The Journal of Animal & Plant Sciences, 24(1): 2014, Page: 179-189
ISSN: 1018-7081

MICROARRAY GENE EXPRESSION MIAME-PLANT SUPPORTIVE TAB DELIMITED DATA FORMAT: MAGE-TAB-PLANT

S. U. Rehman, Atif A. Mirza[1], K. Masood[1], B. Rashid[1] and T. Husnain[1]

Department of Poultry Science, University of Agriculture, Faisalabad, Pakistan
[1]National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan
Corresponding Author Email: shahidurrehman@uaf.edu.pk

ABSTRACT

The goal of Minimum Information about a Microarray Experiment (MIAME) is to outline the minimum information that can clearly state the microarray based gene expression experiment in such a way that the details of experiment design, description and sample annotation can help to check and reproduce the results in some other laboratory. MIAME/plant complements the MIAME standard with guideline for experiment design, description and sample preparation protocols using controlled vocabularies. MAGE-TAB-Plant format has been structured on the similar lines as MAGE-TAB format. It explains the main chain of all events in a typical Microarray based experiment such as steps involved in preparation and handling of samples, labelling and hybridization protocols, scanning of images, normalization of raw data and analysis of images. The physical entities in this chain are Biosource, Sample, Extract, labelled Extract, hybridization, scan image, raw, normalized and combined data matrix. It is hoped that little more effort may result in better precisely replicated, analyzed, and better interpreted reproducible plant based microarray experiments. Data mining capabilities of data submitted using MAGE-TAB-Plant format may generate breakthroughs in research and the use of controlled vocabularies may encourage efficient sharing and annotation of the experiment data.

Key words: Microarray, Data format, Tab delimited, MIAME

INTRODUCTION

The goal of Minimum Information about a Microarray Experiment (MIAME) is to outline the minimum information that can clearly state the microarray based gene expression experiment in such a way that the details of experiment design, description and sample annotation can help to check and reproduce the results in some other laboratory. These details of experiment can vary significantly, the effort of the MIAME guidelines is to record the core details which are common to most experiments. Numerous software tools and public databases have adopted MIAME guidelines to handle and store microarray based datasets (Brazma *et al.,* 2001). These public repositories like Gene Expression Omnibus, developed by National Centre for Biotechnology Information, the DNA Data Bank of Japan and the European Bioinformatics Institute's Array Express also provide the support to the researchers about the format on which microarray based experiments' data can be submitted even array express also provide tools to convert the Microarray data on the guidelines of MIAME.

MIAME standard has been very successful in the past and is widely adopted by many scientists as increasing number of scientific journals require data regarding microarray based experiments to be submitted using this standard. This has also facilitated the grouping of microarray based studies on the basis of standard categories of MIAME i.e. technology platform, labeling and hybridization procedures, measurement data, and array design. But the experiment description which is submitted in the form of free text remained controversial because it is very much difficult to use the same yard stick to measure highly variable domain specific details of experiment annotation. Moreover most of the software tools and public repositories that handle and store microarray based data emphasize on hybridization and normalization protocols. To have reproducible microarray experiment data especially with reference to plant based studies experiment details like how were the plants grown (light intensity, light duration, rain fall etc.), soil was natural or artificial what were the nutrient levels in the soil, which plant organs were used and what was the age of the plant at the time of sampling etc. is more important. Structured storage of such information on domain specific basis is requird and MIAME/Plant, MIAME/Env and MIAME/Tox aresuch extensions (Anonymous, 2004; Zammerman *et al.,* 2006; Field *et al.,* 2005; Bao *et al.,* 2005). MIAME/plant complements the MIAME standard with guideline for experiment design, description and sample preparation protocols using controlled vocabularies.

Gene Expression Specifications: Another objective of MIAME was to guide the development of microarray databases and data management software. Rayner *et al*. (2006) proposed a simple tab-delimited, spread sheet based format, MAGE-TAB, which is a part of the MAGE microarray data standard and is used to annotate and communicate microarray data in a MIAME compliant fashion. MAGE-TAB is helpful for laboratories without bioinformatics experience or support to manage, exchange and submit, well annotated data in a standard format using a spreadsheet. Moreover the public repositories like Gene Expression Omnibus and MIAME Express at EBI Array Express etc. also support the spreadsheet based data submission. But the main limitation with MAGE-TAB and even the web based Microarray Data Submission System of Array Express is that it mainly emphasizes on the data submission related to experiment design and array design but information about experiment design, description and sample annotation is compromised especially with reference to the Plant based microarray experiments.

AtMIAM Express was an open-source Web-based application for the submission of Arabidopsis-based microarray data to Array Express. AtMIAM Express was developed primarily to providea data submission tool for the Compendium of Arabidopsis Gene Expression (CAGE) project which was merged to MIAM Express at Array Express (Mukherjee *et al.*, 2005). Currently no independent software tool is available that can handle/store MIAME/plant based microarray experiment data.

