AFFYMETRIX SOYBEAN GENOME ARRAY DESIGN, ANNOTATION, AND ANALYSIS

Original summary generated in 2004
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 array and general interpretation of features on the array.

A. Common Definition & Terminology

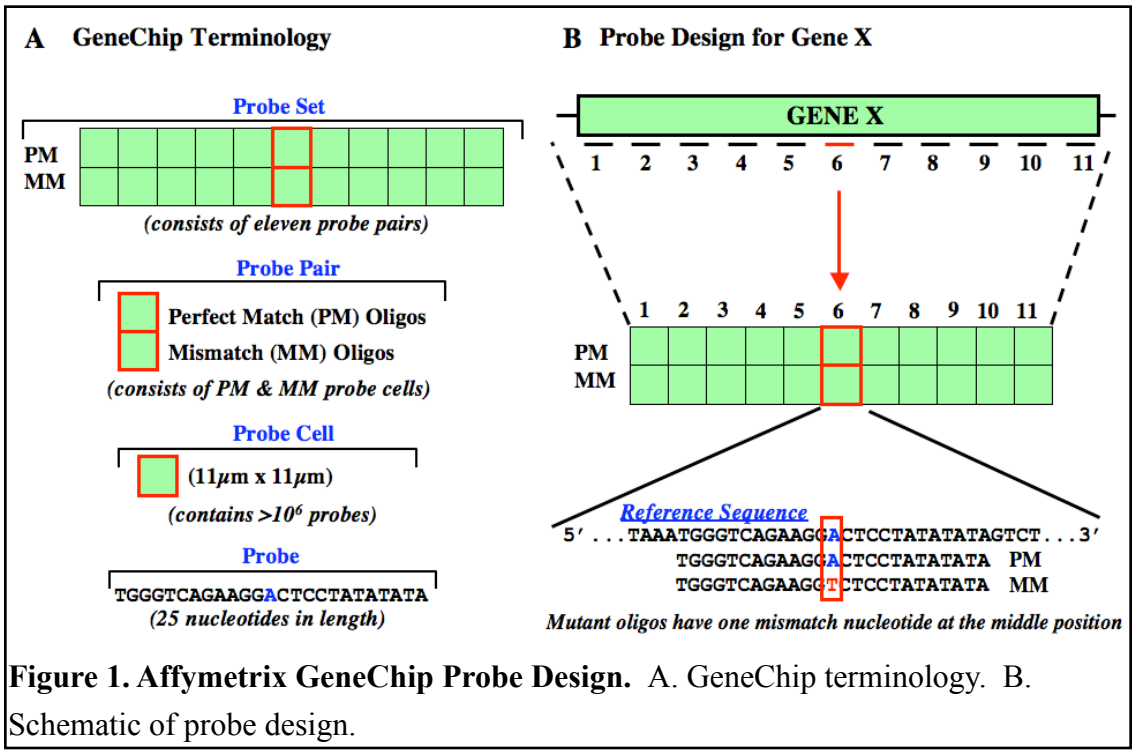

Figure 1. Affymetrix GeneChip Probe Design. A. GeneChip terminology. B. Schematic of probe design.

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 in the soybean array is 11 µm. 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. For the soybean array, a probe set is reduced down to 11 probe pairs consisting of 11 PM and 11 MM probe cells for a total of 22 probe cells (**Figure 1**).

B. Array Information

Information about the array was taken from the Soybean Array Data Sheet provided by Affymetrix (http://www.affymetrix.com/support/technical/byproduct.affx?product=soy).

1. The Soybean Genome Array contains probe sets interrogating three genomes:

    a. *Glycine max* (Soybean)
    b. *Phytophthora sojae* (a water mold that commonly attacks soybean crops)
    c. *Heterodera glycines* (cyst nematode pathogen)

2. The array contains 11µm features (**Figure 1**) containing 11 probe pair per sequence. Each probe pair contains a perfect match and mismatch probe consisting of 25-mer oligonucleotides.

3. Sequence information for the array was obtained from the public domain (GenBank, dbEST, etc…). Sequence clusters were created from UniGene Build 13 (November 5, 2003). Excerpts taken from the UniGene document: "For a sequence to be included in UniGene, the clone insert must have at least 100 base pairs that are of high quality and not repetitive. In addition, UniGene clusters are required to contain at least one sequence carrying readily identifiable evidence of having reached the 3' terminus, (i.e. anchored at the 3' end of a transcription unit.) Resulting clusters may contain more than one alternative splice form. Therefore, not all uncontaminated sequences in dbEST appear in UniGene clusters. (Pontias et al. 2002, UniGene: A Unified View of the Transcriptome, NCBI Handbook). Affymetrix carried out an in-house cluster analysis of the uncontaminated sequences that does not appear in Unigene clusters. ***Therefore, the array contains UniGene sequences (represented by probe sets with the prefix "Gma") and Affymetrix in-house cluster sequences (represented by probe sets with the prefix "GmaAffx")***.

4. The array was developed under the Consortia Program between the Soybean community (lead by Randy Shoemaker and Gary Stacey) and Affymetrix (lead by Alan Williams – alan_williams@affymetrix.com)

5. EST data from dbEST are derived from more than 80 soybean cDNA libraries including most developmental stages, plant tissues, and organs (flowers, leaves, roots, seedlings, stems, pods, seeds).

C. Probe Set Nomenclature

The probe set nomenclature is different from previous Affymetrix arrays like *Arabidopsis* ATH1 and Human U133 arrays. Previously, probe sets were named according to Affymetrix serial number ID. However, probe sets on the soybean array were named according to the UniGene ID or Affymetrix accession number created at the time of design. Below is the guideline for probe set nomenclature for the soybean array (taken from Laura Ramsundar, Affymetrix Field Application Scientist, email communication). Soybean probe set nomenclature has the following format:

HEADER.CLUSTER NO.SUB-CLUSTER.ORIENTATION_SUFFIXES

Table 1. Description of probe set headers

HEADERS	DESCRIPTION
Gma (e.g. **Gma**.6207.1.S1_s_at)	Probe ID representing Soybean UniGene clusters obtained from the public domain (Genbank) – Build #13 (November 2003)
GmaAffx (e.g. **GmaAffx**.12591.1.A1_s_at)	Probe ID representing Affymetrix de-novo clustering of the same Soybean EST data set obtained from the public domain (these ESTs are not included in the UniGene data set - **see part B** above for more details)
HgAffx (e.g.**HgAffx**.10017.1.S1_at)	Probe ID representing Affymetrix de-novo clustering of *H. glycines* EST data
PsAffx (e.g.**PsAffx**.C100000011_at) (e.g.**PsAffx**.Avh1b-20_at)	Probe ID representing Affymetrix de-novo clustering of *P. sojae* sequences provided by Brett Tyler's Lab. The "C__" numbers were from gene predictions vs the genomic sequence. Other names are derived from EST/cDNA identifier.
AFFX (e.g.**AFFX**.-r2-Ps-actin-5_s_at) (e.g.**AFFX**.r2-Hg-gadph-M_at)	"Control" probe set. Includes Actin, GADPH, BioB, C, D, and Cre. Some control probe sets represent the 5', middle, and 3' of the transcript (used to determine the integrity of the RNA). Good 3'/5' ratio (~2-3 indicates the entire length of the transcript was present). Low 3'/5' ratio indicates truncated RNAs.

Table 2. Description of cluster and orientation

CLUSTERS & ORIENTATION	DESCRIPTION
CLUSTER NUMBER (e.g. Gma.**6207**.1.S1_s_at) (e.g. GmaAffx.**12591**.1.A1_s_at)	This number corresponds to the archival UniGene Cluster ID (Build #13) or Affymetrix de-novo clustering ID. The UniGene Cluster ID does not necessarily correspond to the current UniGene Build #30 (July 30, 2007) for soybean.

