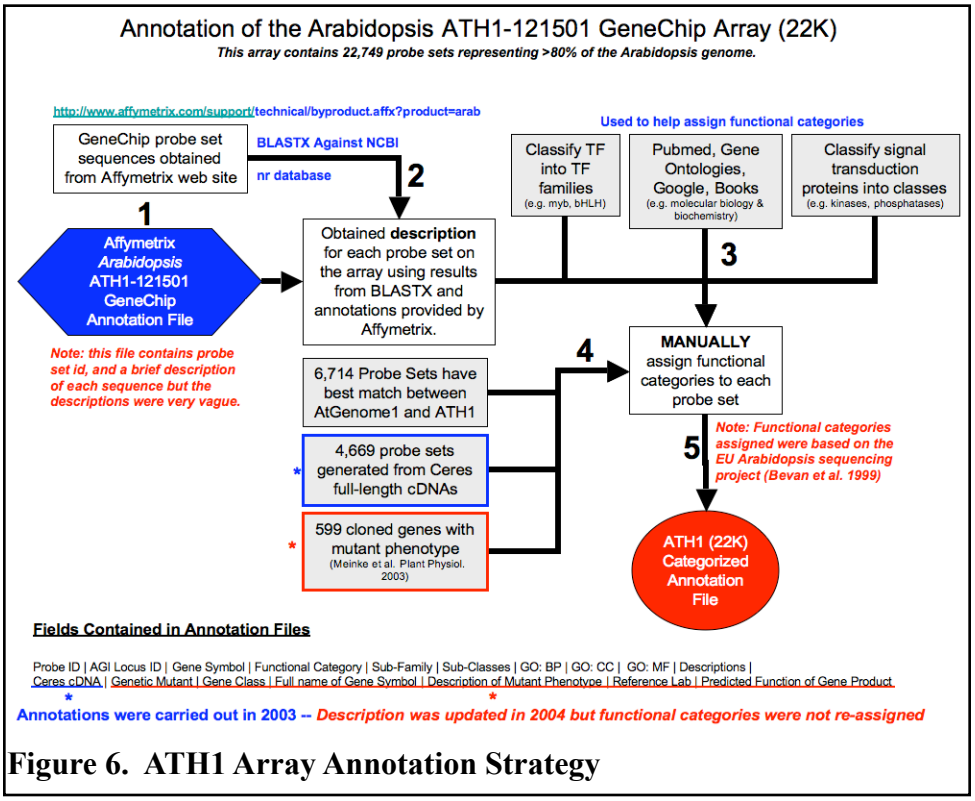

Figure 6. ATH1 Array Annotation Strategy

D. Array Annotation Summary

Summaries of functional category distribution on the array were tallied up using Microsoft Excel and are shown in **Tables 8 & 9** and **Figures 7 and 8**.

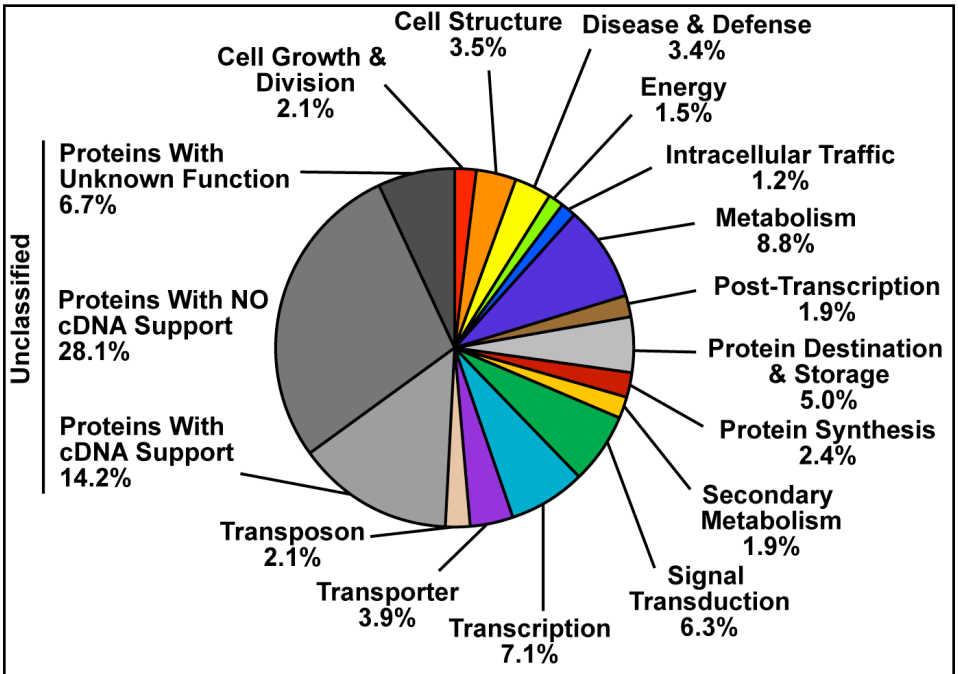

Figure 7. Distribution of 22,746 Features on the ATH1 Array.