MATERIALS AND METHODS

The guidelines proposed in MIAME/plant are the base of the current proposed data format for Microarray Gene Expression Data. Plant genomics laboratory at National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan is endeavoring for stress responsive genes finding in mainly Cotton crop using cDNA Microarray technology. The current study was planned in the Bioinformatics laboratory of National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan to help the scientists working in Plant genomics laboratory by developing a Laboratory Information Management System to handle Microarray gene expression data using MIAME/Plant guidelines as a Ph.D. Research Thesis. As there is no specific format of data submission on MIAME/Plant guidelines so a tab delimited format for data submission was proposed keeping MAGE-TAB format as guideline which is well adapted format at Array Express – the Microarray data repository of EBI and even Gene Expression Omnibus of NCBI also encourages a spreadsheet based data submission. The current publication discusses the proposed MAGE-TAB-Plant format.

The aim of current study was to design tab delimited format for microarray data submission. MAGE-TAB-Plant format has been structured on the similar lines as MAGE-TAB format. Moreover it has similarly structured Investigation Description Format (IDF), Array Design Format (ADF), Sample and Data Relationship Format (SDRF), Raw and Processed data files. SDRF is the file which stores the main information about the experiment by having the reference to various sample annotation files and protocol files used in the experiment (Fig.1). It explains the main chain of all events in a typical Microarray based experiment such as steps involved in preparation and handling of samples, labelling and hybridization protocols, scanning of images, normalization of raw data and analysis of images. Organization of the various elements related to the SDRF are recorded in more detail with specific plant related parameters (Fig. 2). The physical entities in this chain are Biosource, Sample, Extract, labelled Extract, hybridization, scan image, raw, normalized and combined data matrix are shown by dotted grouped elements. This is almost the same as described by MAGE-TAB format except the biosource is made as part of this chain. The other elements in the chain of events are termed as ‘protocols’ in both the formats except that in the current proposed format the sample annotating protocol events like biomaterial manipulation, treatment, sample pooling and separation technique are re-structured with more details. The protocol having the details of extraction method (both kit and published reports is stored) and amplification protocol are required for the annotation of extract entity of the chain. Labelling and hybridization protocols are required for annotation of both labelled extract and hybridization event respectively. Scanning protocol, scanned image, image analyses are grouped with the raw data file. Normalization protocol and normalized data are grouped to represent the combined data matrix.

MAGE-TAB-Plant format data storage starts with the basic information about the experiment, scientist(s), publication and database etc. The detailed attributes of the IDF are shown in Table 1. The biosource properties are stored and the dependent files are selected in the pattern explained by Directed Acyclic Graph (DAG) in Figure 5. The sample annotation is represented by the DAG in Figure 6. The DAG for extract, labelled extract, hybridization and data files are explained in Figures 7 and 8 respectively. It is anticipated that this logical pattern of files may facilitate the scientists to submit the microarray based experiment’s data. Moreover these tab delimited files

can guide the scientists as template for recording the experiment data.

RESULTS AND DISCUSSION

Like MAGE-TAB format MAGE-TAB-Plant format is structured on the same back bone of mainly four types of files namely, Investigation Description Format (IDF), Array Design Format (ADF), Experiment Description Format (EDF) and Raw and Processed data files along with files for structured annotation of various physical object in the experiment/investigation (where needed). The major change is that in IDF file the reference of EDF file is given replacing the reference of Sample and Data Relationship Format file (described in MAGE-TAB format). EDF is the file which stores the main information about the experiment by having the reference to various sample annotation files and protocol files used in the experiment. Figure 1 explains the main chain of events in a typical Microarray based experiment. The physical entities in this chain are Biosource, Sample, Extract, labeled Extract, hybridization, scan image, raw, normalized and combined data matrix. Which is almost same as described by MAGE-TAB format except the biosource is made a part of this chain in MAGE-TAB-Plant format. The other elements in the chain of events in Figure 1 are termed as protocols in both MAGE-TAB as well as MAGE-TAB-Plant format except that in the current proposed format the sample annotating protocol events like biomaterial manipulation, treatment and separation technique are structured and detailed, while the protocol having the details of sample pooling can be linked to sample as well as extract depending upon the stage of pooling of the sample, mostly the sample pooling is done at extract level so in this format sample pooling, extraction method (both kit and publication reference is stored) and amplification protocol are required for the annotation of extract entity of the chain. Labeling protocol and hybridization protocol are required for annotation of both labeled extract and hybridization event respectively. In hybridization protocol the composition of various solutions is stored in structured manner similarly information about binding agents, washing procedures, quantity of labeled target, hybridization time, concentration, volume, temperature, instrument and protocol description is recorded. Scanning protocol, image analysis algorithm as well as instrucment and software description and various parameters like scanning protocol id, name, protocol, hardware, software, laser power, spatial resolution, pixel space and photomultplier tube voltage etc. are recorded for the annotation of raw data file. The storage of scanned image is optional however if required scan protocol also stores the names of scanned images related to each hybridization. Normalization protocol having information about normalization algorithm, normalization type /strategy and protocol annotates the normalized data file related to each hybridization and finally the combined data matrix is stored in a similar manner of MAGE-TAB. The main emphasis of this format is that it should provide precise structured information about experiment, experimental protocols and sample annotation along with machine readability.