Clusters & Orientation	Description
SUB-CLUSTER NUMBER (e.g. Gma.1234.**1**.S1_s_at) (e.g. Gma.1234.**2**.S1_s_at)	This number indicates alternative transcripts from the same gene. Therefore, *.1 and *.2 are putative transcript variants of the same gene.
ORIENTATION (e.g. Gma.1234.1.**S1**_s_at) (e.g. Gma.1234.1.**A1**_s_at)	The orientation of an EST cluster is denoted by "S - sense" and "A - anti-sense" (**see Figure 2**). Orientation is established via EST annotation, polyA tail, polyA signal, and canonical splice junction sequence if exonic structure is available from genome sequence. If orientation could not be determined from sequence information, arbitrary direction of A1 or S1 is assigned. Furthermore, two independent probe sets are tiled for each EST cluster. The number appended to A or S refers serially to alternative polyadenylation sites when EST "stacking" is observed in the context of the assembled cluster of ESTs. Often there is supporting polyA or polyA signals to substantiate the alternative 3' end of the transcript. A1, for example, is the most proximal polyA site relative to the translation stop codon. A2 would be distal to A1.

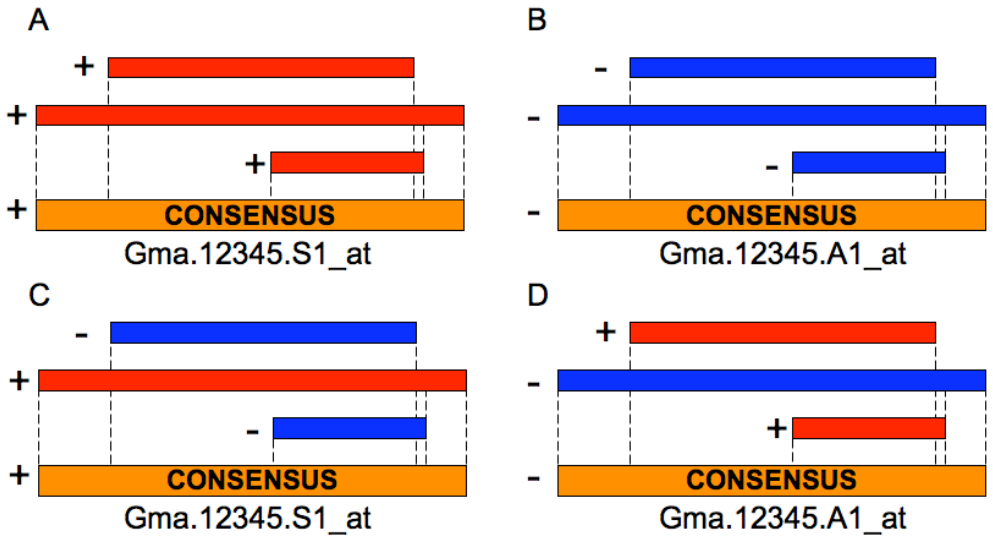

Figure 2. Conceptual Diagram of Possible Outcomes From EST Clustering. All ESTs in the same orientation, either sense (+) or anti-sense (-) will have a consensus sequence that is sense (A) and anti-sense (B), respectively. Clusters containing ESTs in various orientations (C and D) will have a consensus sequence with orientation based on the orientation of the longest EST in the cluster. **This information is used to design the probes on the array (i.e. if the consensus sequence is in the sense orientation, then probes will be designed according to the sense sequence (Figure 1). If the consensus sequence is in the anti-sense orientation, then probes will be designed according to the reverse-complement of the consensus sequence.**

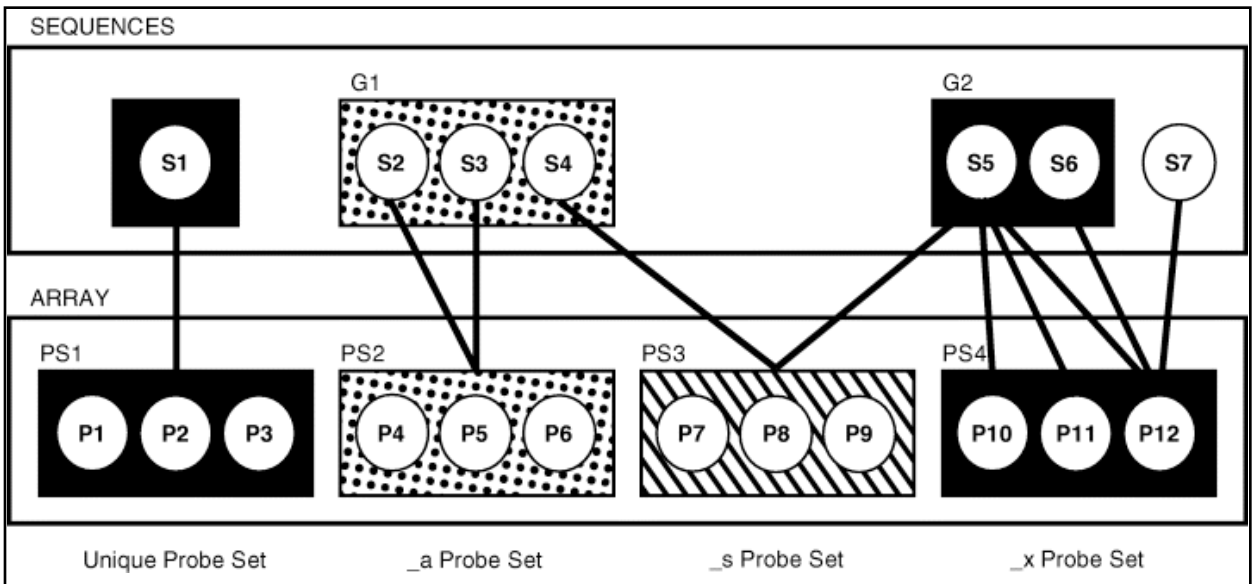

Figure 3. Possible Relationships Between Sequences And Probe Sets on an Array. G1, G2, etc., represent families of related gene sequences; S1, S2, S3, etc., represent individual gene sequences; PS1, PS2, PS3, etc., represent probe sets; P1, P2, P3, etc., represent probes. Solid lines show how sequence 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)).

Table 3. Description of probe set suffixes (refer to Figure 3)

SUFFIXES	DESCRIPTION
_at	A unique probe set directed to anti-sense transcript relative to cRNA. ALL probe sets on the array have an _at designation.
_a_at	A probe set that recognizes alternative transcripts from the same gene (**Figure 3**).
_s_at	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	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.**

D. Probe Set Distribution On The Array

D1. How many probe sets are represented on the soybean array?

There are 61,170 total probe sets on the soybean array. The breakdown of the array is summarized in **Table 4**. The different genomes were distinguished based on the probe set headers mentioned in **Table 1** except for rRNA. We filtered probe set IDs that contained rRNA and counted the number of probe sets representing *G. max* and *P.sojae*. There are no rRNA probe sets representing *H. glycines* on the array.

Table 4. Distribution of soybean array features

Genome	# Probe Sets
G. max	37,593
G. max (rRNA)	48
P. sojae	15,820
P. sojae (rRNA)	44
H. glycines	7,530
Controls	135
Total	**61,170**

Note: In all the analysis carried out below, **only probe sets for soybean are examined.** Probe sets from *P. sojae, H. glycines* and Affymetrix controls are excluded from the analysis.

Conclusions: More than 50% of probe sets on the array is used to interrogate soybean transcripts.

D2. How many probe sets are derived from Unigene or Affymetrix clusters?

From the 37,593 probe sets representing *G. max* sequences (**Table 4**), we further divided the probe sets into two categories: probe sets representing UniGene clusters as denoted by the "Gma" header (**Table 1**) and probe sets representing Affymetrix clusters as denoted by the "GmaAffx" header (**Table 1**). These results are summarized in **Table 5**.

Table 5. Distribution of *G. max* probe sets from UniGene and Affymetrix

Probe ID Containing	# Probe Sets	# Unique Clusters*
UniGene (Gma)	14,928	11,297
Affymetrix (GmaAffx)	22,665	19,633
Total	**37,593**	**30,930**

*Number of unique clusters represents the sum total of Unigene and Affymetrix de-novo cluster IDs only. Probe sets containing sub-cluster (e.g. Gma.12345.1.S1_at , Gma.12345.2.S1_at) or

different orientation of the same clusters (e.g. Gma.12345.1.A1_at) were counted as one unique cluster.