Table 8. Functional category distribution of all probe sets on the ATH1 array

Functional Category	# Probe Sets	% Total
Cell Growth & Division	**476**	**2.1%**
Cell Structure	**793**	**3.5%**
Disease & Defense	**774**	**3.4%**
Energy	**347**	**1.5%**
Intracellular Traffic	**272**	**1.2%**
Metabolism	**1997**	**8.8%**
Post-Transcription	**423**	**1.9%**
Protein Destination & Storage	**1130**	**5.0%**
Protein Synthesis	**539**	**2.4%**
Secondary Metabolism	**424**	**1.9%**
Signal Transduction	**1430**	**6.3%**
Transcription	**1612**	**7.1%**
Transporter	**897**	**3.9%**
Transposon	**487**	**2.1%**
Unclassified - Proteins With cDNA Support	**3229**	**14.2%**
Unclassified - Proteins With NO cDNA Support	**6381**	**28.1%**
Unclassified - Proteins With Unknown Function	**1535**	**6.7%**
Total	**22,746**	

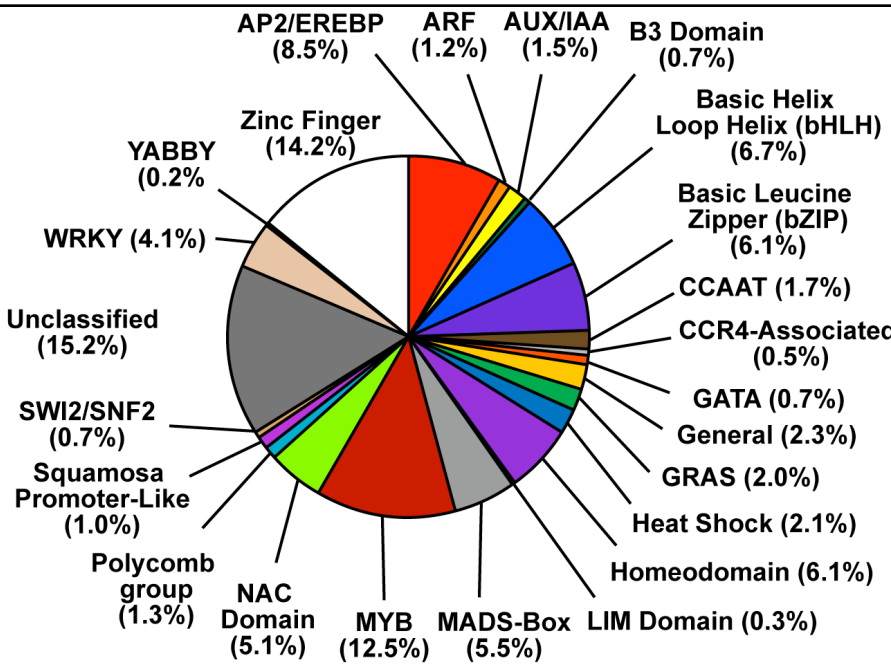

Figure 8. Distribution of Transcription Factor Families Represented on the ATH1 Array.

Table 9. Distribution of transcription factor families on the ATH1 array

Functional Category	# Probe Sets	% Total
AP2/EREBP	126	8.5%
ARF	18	1.2%
AUX/IAA	22	1.5%
B3 Domain	10	0.7%
Basic Helix-Loop-Helix (bHLH)	100	6.7%
Basic Leucine Zipper (bZIP)	90	6.1%
CCAAT	25	1.7%
CCR4-Associated	8	0.5%
GATA	11	0.7%
General	34	2.3%
GRAS	30	2.0%
Heat Shock	31	2.1%
Homeodomain	90	6.1%
LIM Domain	4	0.3%
MADS-Box	82	5.5%
MYB	185	12.5%
NAC Domain	75	5.1%
Polycomb group	19	1.3%
Squamosa Promoter-Like	15	1.0%