Biosource used in the Experiment is described by Biosource Description Format which make a tab delimited file/spreadsheet having references to the various properties of the source material used in the experiment (Figure 2). The structured attributed stored by Biosource and the sub files are listed in Table1. Biosource ID is referred in EDF and it creates the reusability of the biosource in other experiments hence reduce the chances of data redundancy in case of database entry. Biosource ID changes with any change in dependent tables/properties of the material and takes a new name. The dependent properties also bear unique IDs with unique Names and can be reused in different Biosource IDs. All Biosource related files or any one or any combination of them can be used to define the biosource depending upon the Biomaterial used in the experiment in a structured manner and were ever additional information is required to be stored it can be stored in columns like description.

Biomaterial manipulation is the protocol or the manipulation which are made intentionally or unintentionally to the biosource before being selected as sample, these may be the growth substrates which are used for the plant materials' growth in presampling period it may be the environmental conditions in before the sampling or it may be the harvesting conditions before samples being collected recorded in order to enhance the reproducibility of the experiment. Figure 3 shows the major components of the biomaterial manipulation file, and Table 1 enlists the various sub and sub sub files and their stored attributes. Biomaterial manipulation ID and Name columns uniquely defines a set of conditions in which the sample was collected this can be reusable in the EDF where the ID is referred to define the Biomaterial Manipulation conditions. The Growth Substrates ID reference in the Biomaterial Manipulation file is another set of conditions which defines the various growth agents grouped by their physical properties like Liquid, solid, soil and Aeroponics etc. (Figure 4). The General Environment file stores the general environmental conditions which can be common to all growth environments along with the reference to the Green house environment, cell culture environment, field environment and growth control agents (Figure 2).

Figure 2 shows the information regarding the treatment applied to the samples it stores ID, Name, Description along with the references to the IDs of abiotic treatment, biotic treatment, stress treatment, seed stratification treatment, seed sterilization treatment and vernalization treatment etc. Abiotic treatment file stores the attributes like, ID, Name, Description, Temperature, Mechanical, atmospheric pressure, osmotic pressure, water, plant nutrient, chemical type, chemical amount and reference to light ID. The light file stores information in the columns of ID, Name, photoperiod, light, wave length, light intensity and description etc. Biotic treatment stores information in the attributes of ID, Name, description and references of IDs of Pathogen, animal and other plants etc. Pathogen file contains information regarding ID, Name, organism, organism type, strain, dosage and incubation conditions. Animal file stores ID, Name organism and type of effect. Other plant file handles information about ID, Name, concurrence, shading and plant parasites. Stress treatment file attributes to ID, Name, description, type of stress, degree of stress, timing of stress and stage of growth of plant. Seed stratification treatment file stores information about ID, Name, description, temperature, hormonal, duration, humidity, intensity and method. Seed sterilization treatment holds information regarding ID, Name, description and method. Vernalization treatment file stores information like ID, name, description, temperature, length of vernalization and growth environment.

Various protocols like separation technique, sample pooling, extraction method, extraction method kit, extraction method publication, labeling protocol, hybridization, protocol, scanning protocol, scan image, image analysis are also stored in structured format tamplates for which are available. ADF, Raw, normalized and combined data matrix are stored in the format already described in MAGE-TAB (Rayner *et al.*, 2006).

The format proposed is more complex than already defined by MAGE-TAB but is simpler than the MAGE-OM and MAGE-ML specifications and this requires no bioinformatics or MAGE-OM or MAGE-ML knowledge for the scientist to submit the data involving microarray experiments involving plants according to MIAME-Plant guidelines. This format helps logical storage of data which makes it easier for the researcher to follow. The templates of this format can be used as Laboratory information management system as the researcher can fill in the information during the course of the experiment. A laboratory information management system is already developed in National Centre of Excellence in Molecular Biology using the MAGE-TAB-plant format. It is hoped that little more effort may result in better precisely replicated, analyzed, and better interpreted reproducible plant based microarray experiments. Data mining capabilities of data submitted using MAGE-TAB-Plant format may generate breakthroughs in research and the use of controlled vocabularies may encourage efficient sharing and annotation of the experiment data. The proposed MAGE-TAB-Plant format is not the end, laboratories however, can collect additional information using this format for better interpretation of experiments.

Table 1 Data Variables for Sample annotation stored in various files of MAGE-TAB-Plant Format

Stem Protocol File	Sub Protocol File	Sub Sub Protocol File	Data Variables
Biosource		-	biosource ID, Name, organism, contact details, Germplasm ID, Ecotype ID, mutant ID, transgenic ID, starting material ID, development stage ID and organism part ID
Biosource	Germplasm	-	Germplasm ID, Name, Centre, genus, species, accession number, subspecies, cultivar, strain, genotype, haplotype, polymorphism, allele
Biosource	Ecotype	-	Ecotype ID, Name, habitat, location, date, collector name
Biosource	Mutant	-	Mutant ID, Name, Mutagene, locus, mode of inheritance.
Biosource	Transgenic	-	Gene ID, Transgene Name, Gene name, insert type, construct type, transgene type, promoter, reporter, selection marker, vector name, vector accession number
Biosource	Starting material	-	Starting material ID, Name, type, source of starting material, age, sex, disease type, additional clinical information, tissue culture, cell culture, protoplast
Biosource	Development stage	-	Development stage ID, Name, description, reference
Biosource	Organsim part	-	Organism part ID, name, description, targeted cell type, isolation protocol, reference