Note: The total number of unique clusters (30,930) does not truly represent the number of unique transcripts on the array.

Conclusions: UniGene and Affymetrix clusters represent 40% and 60% of probe sets, respectively.

D3. What is the representation of probe sets targeting unique sequences (_at), alternative transcripts (_a_at), gene families (_s_at), and others (_x_at)?

We examined in greater detail the representation of different suffixes (**Table 3**) to determine the number of probe sets representing unique transcripts and number of probe sets targeting multiple members, splice variants, etc. **Table 6** summarizes the distribution of different suffixes on the array.

Table 6. Detailed representation of different probe set suffixes on the array.

Probe ID suffixes*	**_a_at**		**_at**		**_s_at**		**_x_at**		**_at**		
Orientation*	A1	S1	A1	S1	A1	S1	A1	S1	A2	S2	
Unigene (gma)	83	1,048	3,137	8,980	217	1,002	32	427	0	2	**14,928**
Affymetrix (gmaAffx)	0	0	3,475	16,745	295	1,837	34	278	1	0	**22,665**
Total	**83**	**1,048**	**6,612**	**25,725**	**512**	**2,839**	**66**	**705**	**1**	**2**	**37,593**

Table 7. Simplified representation of different probe set suffixes on the array.

Probe ID suffixes*	**Gma**	**GmaAffx**	**Total**
_a_at	1,131	0	**1,131**
_s_at	1,219	2,132	**3,351**
_x_at	459	312	**771**
_at	12,119	20,221	**32,340**
Total	**14,928**	**22,665**	**37,593**

Conclusions: Approximately 89% of the probe sets on the array targets a unique sequence within an RNA population. Less than 11% of the probe sets targets members of gene families (3,351) and other ambiguous sequences (771).

E. Analysis of Unique Sequences on the Array Using CAP3

E1. How many UNIQUE sequences are represented on the array?

To address the issue of how many unique sequences are represented on the array, we carried out CAP3 (contig assembly program) to cluster all the target sequences from the probe set (37,593)

represented on the array to determine the number of contigs (cluster containing more than one sequence) and singletons (cluster containing exactly one sequence). The CAP3 program has been shown to correctly distinguish gene family members up to 96% identity. Therefore, the sum total number of CAP3 generated contigs and singletons should give us a good indication of the total number of unique sequences on the array.

Table 8. CAP3 Clustering Analysis Summary

	# Clusters*	# Probe Sets
# Singleton	28,821	28,821
# Contig	3,803	8,772
Total	**32,624**	**37,593**

* Number of clusters (contigs and singletons) generated from a CAP3 analysis.

Note: The total number of clusters (32,624) representing unique sequences is different from the total number of unique clusters (30,930) summarized in Table 5 that was based only on UniGene and Affymetrix cluster IDs. One explanation is that multiple probe sets containing the same UniGene or Affymetrix cluster IDs are counted as one unique sequence in Table 5 but are counted as multiple unique sequences in Table 8.

Conclusions: There are more than 32,000 unique sequences on the array. This does not equate to 32,000 unique proteins as many probe sets might represent different regions of the same transcript.

II. ANNOTATION OF THE SOYBEAN ARRAY (2004)

A. Strategy for the Annotation of the Soybean Array

There are 37,593 features on the soybean array and it would take considerable amount of time to identify what gene each feature represents in order to assign functional category. This was manually carried out for the *Arabidopsis* AtGenome1 and ATH1 genome arrays. We developed a strategy to expedite the annotation of the soybean array using the manually annotated *Arabidopsis* ATH1 genome array as a guide (**Figure 4**). First, soybean array sequences were BLASTed against all predicted *Arabidopsis* proteins to identify putative homologs and orthologs (28,498). Second, using the *Arabidopsis* protein ID, determine which proteins are represented on the ATH1 array (25,993). Third, functional category were assigned to soybean sequences that have a match to *Arabidopsis* proteins represented on the ATH1 array using the manually annotated categories from the ATH1 array. Sequences with no homology to an *Arabidopsis* protein were BLASTed against NCBI non-redundant (nr) protein database. Functional categories were assigned manually to the remaining sequences with homology to *Arabidopsis* proteins not represented on the ATH1 array (2,505) and sequences with homology to sequences in Genbank (766). (**Figure 4**).

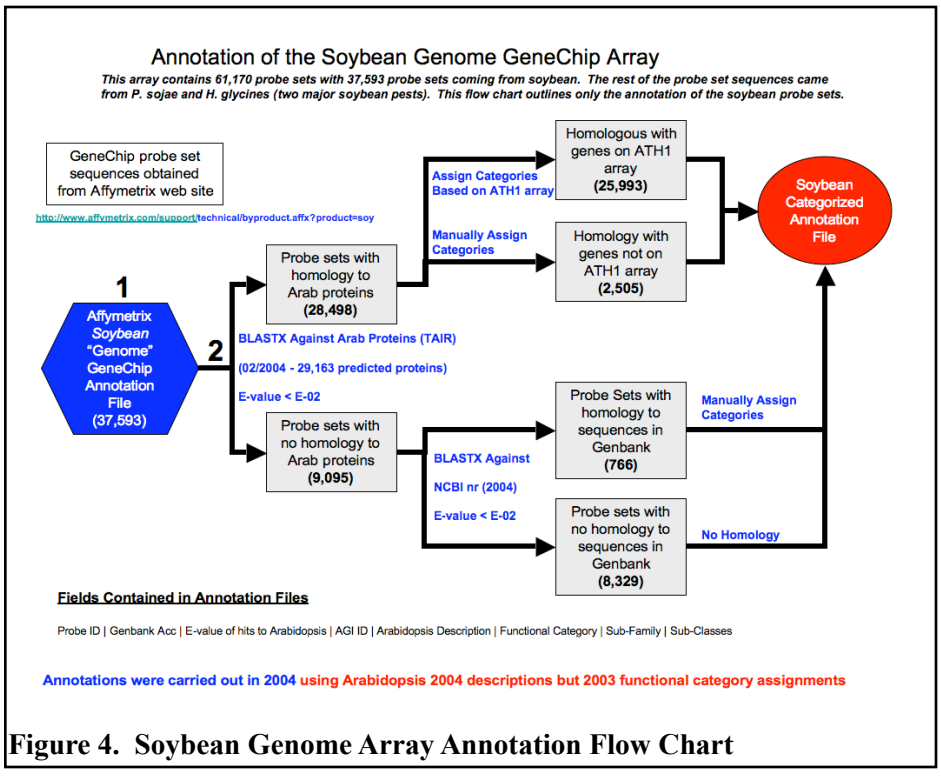

Figure 4. Soybean Genome Array Annotation Flow Chart

B. BLAST Analysis to Annotate Sequences on the Array

B1. How many soybean probe sets have homology to *Arabidopsis* proteins?

We determined the number of soybean probe sets with homology to *Arabidopsis* proteins by BLASTX analysis of the soybean consensus sequence (query) against the *Arabidopsis* protein database (**ATH1_pep_cm_20040228)** (subject). We filtered for soybean probe sets with homology to Arabidopsis protein (e-value <0.01). We took the top hit (i.e. *Arabidopsis* protein with lowest e-value match to the soybean sequence) for further analysis (**see Table 11**). The e-value e^{-02} was set arbitrarily to detect related proteins from *Arabidopsis* and might not be the most stringent criterion for sequence homology.

Table 9. BLASTX analysis against *Arabidopsis* protein database

BLASTX Results	# Probe Sets
Homology to *Arabidopsis* proteins (e-value < 0.01)	**28,498**
No Homology to *Arabidopsis* proteins*	**9,095**
Total	**37,593**

* This group includes probe sets with no hits to *Arabidopsis* proteins and probe sets with hits to *Arabidopsis* proteins with e-value > 0.01.