Functional Category	# Probe Sets	% Total
SWI2/SNF2	**10**	**0.7%**
Unclassified	**225**	**15.2%**
WRKY	**61**	**4.1%**
YABBY	**3**	**0.2%**
Zinc Finger	**210**	**14.2%**
Total	**1484**	

IV. COMPARISON OF THE ATGENOME1 AND ATH1 ARRAYS

A. Comparison of Array Features

We carried out experiments using both the first (AtGenome1) and second (ATH1) generation arrays for RNA profiling analysis. In order to correlate the results generated on the AtGenome1 versus ATH1 arrays, we needed to find the homologous sequence or probe sets represented on both arrays. Fortunately, Affymetrix carried out an extensive same-species array comparison (see the comparison sheet manual for more details: http://www.affymetrix.com/support/technical/manual/comparison_spreadsheets_manual.pdf). The similarities and differences between the AtGenome1 and ATH1 arrays are summarized in **Table 10**.

Table 10. AtGenome1 and ATH1 arrays comparison

Array Features	AtGenome1 (8K)	ATH1 (22K)
Number of Probe Sets	**8,247**	**22,746**
Oligonucleotide Length	**25-mer**	**25-mer**
Number of Probe Pairs	**16**	**11**
Probe Pair Distribution	**Contiguous**	**Scattered**

Note: Probe pairs for a given probe set were designed to be scattered across the ATH1 array to ensure that information for a probe set would not be lost if a GeneChip is "damaged" locally in one section of the array. However, in the AtGenome1 array, since all probe pairs are next to each other, if the probe set is in a "damaged" area, then the probe set can no longer be used.

B. Mapping of Probe Sets Across the Two Arrays

Due to the dynamic nature of the public databases, probe sets between the two *Arabidopsis* arrays will not be identical. In some cases the same sequences will be represented by completely different probe sets, creating a challenge when comparing data sets generated on different generations. **Table 11** shows results from the comparison analysis by Affymetrix.

Table 11. Affymetrix Array Comparison

Match Description*	# Probe Sets
Best Probe Match	**6,714**
Good Probe Match	**6,731**
Complex Probe Match	**8,312**
No Probe Match	**516**

* A detailed description of each probe match criteria is available from Affymetrix (http://www.affymetrix.com/support/technical/manual/comparison_spreadsheets_manual.pdf). Figure 9 illustrates the relationships between probe sets from two arrays and the designation for each probe set.