Biomaterial Manipulation			Biomaterial manipulation ID, Name, description, Growth substrates ID, general environment ID, harvesting conditions ID
Biomaterial Manipulation	Growth Substrates		Growth substrates ID, Name, Description, liquid ID, solid ID, soil ID, aeroponics ID
Biomaterial Manipulation	Growth Substrates	Liquid	Liquid ID, Name, hydroponics, constituents, sterilization
Biomaterial Manipulation	Growth Substrates	Solid	Solid ID, Name, Agar, Filter paper, Nylon Membrane, quartz sand, sterilization
Biomaterial Manipulation	Growth Substrates	Soil	Soil ID, Name, Description of the soil, soil type, nutrient contents, pH, Size distribution, organic matter content, manufacturers, soil source
Biomaterial Manipulation	Growth Substrates	Aereoponics	Aereoponics ID, Name, constituent and concentration
Biomaterial Manipulation	General Environment		General environment ID, Name, photoperiod, light intensity, light wave length, light type, light manufacturer, watering conditions, relative humidity Day and night, Temperature day and night, carbon dioxide concentration, plant density, selection criteria, Green House Environment ID, Cell culture environment ID, Field environment ID, Growth control agents ID
Biomaterial Manipulation	General Environment	Green house environment	Green House Environment ID, Name, plastic cover, pot size, pot manufacturer, aeration
Biomaterial Manipulation	General Environment	Cell culture environment	Cell culture environment ID, Name, media name, publication, media manufacturer, modification of media, media pH, sugar contents, vitamins, minerals, antibiotic type
Biomaterial Manipulation	General Environment	Field environment	Field environment ID, Name, duration of rain fall, timing of rain fall, vapor pressure deficit, temperature, relative humidity, irrigation method, soil fertility, soil tillage
Biomaterial Manipulation	General Environment	Growth control agents	Growth control agents ID, Name, type, concentration, amount used.
Biomaterial Manipulation	Harvesting Conditions		Harvesting Conditions ID, Name, photoperiod, light intensity, light wave length, light type, light manufacturer, watering conditions, relative humidity Day and night, Temperature day and night, carbon dioxide concentration, plant density, selection criteria

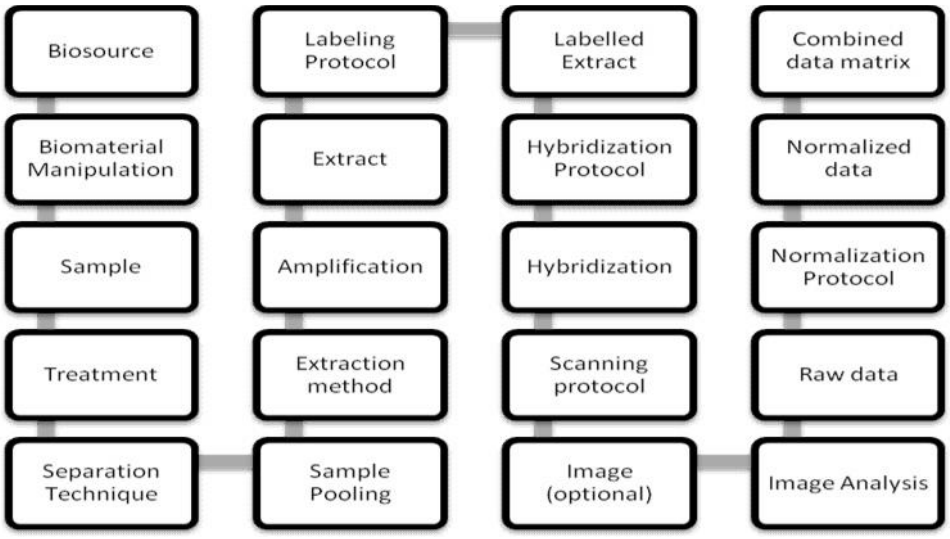


Figure 1 The Data Flow which defines the Experiment Description file.