Conclusions: Approximately 76% of probe set sequences have homology to Arabidopsis proteins (e-value <0.01).

B2. How many soybean probe sets have no homology with any proteins in the public domain (Genbank)?

We identified 9,095 probe sets with no homology (e-value > 0.01) to any *Arabidopsis* proteins (**Table 9**). We want to know if these 9,095 probe sets have homology to any proteins in the NCBI nr protein database. We carried out BLASTX of the 9,095 sequences against the NCBI protein database and took the top hit (**see B1 above for definition**) for annotations.

Table 10. BLASTX analysis against GenBank non-redundant database

BLASTX Results	# Probe Sets
Homology to Genbank proteins (e-value < 0.01)	766
No Homology to Genbank proteins*	8,329
Total	9,095

* This group includes probe sets with no hits to any proteins in the public database (GenBank 2004).

Conclusions: Approximately 22% of soybean probe sets have no homology with any proteins in the public domain. These sequences might represent soybean-specific transcripts or transcripts of unknown function.

B3. How many soybean probe sets have homology to *Arabidopsis* genes represented on the ATH1 array?

In **section B1**, we identified 28,498 soybean probe sets with homology to *Arabidopsis* proteins with an e-value < 0.01. We compared the AGI locus ID of the top *Arabidopsis* hit to the AGI locus ID of genes represented on the *Arabidopsis* ATH1 array. There are 2,505 soybean probe sets that has homology to an *Arabidopsis* protein not represented on the *Arabidopsis* ATH1 Array. The remaining 25,993 probe sets were assigned functional category based on the categories assigned for the *Arabidopsis* counterpart. *Note: The 2,505 probe sets not represented on the Arabidopsis ATH1 Genome array and the 766 probe sets (previous section) with homology to a protein in GenBank were manually annotated and assigned into functional categories.*

Table 11. Soybean Probe Sets with Homology to *Arabidopsis* Proteins Represented on the *Arabidopsis* ATH1 Genome Array

	# probe sets
Homology to Proteins on *Arabidopsis* Array	**25,993**
Homology to Proteins not on *Arabidopsis* Array	**2,505**
Total	**28,498**

B4. Functional category assignment

Using *Arabidopsis* as a reference, we identified 25,993 probe sets on the soybean array that can be automatically assign a functional category based on the *Arabidopsis* counterparts on the *Arabidopsis* ATH1 array. There are an additional 2,505 probe sets that have homology to an *Arabidopsis* protein but the protein is not represented on the ATH1 array. Furthermore, we identified 766 probe sets that have homology to a protein in GenBank and 8,329 probe sets with no homology to any proteins in GenBank. The results are summarized in **Table 12**.

Table 12. Classification of soybean probe sets into functional categories

	# Probe Sets	Classification
Homology to Proteins on *Arabidopsis* Array	**25,993**	**Automated**
Homology to Proteins not on *Arabidopsis* Array	**2,505**	**Manual**
Homology to Proteins in GenBank	**766**	**Manual**
No Homology to any Proteins in GenBank	**8,329**	**Manual**
Total	**37,593**	

Conclusions: Approximately 69% of the probe sets on the array were automatically assigned functional category based on the Arabidopsis counterpart and another 9% were manually assigned functional category. The remaining 22% with no homology to any proteins in GenBank were assigned to the "No homology to any known protein" category.

C. Complete Annotation of the Soybean Array

The 37,593 probe sets on the Soybean GeneChip Array were assigned functional categories based on the EU *Arabidopsis* sequencing project (Bevan et al., 1999) (**Table 13 and Figure 5**). Approximately 31% of the probe sets (11,790) were classified collectively as unclassified while 22% of the probe sets (8,329) had no homology to any proteins in the NCBI non-redundant (nr) database (**Table 12**). Collectively, more than 50% of the sequences from the array are unknown or cannot be assigned a category. Additionally, we annotated 2,178 putative transcription factors and assign these probe sets into distinct transcription factor families (**Table 14** and **Figure 6**). Approximately 350 (16%) of those transcription factor probe sets could not be assign into a transcription factor family and was assigned as unclassified (**Table 14**).

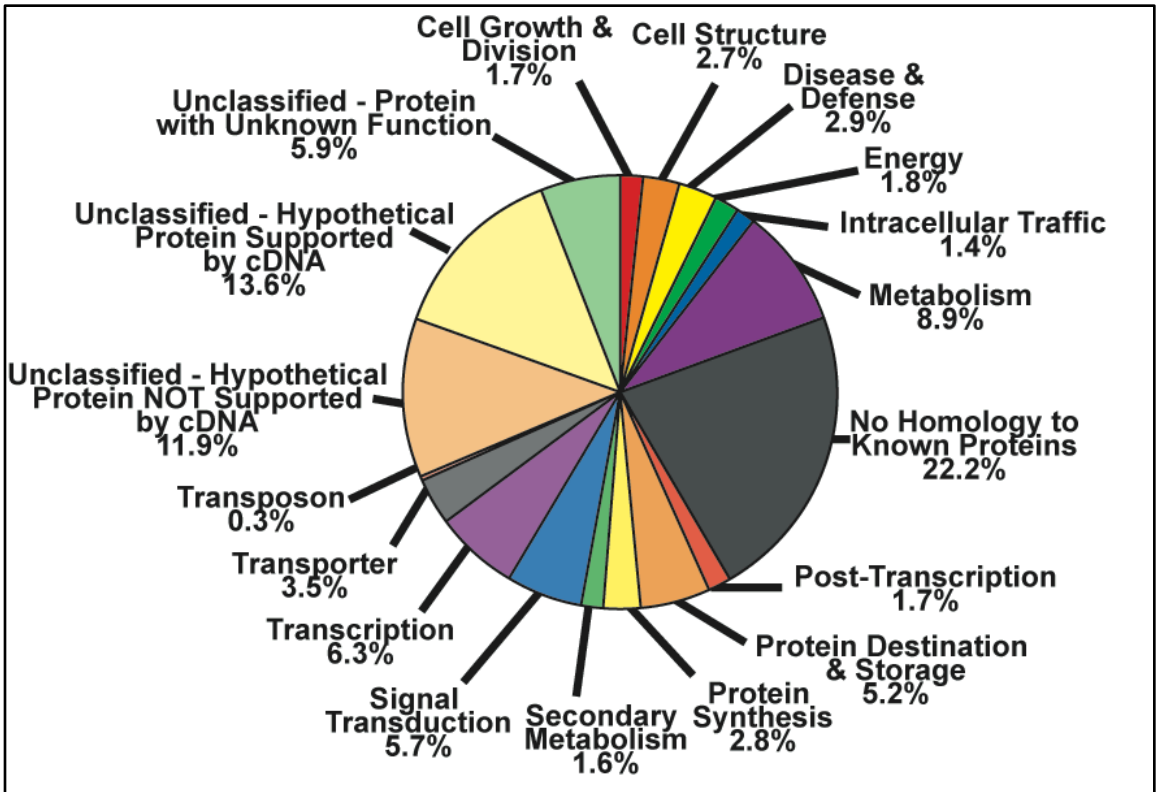

Figure 5. Distribution of All Probe Sets on the Soybean Array (2004).