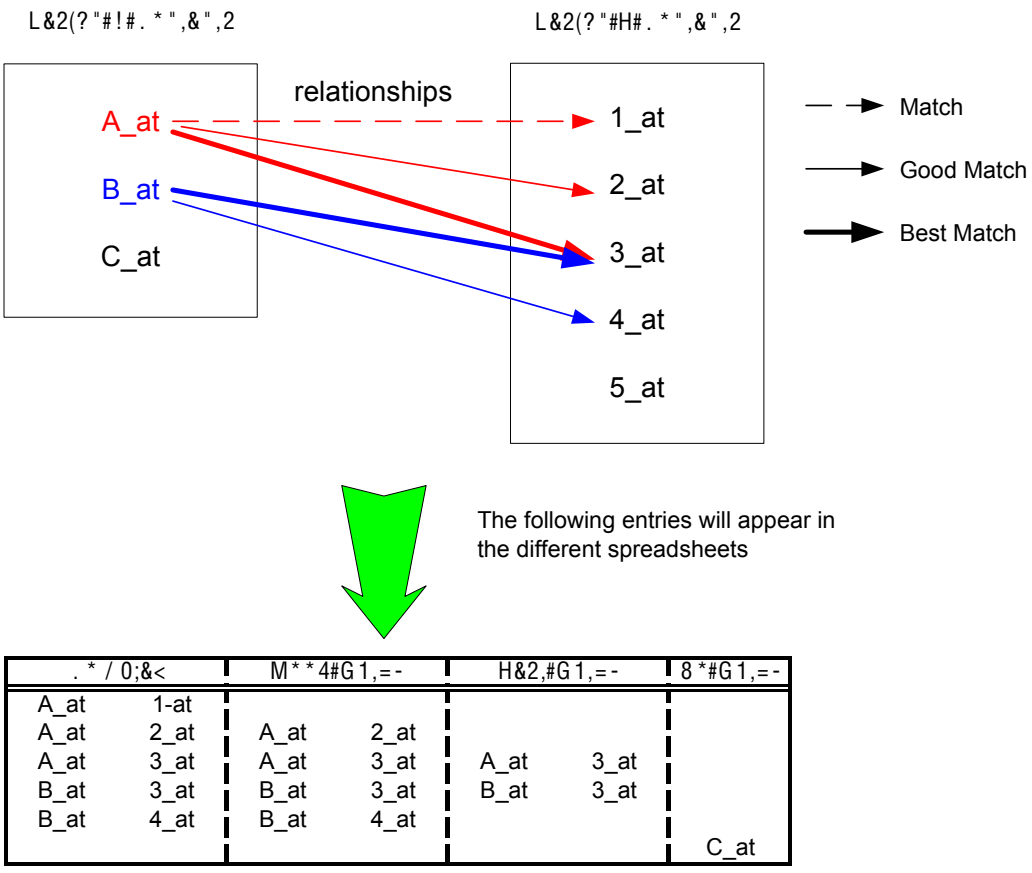

Figure 9. Possible Probe Set Relationships Between Arrays. In the case of the *Arabidopsis* arrays, Design A and B refers to the AtGenome1 and ATH1 arrays, respectively. The type of match (Best, Good, Complex) depends on the level of stringency applied to a probe set. This analysis was done at the probe level (25-mer) and not at the consensus or target sequence level.

V. RE-ANNOTATION OF THE ATH1 ARRAY (2007)

A. Motivation for Re-Annotation Efforts

The *Arabidopsis* ATH1 array was annotated in 2003 using all the publicly available resources at the time (see **Section III**). The annotations were used to quickly annotate the Soybean Genome array (see the Soybean Genome Array Annotation summary for more details). In order to keep up with the increasing amount of information generated within the past four years since the annotation of the ATH1 array, we decided to re-annotate the ATH1 array in parallel with the soybean genome array.

B. Array Re-Annotation Strategy

The strategy for the re-annotation of the ATH1 array is summarized in **Figure 10** below. Our strategy is as follows:

1. We updated the descriptions for each probe set on the array using TAIR Affy array descriptions (**affy_ATH1_array_elements-2007-5-2.txt**). The description file was downloaded from the TAIR web site: ftp://ftp.arabidopsis.org/home/tair/Microarrays. Descriptions were based on the latest release of the *Arabidopsis* genome TAIR 7 (released 04-11-07). **Note from TAIR:** The mapping to the TAIR7 Transcripts was performed using the BLASTN program with e-value cutoff < **9.9e-6**. For the 25-mer oligo probes used on the Affy chips, the required match length to achieve this e-value is **23 or more identical nucleotides**. To assign a probe set to a given locus, **at least 9 of the probes** included in the probe set were required to match a transcript at that locus. Otherwise, the probe set was not assigned a locus and was given the description "no match".

2. In addition to updating the descriptions for each probe set, we also updated gene ontology (GO) information provided by Affymetrix.

3. We gathered information about putative transcription factors from many publicly available TF database for *Arabidopsis* including:

   a. **AGRIS** - ***A**rabidopsis* **G**ene **R**egulatory **I**nformation **S**erver (http://arabidopsis.med.ohio-state.edu/)

   b. **DATF** - **D**atabase of ***A**rabidopsis* **T**ranscription **F**actors (http://datf.cbi.pku.edu.cn/)

   c. **RARTF** - **R**iken ***A**rabidopsis* **T**ranscription **F**actor Database (http://rarge.gsc.riken.jp/rartf/)

   d. **ArabTFDB** - ***Arab**idopsis* **T**ranscription **F**actor **D**atab**a**se (http://arabtfdb.bio.uni-potsdam.de/v1.1/)

Transcription factors and transcription factor families were associated with each probe set on the array.

Information obtained from points 1-3 were compiled together into an annotation file containing the 2003 ATH1 annotations. Transcription factors were automatically updated based on the information obtained from the databases in point 3.

4. We focused on probe sets that were previously assigned into the "unclassified" category. The rationale is that many of the sequences in the "unclassified" category might have update information that can be used to re-assign into a different category. Sequences previously assigned categories of "protein synthesis" or "metabolism" most likely will not change. Therefore, we first focused on re-assigning the 11,145 probe sets classified as "unclassified" in 2003.

5. After the "unclassified" category was re-examined, we decided to re-examine the entire 22,746 probe sets on the array for consistent assignment of functional categories. We sorted all the probe sets by their description and made sure that probe sets with similar descriptions are assigned the same functional category.