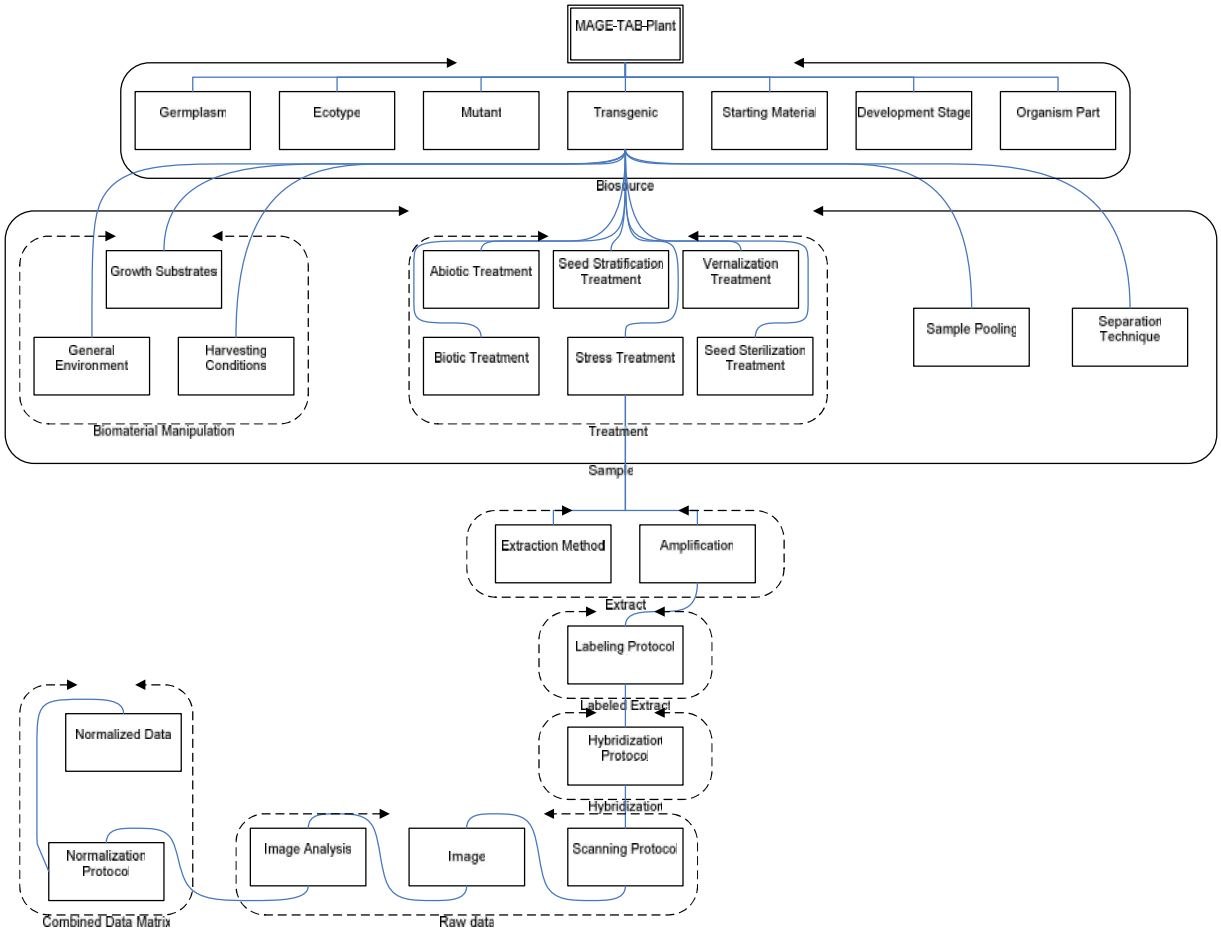


Figure 2 Organization of various files in MAGE-Tab/Plant Format

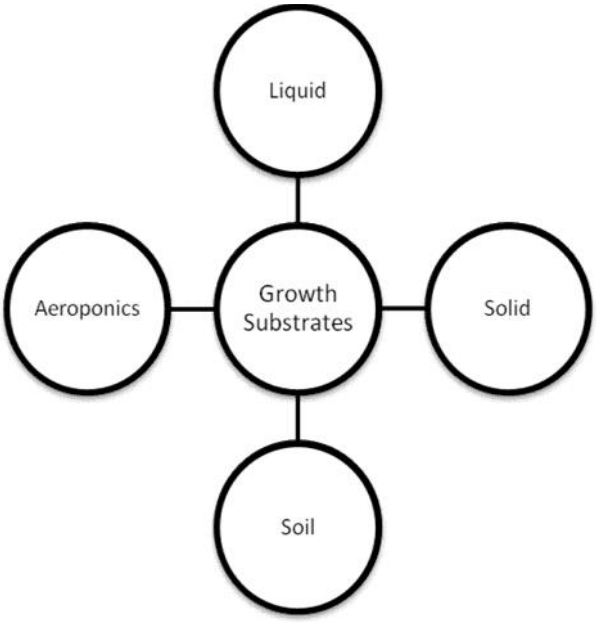


Figure 3 Growth Substrates Description Format.