Table 13. Distribution of 37,593 probe sets into functional categories

Functional Categories	Total	%
Cell Growth & Division	639	1.7
Cell Structure	1020	2.7
Disease & Defense	1081	2.9
Energy	687	1.8
Intracellular Traffic	533	1.4
Metabolism	3334	8.9
No Homology to Known Proteins	8329	22.2
Post-Transcription	643	1.7
Protein Destination & Storage	1959	5.2
Protein Synthesis	1042	2.8
Secondary Metabolism	616	1.6
Signal Transduction	2133	5.7
Transcription	2354	6.3
Transporter	1313	3.5
Transposon	120	0.3
Unclassified - Hypothetical Protein NOT Supported by cDN	4455	11.9
Unclassified - Hypothetical Protein Supported by cDNA	5103	13.6
Unclassified - Protein with Unknown Function	2232	5.9
Total	**37593**	**100.0**

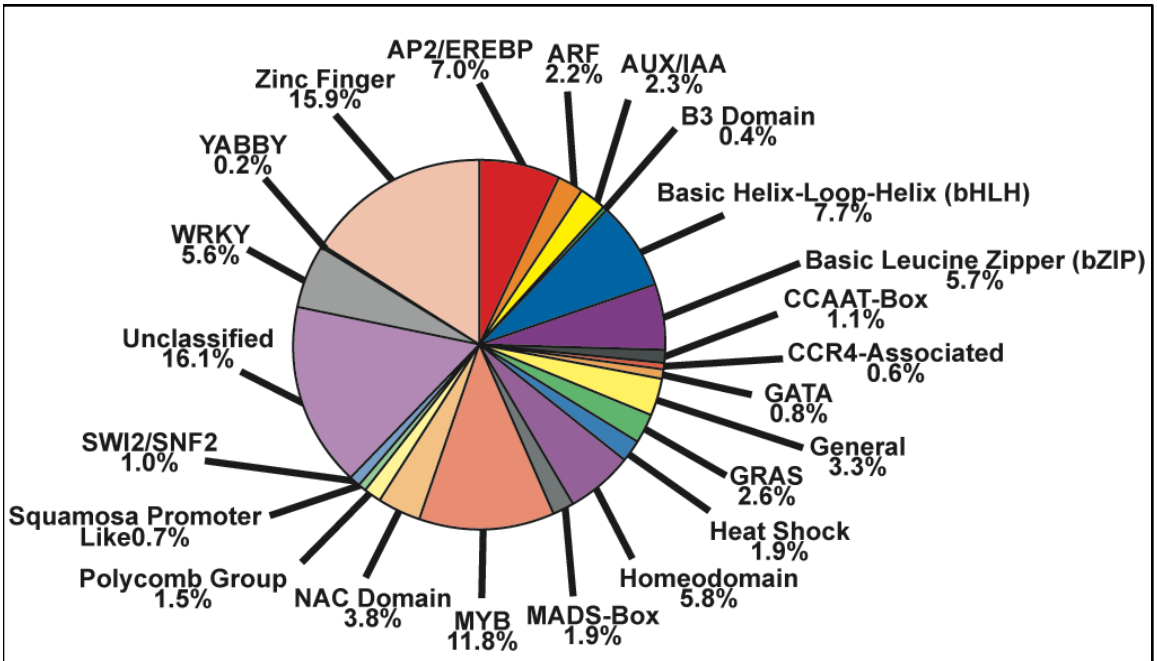

Figure 6. Distribution of All Transcription Factor Probe Sets on the Soybean Array (2004)

Table 14. Distribution of 2,178 probe sets into transcription factor families

Transcription Factor Families (2003)	Total	%
AP2/EREBP	153	7.02
ARF	49	2.25
AUX/IAA	51	2.34
B3 Domain	8	0.37
Basic Helix-Loop-Helix (bHLH)	168	7.71
Basic Leucine Zipper (bZIP)	125	5.74
CCAAT-Box	25	1.15
CCR4-Associated	12	0.55
GATA	18	0.83
General	71	3.26
GRAS	56	2.57
Heat Shock	42	1.93
Homeodomain	127	5.83
LIM Domain	1	0.05
MADS-Box	41	1.88
MYB	256	11.75
NAC Domain	82	3.76
Polycomb Group	33	1.52
Squamosa Promoter-Like	15	0.69
SWI2/SNF2	21	0.96
TCP	1	0.05
Unclassified	350	16.07
WRKY	121	5.56
YABBY	5	0.23
Zinc Finger	347	15.93
Total	**2178**	**100.00**

III. RE-ANNOTATION OF THE SOYBEAN ARRAY (2007)

A. Motivation for the Re-Annotation of the Soybean Array

The Affymetrix Soybean Genome GeneChip Array consists of 61,170 probe sets from which 37,593 probe sets belong to Soybean (*Glycine max*) and the remaining represent sequences from the soybean pathogens *Phytophthora sojae* and *Heterodera glycines*. The soybean sequences were obtained from public EST data derived from approximately 85 cDNA libraries representing most plant organs at different developmental stages (i.e seeds, seedlings, leaves, stems, roots, flowers, etc.). **See Sections I and II** for more detailed information.

The Soybean GeneChip Array sequences were annotated in 2004. (see **Section II**). During that process, more than 50% of the features (20,116) remained unclassified or had no homology to known proteins (**Table 13** and **Figure 5**). In the past three years, there has been an increase availability of genomic resources and sophisticated programs and tools developed to aid in the identification and classification of novel transcripts . In addition, several databases such as AGRIS TFDB (http://arabidopsis.med.ohio-state.edu/AtTFDB/) and Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/) have been created to classify the array of transcription factors and transcription factor families identified in both animals and plants as well as some that are plant-specific. Furthermore, classification of the soybean array was based on the *Arabidopsis* ATH1 array 2003 classification containing 2004 *Arabidopsis* descriptions. Therefore, this is a good time to revisit the soybean array annotation to generate a more accurate, complete, and up to date version.

B. Goals and Approaches

The goals of the re-annotation effort are to 1) update the probe set information available in the databases, 2) generate a consensus on the criteria used for the classification of each probe set, and 3) reduce the more than 50% probe sets unclassified or unknown. To achieved these goals, we took the following steps:

1) Re-annotate the Soybean GeneChip Array using updated information from TAIR (http://arabidopsis.org/) and TIGR Plant Transcript Assembly (TA) database: (http://plantta.tigr.org/)
2) Re-classify the 'Unknown' and 'Unclassified' soybean array probe sets based on the new annotation available.
3) Re-annotate and classify the 'Transcription Factor' probe sets using information from the Soybean transcription factor database (Soybean TFDB - http://planttfdb.cbi.pku.edu.cn/web/index.php?sp=gm).

4) Re-classify the entire array (37,593 probe sets) establishing a consensus classification for similar proteins.

C. Soybean Re-Annotation Efforts - A Step-By-Step Account

In this section, we will describe a step by step account of how we re-annotated the Soybean array as conceptualized by the flow chart in **Figure 7**.