6. We further examined the "unclassified" category that is divided into three groups as mentioned in **Section IIIC**. We obtained several files from TAIR that will distinguish the different sequences within the unclassified category. We downloaded several files from the TAIR site including:

a. TAIR7_protein_coding_no_transcript_support_09_30_07
b. TAIR7_protein_coding_with_transcript_support_09_30_07
c. TAIR7_unknown_proteins_no_transcript_support_09_30_07
d. TAIR7_proteins_of_undefined_function_03_07
e. TAIR7_unknown_proteins_03_07
f. TAIR7_locus_type

These files were compiled into one main table listing all the transcripts detected and/or predicted in the *Arabidopsis* genome. This list helps distinguish if a sequence has cDNA support, represents a pseudogene/transposon, or is unknown. These files help re-assign the probe sets into appropriate unclassified categories.

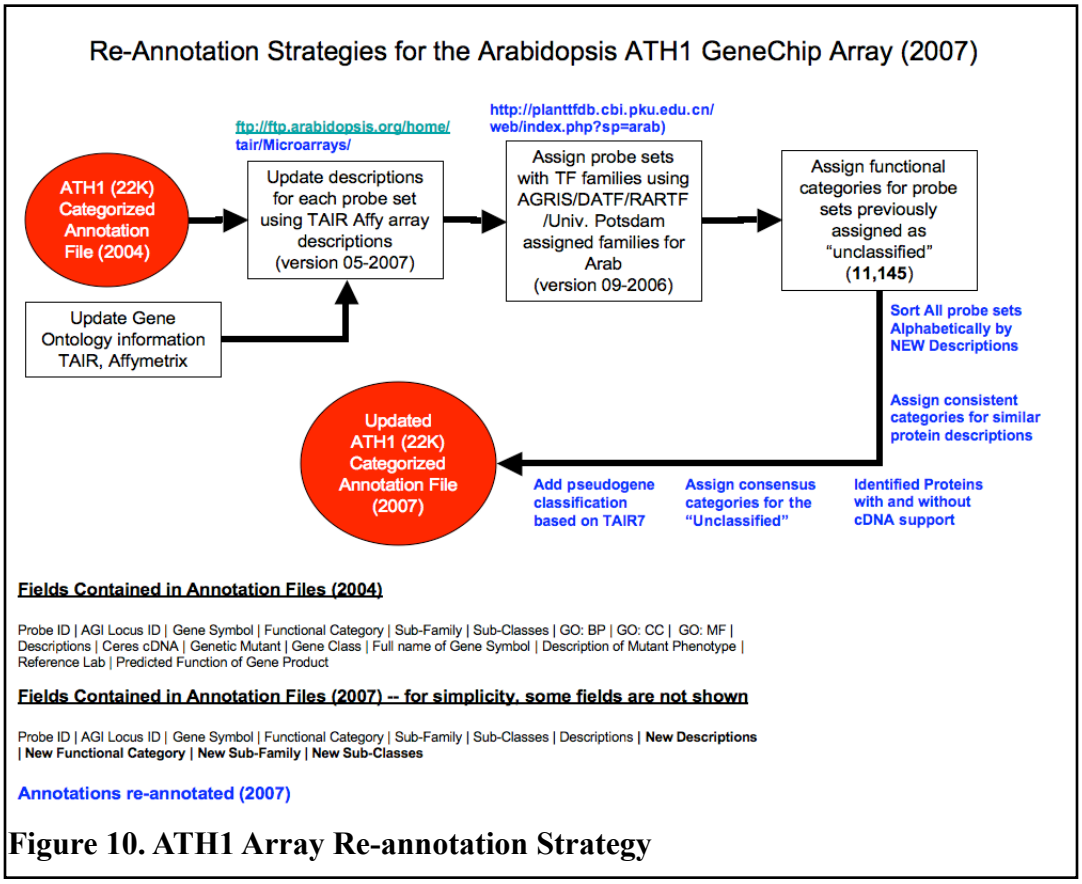

Figure 10. ATH1 Array Re-annotation Strategy

C. Array Re-Annotation Summary

The final annotation file contains a list of all the probe sets with annotation information from 2003 and the latest information (2007). ***Arabidopsis* ATH1 Array Annotation version 2.4 is the current version used for all summaries.** **Tables 12** and **Figure 11** summarizes functional category assignment of all probe sets on the array. **Table 13** and **Figure 12** summarizes all identified transcription factor probe sets, on the array.