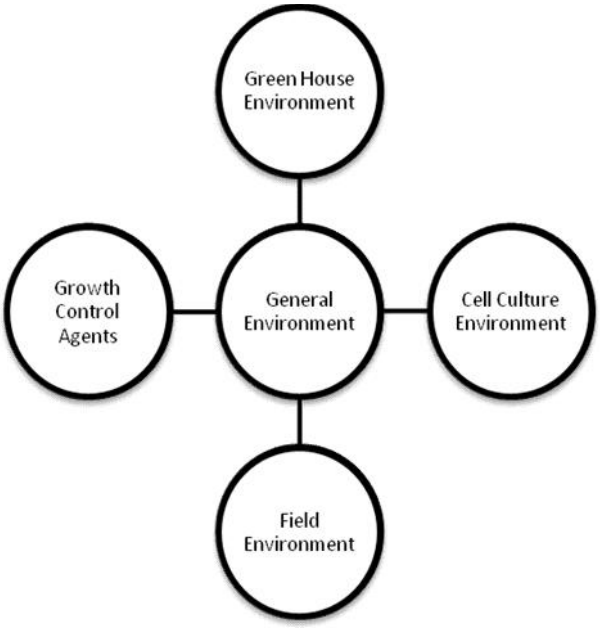


Figure 4 General Environment Description Format