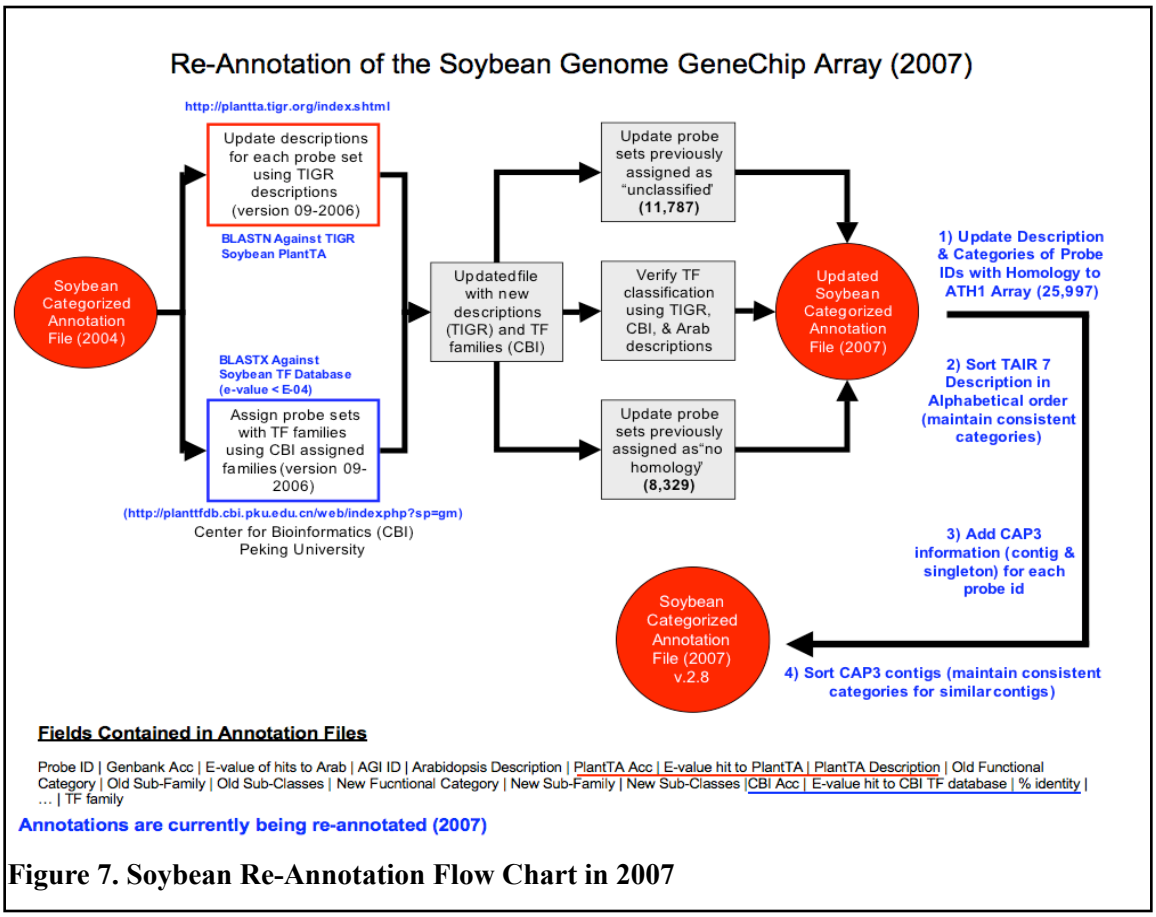

Figure 7. Soybean Re-Annotation Flow Chart in 2007

1. Update descriptions for each probe set using TIGR Plant TA descriptions

TIGR PlantTA is a database of transcript assemblies for many plant ESTs including soybean (http://plantta.tigr.org/index.shtml). For each PlantTA ID, there is an associated description based on BLAST analysis against UniProt UniRef database. In order to use the description, we need to determine which probe set sequence corresponds with each PlantTA ID for soybean. Target sequence for each probe set from the array was BLASTN against the TIGR Soybean PlantTA sequences. The TIGR soybean PlantTA sequences were downloaded from the PlantTA web site. We downloaded Glycine_max_release_2.fasta.zip, which is the current release for soybean generated in 2006-09-28. In this release there are 36,399 clusters and 80,566 singletons. Both files were imported to the fructose server (fructose.ribs.ucla.edu) where BLASTN analysis

was carried out. The associated PlantTA ID with homology to the soybean probe set is included in the annotation file along with results from the BLAST analysis (% identity, number of nucleotides matched, total number of nucleotides, etc...).

2. Identify putative transcription factors based on the Plant Transcription Factor Database (PlantTFDB).

The Center for Bioinformatics at Peking University created a database of putative transcription factors in soybean (http://planttfdb.cbi.pku.edu.cn/web/index.php?sp=gm). To-date, this is the only public database for soybean transcription factors available that I'm aware of. This database used the assembled transcripts data generated from the Plant Genome Database (PlantGDB) web site (http://www.plantgdb.org/) to identify putative transcription factors. Putative transcription factors were identified based on DNA binding domains that exists in Pfam. For transcription factor families without DNA binding domains, multiple alignment of sequences from Arabidopsis, rice, and poplar was used to generate a hidden markov model (HMM) profile for conserved domains that was used to search for putative TFs. The predicted TFs were obtained from transcripts assembled by PlantGDB on Sept 2006 (GenBank release 155).

The putative amino acid sequences from the soybean PlantTFDB was uploaded to the fructose server for BLASTX analysis. See part "D. Criteria for Functional Category Assignment" for information on what were annotated as TFs. The results from the BLASTX analysis was incorporated into the annotation file as noted in **Figure 7**.

3. Update the "Unclassified" & "No homology to Known Proteins" categories

We assumed that the probe set features previously assigned with a category other than "unclassified" or "no homology to known proteins" would not change categories over time. Therefore, we focus on re-classifying the probe sets previously assigned as "unclassified" or "no homology to known proteins". Probe sets were sorted based on the updated description from TIGR's PlantTA and 2004 *Arabidopsis* descriptions from the ATH1 array. See part "D. Criteria for Functional Category Assignment" for classification information.

4. Update the description and classification of probe sets with homology to the ATH1 array

The following step was to update the description and the classification of 25,993 probe sets with homology to the ATH1 array. The new *Arabidopsis* description from the ATH1 array was obtained from TAIR's description of the Affymetrix (updated May 2007 using TAIR version 7). Using the AGI ID of the *Arabidopsis* proteins previously determined in **Section II**, we matched the AGI ID to the IDs on the ATH1 array and merged the new description and functional category information to the soybean annotation file. The newly merged Soybean annotation file was then sorted by the *Arabidopsis* description (2007) and the 25,993 probe sets with an *Arabidopsis* homolog were classified automatically according to the ATH1 categories. In the updated ATH1 annotation file,

more than one AGI ID can be associated with a single probe set. Therefore, we identified 61 new soybean probe sets with an equivalent ortholog on the ATH1 array (compared to the 2004 annotation). Therefore, the updated number of soybean probe sets that hit the ATH1 array is **26,054**. The remaining 11,539 probe sets were classified using all the descriptions available from the sources described in the next section.

5. Inclusion of sequence assembly information in the annotation file

Sequences of the soybean array were assembled into contigs and singletons using the CAP3 program (**see Section II**). This contig assembly information was added to the annotation file. To classify the contigs in a consistent manner, the different contigs were sorted by they number and then a single category was assigned to all the probe sets that are included in an individual cluster.

D. Criteria for Functional Category Assignment

The following criteria were used to assign functional category for features on the *Arabidopsis* ATH1 and soybean GeneChip arrays. Overall, there are no general rule that can be applied to every probe sets. However, we try to be consistent in the assignment of functional categories.

1. Probe sets classification was determined using the following sources of information:
    a. TAIR description updated in 2007 (TAIR version 7)
    b. Comparison with descriptions from TAIR (2004) and TIGR PlantTA(2006).
    c. Gene Ontology information obtained from the TAIR web page (http://www.arabidopsis.org/).
    d. Information about protein domains from PFAM 22.0 database (2007) (http://pfam.janelia.org/).
    e. Search for additional information on the web (InterPro, Wikipedia, PubMed).

2. The categories were assigned based on 1) molecular function or 2) biological process in that order of importance.

3. The transcription factor category was assigned based on the following:
    a. First, we filtered sequences with a hit to CBI soybean TF database (e-value < e^{-04}).
    b. Second, we examined the e-value and the % sequence identity for each hit.
        i. Sequences with 90-100% identity alignment and low e-value (< e^{-04} to e^{-10}) were classified as TFs. *Note: Sequences with e-values close to e^{-04}, on average, had non-specific match to TFs and were not classified as TFs.*
        ii. Sequences with low e-values (< $e^{-10\ to}$ e^{-20}) but had insignificant alignment along the entire transcription factor proteins were NOT

classified as TFs. *Note: This situation occurred rarely and mainly due to repetitive amino acids from large TF proteins such as HMG or LEU.*

c. Sequences with homology to *Arabidopsis* TFs but are not identified in the CBI soybean TF database were also assigned as TFs.

d. Sequences with descriptions pertaining to TF domains (e.g. MYB domain containing protein, bZIP-containing protein, etc) were classified as TFs.

4. Proteins with known domains with general function (e.g. WD40) were placed under Unclassified.

5. Protein kinases were assigned signal transduction unless additional information can place the protein in another category. For example, Armadillo-related proteins were also placed under signal transduction.

6. Apoptosis related proteins were placed under cell growth and division.

7. Descriptions related to electron transport, redox, PSI, PSII, glycolysis and associated processes were placed under energy.

8. Cytoskeleton component associated to spindles were placed under cell growth and division. Other cytoskeleton components such as actin and myosin were placed under Cell Structure. Cellulose and cell wall related sugars are classified as cell structure. All other sugars are placed under metabolism.