Table 12. Functional category distribution of all probe sets on the ATH1 array

Functional Category	# Probe Sets	% Total
Cell Growth & Division	621	**2.7%**
Cell Structure	1108	**4.9%**
Disease & Defense	912	**4.0%**
Energy	729	**3.2%**
Intracellular Traffic	624	**2.7%**
Metabolism	2848	**12.5%**

Functional Category	# Probe Sets	% Total
Post-Transcription	668	**2.9%**
Protein Destination & Storage	1573	**6.9%**
Protein Synthesis	616	**2.7%**
Pseudogene	320	**1.4%**
Secondary Metabolism	458	**2.0%**
Signal Transduction	1950	**8.6%**
Transcription	2356	**10.4%**
Transporter	1131	**5.0%**
Transposon	498	**2.2%**
Unclassified - Proteins With cDNA Support	3273	**14.4%**
Unclassified - Proteins With NO cDNA Support	1049	**4.6%**
Unclassified - Proteins With Unknown Function	2012	**8.8%**
Total	**22,746**	

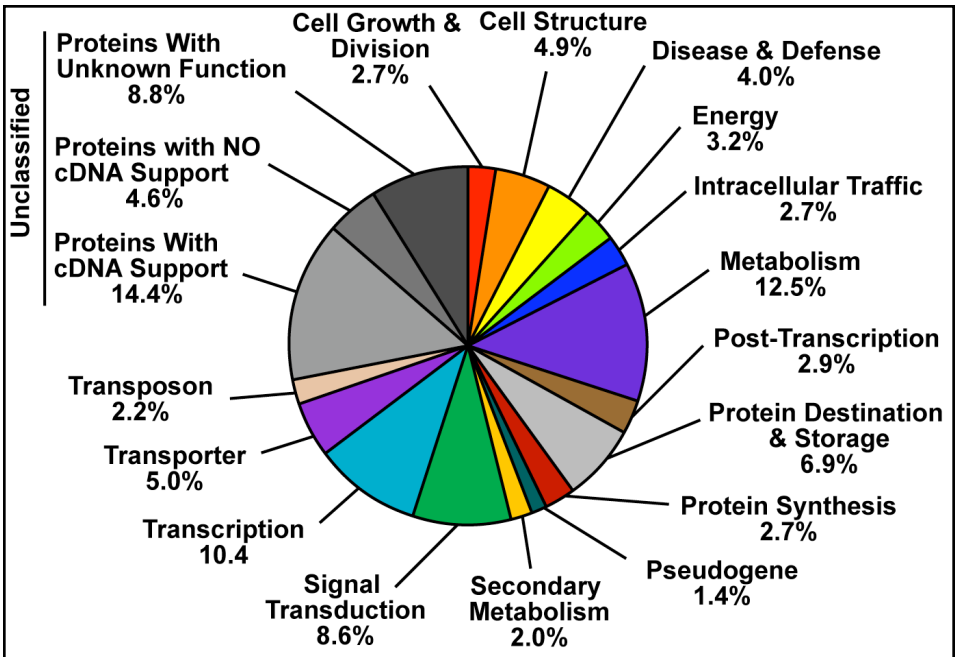

Figure 11. Distribution of 22,746 Features on the Re-Annotated ATH1 Array (2007).

Table 13. Distribution of transcription factor probe sets on the array.

Transcription Factor Families	Pie Groups	# Probe Sets	% Total
ABI3-VP1	ABI3-VP1	45	**2.3%**
AP2-EREBP	AP2-EREBP	131	**6.7%**
ARF	ARF	20	**1.0%**

Transcription Factor Families	Pie Groups	# Probe Sets	% Total
ARR-B	ARR-B	20	**1.0%**
AS2	AS2	26	**1.3%**
AUX-IAA	AUX-IAA	28	**1.4%**
bHLH	bHLH	120	**6.1%**
bZIP	bZIP	79	**4.0%**
CCAAT-Dr1	CCAAT	2	**0.1%**
CCAAT-HAP2	CCAAT	10	**0.5%**
CCAAT-HAP3	CCAAT	11	**0.6%**
CCAAT-HAP5	CCAAT	11	**0.6%**
G2-like	G2-like	38	**1.9%**
General	General	42	**2.1%**
GRAS	GRAS	33	**1.7%**
HB	HB	94	**4.8%**
HSF	HSF	24	**1.2%**
MADS	MADS	89	**4.6%**
MYB-related	MYB	55	**2.8%**
MYB	MYB	149	**7.6%**
NAC	NAC	96	**4.9%**
LFY	Others	1	**0.1%**
NZZ	Others	1	**0.1%**
SAP	Others	1	**0.1%**
ULT	Others	1	**0.1%**
LUG	Others	2	**0.1%**
GIF	Others	3	**0.2%**
MBF1	Others	3	**0.2%**
S1Fa-like	Others	3	**0.2%**
Whirly	Others	3	**0.2%**
CSD	Others	4	**0.2%**
BZR-BES1	Others	5	**0.3%**
Pseudo ARR-B	Others	5	**0.3%**
BBR-BPC	Others	6	**0.3%**
CPP	Others	6	**0.3%**
EIL	Others	6	**0.3%**
Sigma70-like	Others	6	**0.3%**
CAMTA	Others	7	**0.4%**
E2F-DP	Others	7	**0.4%**
PLATZ	Others	7	**0.4%**
GRF	Others	8	**0.4%**
TAZ	Others	8	**0.4%**
ARID	Others	10	**0.5%**
Nin-like	Others	11	**0.6%**
TUB	Others	11	**0.6%**