Figure 5 BioSource Annotation Files

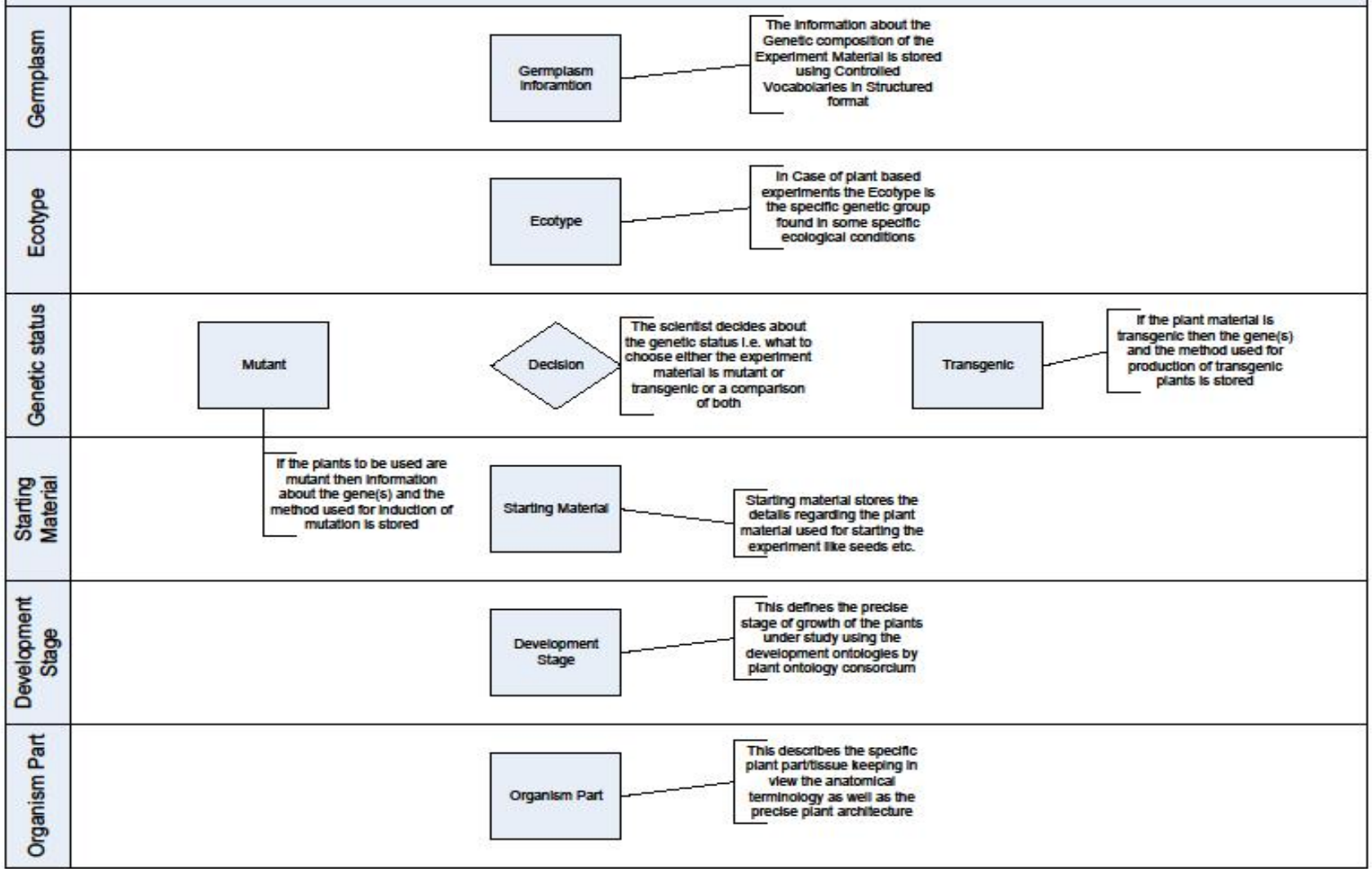
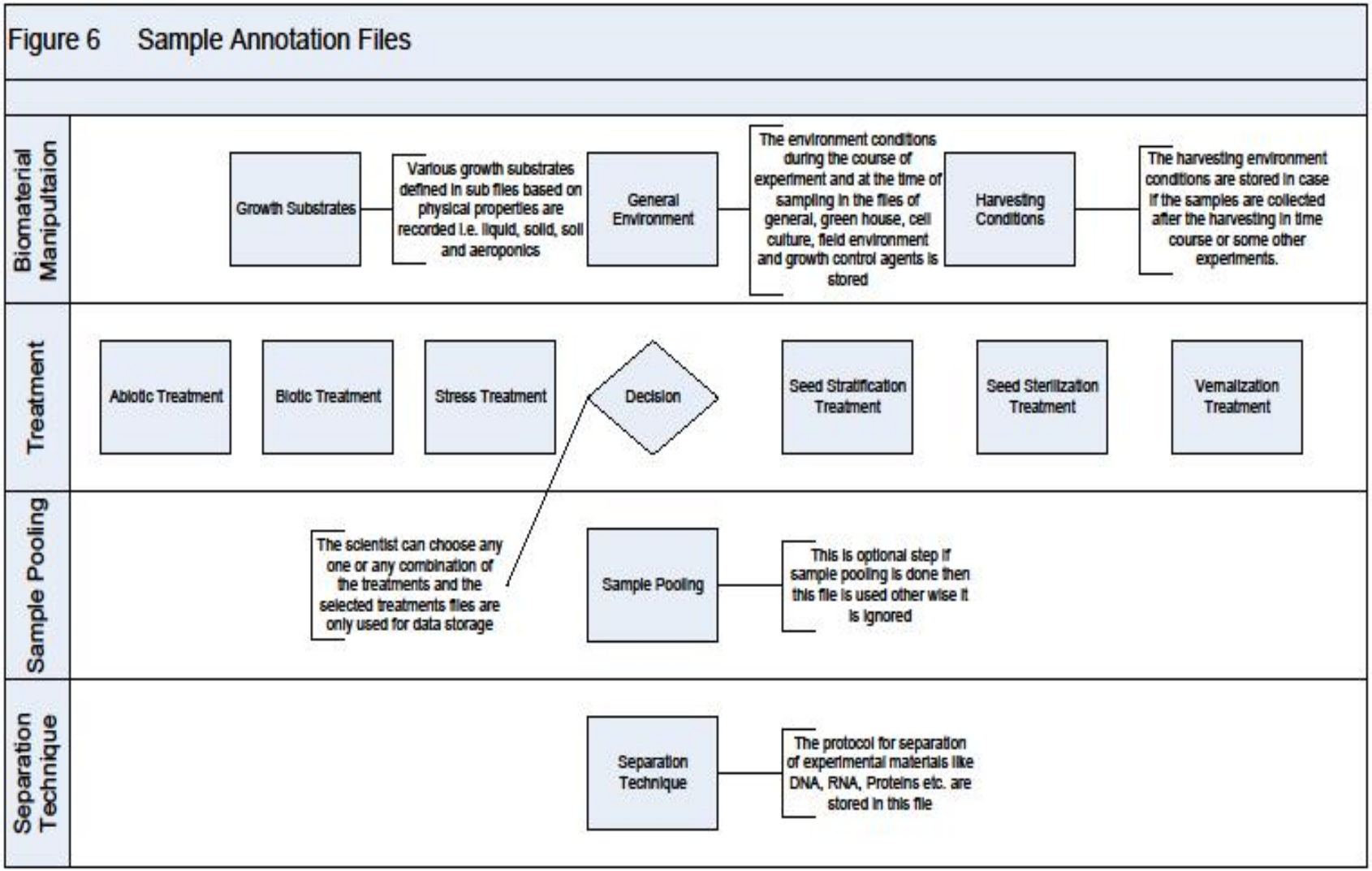