9. Histones were placed under cell structure, but histone modification proteins were placed under transcription.

10. Proteins were placed under transporters if they are associated to membrane and transport of any component from one side of the membrane to another. Proteins associated with SNAREs, ER, Golgi, and vesicles were placed under intracellular trafficking. Chaperone, heat shock proteins, proteases, ubiquitin and proteasome related were all placed under protein destination & storage.

11. An ambiguous description was left as Unclassified (e.g. DNA-binding, helicase, etc).

E. Re-Annotation Results of the Soybean Array

Results from the re-annotation of the Soybean GeneChip Array are summarized in **Table 16** and **Figure 7. Table 16** shows the distribution of 37,593 probe sets assigned into 16 functional categories according to the 2004 annotation. Previously, the unclassified category was divided into three groups (Unclassified - hypothetical proteins with no cDNA support, Unclassified - hypothetical proteins with cDNA support, and Unclassified - proteins with unknown function). However, the soybean array was designed based on EST sequences and therefore the "Unclassified - hypothetical proteins with no cDNA support" category no longer applies. Furthermore, for simplicity, we decided to group the remaining unclassified groups into one unclassified category.

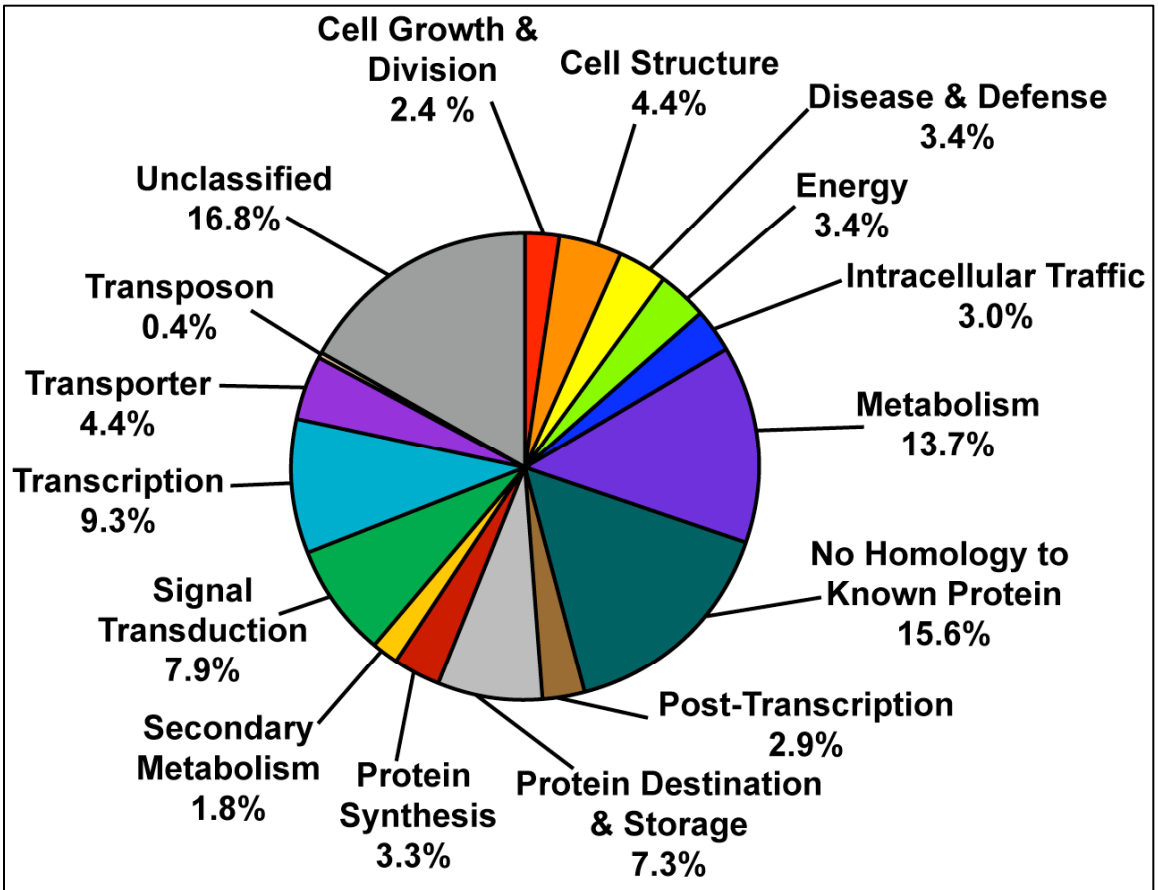

Figure 8. Distribution of All Probe Sets on the Soybean Array (2007).

Table 15. Distribution of all probe sets on the soybean array (2007)

Functional Categories	Total	%
Cell Growth & Division	892	2.4
Cell Structure	1638	4.4
Disease & Defense	1272	3.4
Energy	1289	3.4
Intracellular Traffic	1132	3.0
Metabolism	5138	13.7
No Homology to Known Proteins	5879	15.6
Post-Transcription	1099	2.9
Protein Destination & Storage	2736	7.3
Protein Synthesis	1227	3.3
Secondary Metabolism	695	1.8
Signal Transduction	2952	7.9
Transcription	3508	9.3
Transporter	1668	4.4
Transposon	144	0.4
Unclassified	6324	16.8
Total	**37593**	**100.0**

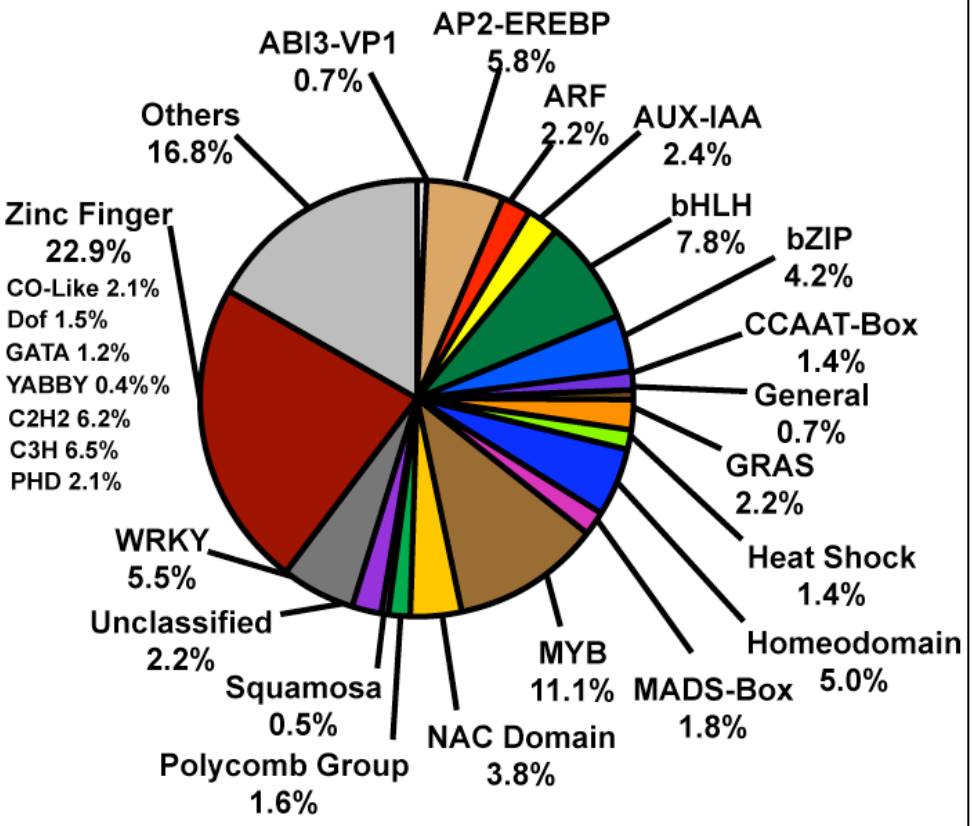

Figure 9. Distribution of All Transcription Factor Probe Sets on the Soybean

Table 16. Distribution of probe sets into transcription factor families.