Transcription Factor Families	Pie Groups	# Probe Sets	% Total
HMG	Others	12	**0.6%**
FHA	Others	13	**0.7%**
LIM	Others	14	**0.7%**
GeBP	Others	15	**0.8%**
SBP	Others	16	**0.8%**
TCP	Others	17	**0.9%**
JUMONJI	Others	19	**1.0%**
PcG	PcG	35	**1.8%**
Trihelix	Trihelix	22	**1.1%**
Unclassified	Unclassified	50	**2.6%**
WRKY	WRKY	66	**3.4%**
VOZ	ZF	2	**0.1%**
HRT	ZF	3	**0.2%**
C2C2-YABBY	ZF	5	**0.3%**
SRS	ZF	6	**0.3%**
Alfin	ZF	7	**0.4%**
ZIM	ZF	13	**0.7%**
ZF-HD	ZF	14	**0.7%**
C2C2-GATA	ZF	21	**1.1%**
C2C2-Dof	ZF	32	**1.6%**
C2C2-CO-like	ZF	38	**1.9%**
PHD	ZF	45	**2.3%**
C3H	ZF	116	**5.9%**
C2H2	ZF	125	**6.4%**
Total		1954	

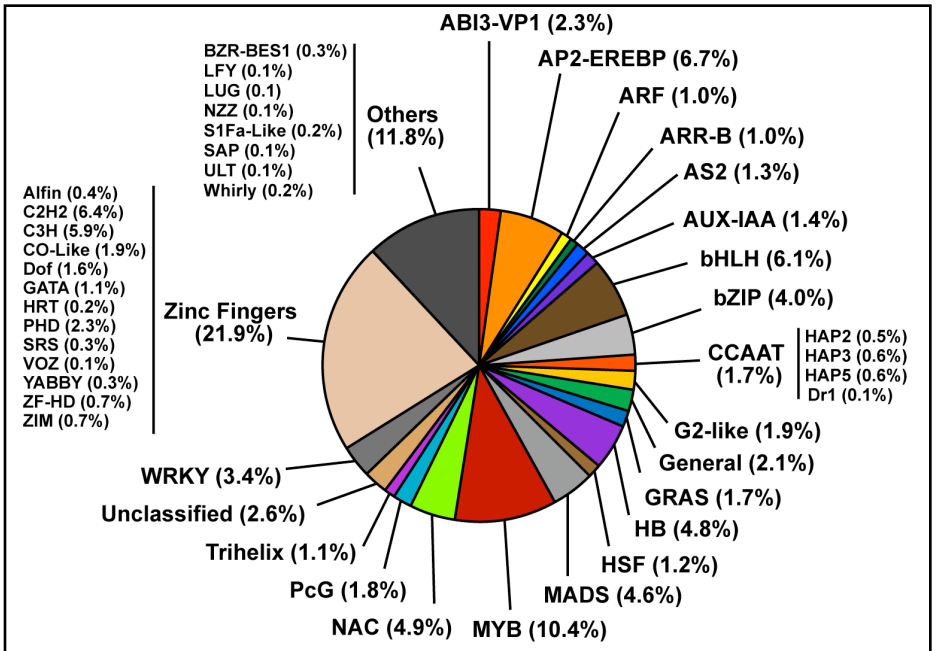

Figure 12. Pie Distribution of Transcription Factors Represented on the array. Other represents TF families with fewer than 20 members represented on the array (see Table13).