Transcription Factor Families	Total	%	Transcription Factor Families	Total	%
ABI3-VP1	20	0.71	HMG	25	0.88
Alfin	17	0.60	HRT	0	0.00
AP2-EREBP	164	5.79	HSF	41	1.45
ARF	62	2.19	JUMONJI	35	1.24
ARID	13	0.46	LFY	1	0.04
ARR-B	16	0.56	LIM	21	0.74
AS2	19	0.67	LUG	9	0.32
AUX-IAA	67	2.37	MADS	51	1.80
BBR-BPC	10	0.35	MBF1	6	0.21
bHLH	220	7.77	MYB	202	7.13
bZIP	118	4.17	MYB-related	111	3.92
BZR-BES1	13	0.46	NAC	108	3.81
C2C2-CO-like	59	2.08	Nin-like	6	0.21
C2C2-Dof	43	1.52	NZZ	4	0.14
C2C2-GATA	33	1.17	PcG	46	1.62
C2C2-YABBY	11	0.39	PHD	61	2.15
C2H2	179	6.32	PLATZ	13	0.46
C3H	184	6.50	Pseudo ARR-B	15	0.53
CAMTA	16	0.56	S1Fa-like	6	0.21
CCAAT-Dr1	3	0.11	SAP	0	0.00
CCAAT-HAP2	11	0.39	SBP	15	0.53
CCAAT-HAP3	11	0.39	Sigma70-like	15	0.53
CCAAT-HAP5	14	0.49	SRS	4	0.14
CPP	8	0.28	TAZ	7	0.25
CSD	4	0.14	TCP	20	0.71
E2F-DP	8	0.28	Trihelix	37	1.31
EIL	17	0.60	TUB	23	0.81
FHA	14	0.49	ULT	1	0.04
G2-like	60	2.12	Unclassified	62	2.19
GeBP	8	0.28	VOZ	5	0.18
General	20	0.71	Whirly	5	0.18
GIF	5	0.18	WRKY	156	5.51
GRAS	63	2.22	ZF-HD	19	0.67
GRF	15	0.53	ZIM	34	1.20
HB	143	5.05	**Total**	**2832**	**100.00**

There are 2,832 probe sets identified as putative transcription factors in the 2007 re-annotation efforts. These transcription factors are classified into 69 transcription factor families according to the classification from the soybean TFDB (**Table 16**). Zinc finger is the largest transcription factor family represented on the array with 23% of the transcription factor probe sets (**Figure 9**). However, the zinc finger transcription factors are now represented by 13 sub-families, and the highly represented sub-families are CO-like (2.1%), Dof (1.5%), GATA (1.2%), YABBY (0.4%), C2H2 (6.2%), and C3H (6.5%). The MYB transcription factors are now divided into two families (myb and myb-related) that together represent the second transcription factor family on the array with 11% of the transcription factor probe sets. Other important transcription factor groups such as bHLH, bZIP, and homeodomain remain well represented in the array as shown in **Table 16** and **Figure 9**. Sixty-two transcription factor probe sets (2.2%) could not be assign a transcription factor family and were classified as unclassified.

F. Comparison of the 2004 and 2007 Annotation Pies

One of the reasons to re-annotate the soybean array was to reduce the more than 50% probe sets that were unclassified or unknown. In fact, 'Unclassified' and 'No homology to Known Proteins' were the only categories with a significant reduction in the number of

probe sets (46% and 29% respectively) after re-annotating the soybean array (**Table 17**). The remaining categories increased their representation on the array. The most significant increases were the intracellular traffic (112%), energy (88%), and post-transcription (71%) categories. Transcription and signal transduction changed around 50% and 40%, respectively (**Table 17**).

Table 17. Changes in Probe Set Distributions

Functional Categories	2004	2007	% Change
Cell Growth & Division	639.0	892	39.6%
Cell Structure	1020.0	1638	60.6%
Disease & Defense	1081.0	1272	17.7%
Energy	687.0	1289	87.6%
Intracellular Traffic	533.0	1132	112.4%
Metabolism	3334.0	5138	54.1%
No Homology to Known Proteins	8329.0	5879	-29.4%
Post-Transcription	643.0	1099	70.9%
Protein Destination & Storage	1959.0	2736	39.7%
Protein Synthesis	1042.0	1227	17.8%
Secondary Metabolism	616.0	695	12.8%
Signal Transduction	2133.0	2952	38.4%
Transcription	2354.0	3508	49.0%
Transporter	1313.0	1668	27.0%
Transposon	120.0	144	20.0%
Unclassified	11790.0	6324	-46.4%
Total	**37593**	**37593**	

The 2004 soybean array had 11,790 probe sets that were assigned the "unclassified" category. **Table 18** shows the distribution of these probe sets into functional categories with 56% of the probe sets being assigned a new category and 44% of the probe sets remained as "unclassified". Notably, around 900 probe sets (7.7%) became part of transcription and approximately 700 probe sets (5.6%) were classified as signal transduction. This result shows that a significant progress was made in terms of sequence annotation and information update on the several databases used for this analysis. However, little progress was made related to the annotation and classification of novel proteins. From the 8,329 probe sets classified as 'Unknown' in 2004, approximately 71% remained as such in 2007 and 12% were moved to 'Unclassified' (**Table 19**). Therefore, only 17% of these probe sets acquired a new category with 300 probe sets (3.5%) assigned to transcription and 139 probe sets (1.7%) assigned to signal transduction.

Table 18. Distribution of "unclassified" probe sets (2004) into functional categories.

Functional Categories	Total	%
Cell Growth & Division	324	2.7
Cell Structure	421	3.6
Disease & Defense	282	2.4
Energy	397	3.4
Intracellular Traffic	457	3.9
Metabolism	1427	12.1
No Homology to Known Proteins	0	0.0
Post-Transcription	318	2.7
Protein Destination & Storage	784	6.6
Protein Synthesis	132	1.1
Secondary Metabolism	89	0.8
Signal Transduction	657	5.6
Transcription	909	7.7
Transporter	381	3.2
Transposon	25	0.2
Unclassified	5184	44.0
Unknown	3	0.0
Total	**11790**	**100.0**

Table 19. Distribution of probe sets with no homology (2004) into functional categories.

Functional Categories	Total	%
Cell Growth & Division	53	0.6
Cell Structure	101	1.2
Disease & Defense	41	0.5
Energy	53	0.6
Intracellular Traffic	64	0.8
Metabolism	281	3.4
No Homology to Known Proteins	5879	70.6
Post-Transcription	64	0.8
Protein Destination & Storage	124	1.5
Protein Synthesis	66	0.8
Secondary Metabolism	29	0.3
Signal Transduction	139	1.7
Transcription	295	3.5
Transporter	103	1.2
Transposon	4	0.0
Unclassified	1033	12.4
Total	**8329**	**100.0**

G. Number of Genes Active in a Single Compartment -- Globular Stage Embryo Proper

We wish to determine an estimate for the number of genes active within a seed compartment. We used the globular stage embryo proper as an example. The goal is to determine the number of probe sets detected within a compartment and find the distribution of different probe set suffixes that might over- or under-estimate the count of genes active within a compartment.

Table 20. Distribution of probe sets detected in globular stage embryo proper.

Suffixes	# Probe Sets (Glob EP)	% Total	# Probe Sets* (Entire Array)	% Total
_at	**14,169**	**83.4%**	**32,340**	**86.0%**
_a_at	**697**	**4.1%**	**1,131**	**3.0%**
_s_at	**1,715**	**10.1%**	**3,351**	**8.9%**
_x_at	**417**	**2.4%**	**771**	**2.1%**
Total	**16,998**		**37,593**	

***** Numbers obtained from **Table 7**.

Table 21. Distribution of TF probe sets detected in globular stage embryo proper.

Suffixes	# TF Probe Sets (Glob EP)	% Total	# TF Probe Sets (Entire Array)	% Total
_at	928	82.1%	2,433	85.9%
_a_at	54	4.8%	100	3.5%
_s_at	108	9.6%	213	7.5%
_x_at	39	3.5%	86	3.1%
Total	1,129		2,832	