D. Comparison of Old and New Annotations

A. We first examined the overall difference in annotations carried out in 2003 and 2007. Notice that there is a 83.6% reduction of the Unclassified - Proteins with NO cDNA Support category. Interestingly, we see more than a 100% increase in probe sets classified in the Energy and Intracellular Traffic categories. Overall, a moderate increase occurred for each functional categories.

Table 14. Comparison of old and new annotation of the ATH1 array

Functional Categories	2003	2007	% Change
Cell Growth & Division	476	621	**30.5%**
Cell Structure	793	1108	**39.7%**
Disease & Defense	774	912	**17.8%**
Energy	347	729	**110.1%**
Intracellular Traffic	272	624	**129.4%**
Metabolism	1997	2848	**42.6%**
Post-Transcription	423	668	**57.9%**
Protein Destination & Storage	1130	1573	**39.2%**
Protein Synthesis	539	616	**14.3%**

Functional Categories	2003	2007	% Change
Pseudogene	0	320	**INF**
Secondary Metabolism	424	458	**8.0%**
Signal Transduction	1430	1950	**36.4%**
Transcription	1612	2356	**46.2%**
Transporter	897	1131	**26.1%**
Transposon	487	498	**2.3%**
Unclassified - Proteins With cDNA Support	3229	3273	**1.4%**
Unclassified - Proteins With NO cDNA Support	6381	1049	**-83.6%**
Unclassified - Proteins With Unknown Function	1535	2012	**31.1%**
Total	22746	22746	

B. We next examined what changed within each of the three unclassified categories shown earlier. Tables 15-17 shows the assignment of probe sets previously classified as "unclassified".

Table 15. Changes within the unclassified - proteins with NO cDNA support category (2003).

Functional Category	# Probe Sets	% Total
Cell Growth & Division	128	**2.0%**
Cell Structure	154	**2.4%**
Disease & Defense	85	**1.3%**
Energy	106	**1.7%**
Intracellular Traffic	135	**2.1%**
Metabolism	341	**5.3%**
Post-Transcription	140	**2.2%**
Protein Destination & Storage	315	**4.9%**
Protein Synthesis	50	**0.8%**
Pseudogene	231	**3.6%**
Secondary Metabolism	22	**0.3%**
Signal Transduction	251	**3.9%**
Transcription	445	**7.0%**
Transporter	169	**2.6%**
Transposon	18	**0.3%**
Unclassified - Proteins With cDNA Support	1851	**29.0%**
Unclassified - Proteins With NO cDNA Support	937	**14.7%**
Unclassified - Proteins With Unknown Function	1003	**15.7%**
Total	**6381**	

Table 16. Changes within the unclassified - proteins with cDNA support category (2003).

Functional Category	# Probe Sets	% Total
Cell Growth & Division	68	**2.1%**
Cell Structure	80	**2.5%**
Disease & Defense	79	**2.4%**
Energy	107	**3.3%**
Intracellular Traffic	108	**3.3%**
Metabolism	299	**9.3%**
Post-Transcription	62	**1.9%**
Protein Destination & Storage	123	**3.8%**
Protein Synthesis	41	**1.3%**
Pseudogene	4	**0.1%**
Secondary Metabolism	17	**0.5%**
Signal Transduction	139	**4.3%**
Transcription	233	**7.3%**
Transporter	109	**3.4%**
Transposon	6	**0.2%**
Unclassified - Proteins With cDNA Support	1222	**37.8%**
Unclassified - Proteins With NO cDNA Support	0	**0.0%**
Unclassified - Proteins With Unknown Function	532	**16.4%**
Total	**3229**	

Table 17. Changes within the unclassified - protein of unknown function category (2003).

Functional Category	# Probe Sets	% Total
Cell Growth & Division	37	**2.4%**
Cell Structure	54	**3.5%**
Disease & Defense	73	**4.8%**
Energy	45	**2.9%**
Intracellular Traffic	44	**2.9%**
Metabolism	207	**13.5%**
Post-Transcription	30	**2.0%**
Protein Destination & Storage	81	**5.3%**
Protein Synthesis	6	**0.4%**
Pseudogene	8	**0.5%**
Secondary Metabolism	19	**1.2%**
Signal Transduction	98	**6.4%**
Transcription	93	**6.1%**

Functional Category	# Probe Sets	% Total
Transporter	45	**2.9%**
Transposon	0	**0.0%**
Unclassified - Proteins With cDNA Support	185	**12.1%**
Unclassified - Proteins With NO cDNA Support	100	**6.5%**
Unclassified - Proteins With Unknown Function	410	**26.7%**
Total	**1535**	