Figure 6 Sample Annotation Files



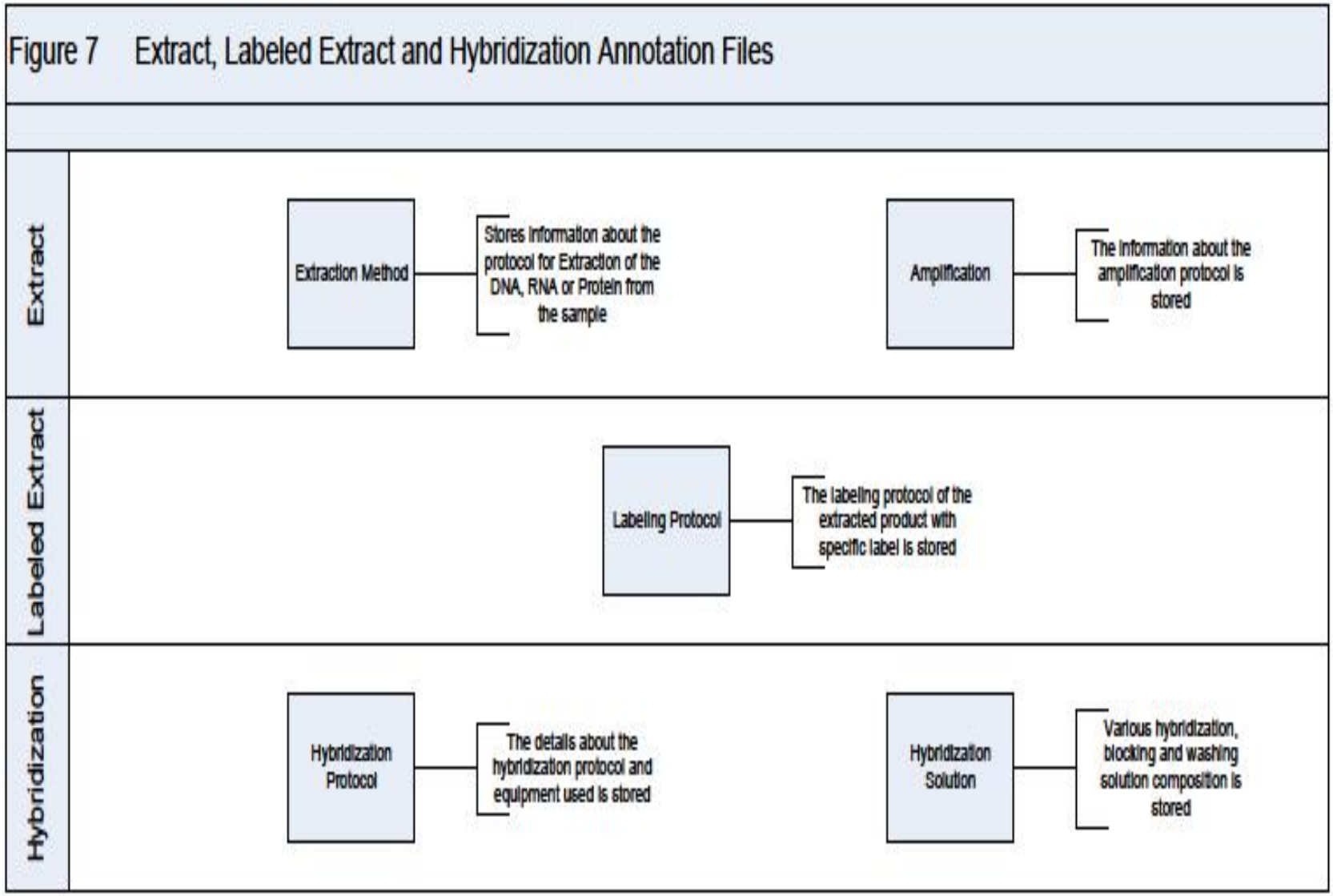
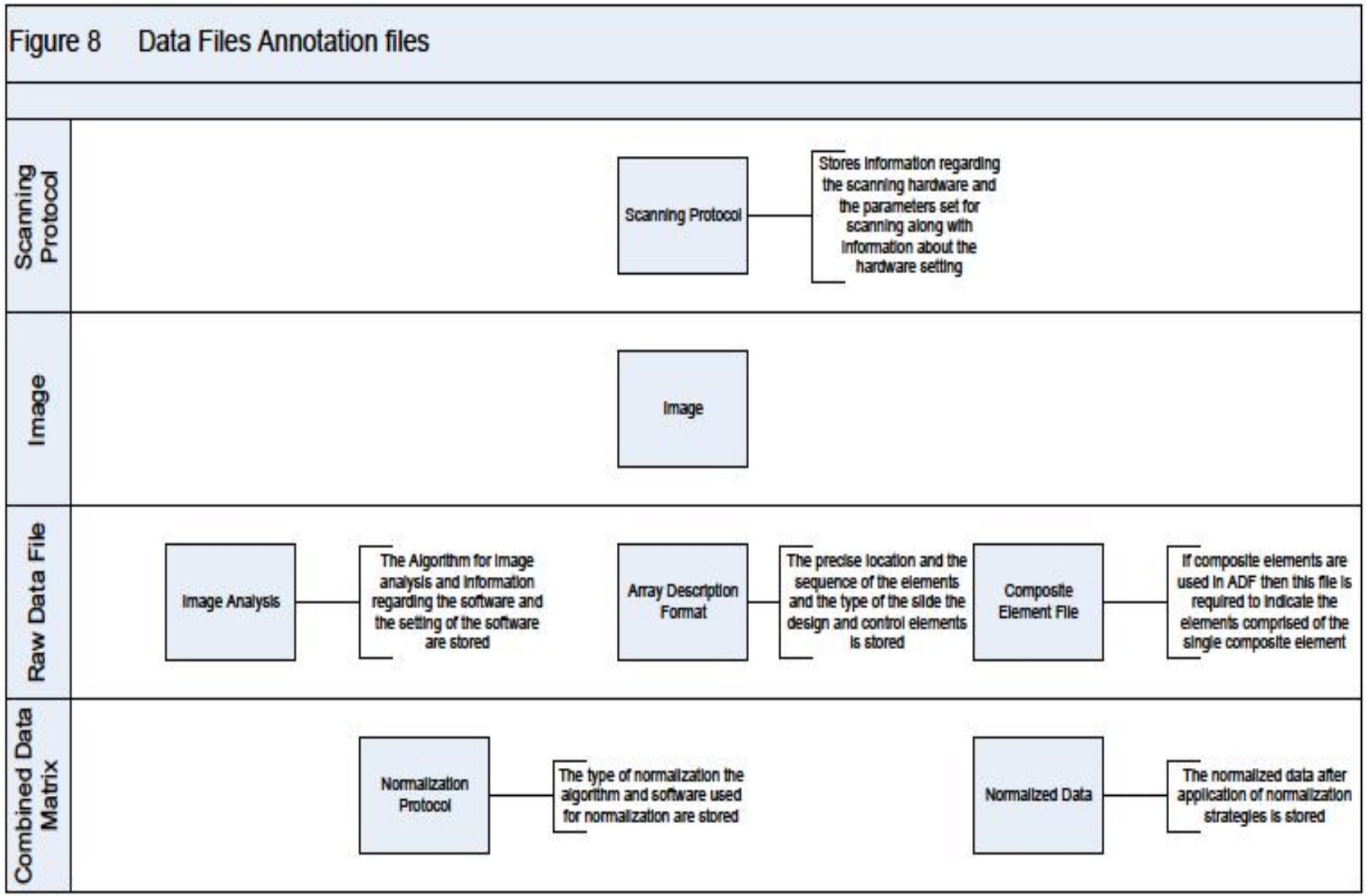


Figure 7 Extract, Labeled Extract and Hybridization Annotation Files
Extract
Extraction Method
Stores information about the protocol for Extraction of the DNA, RNA or Protein from the sample
Amplification
The information about the amplification protocol is stored
Labeled Extract
Labeling Protocol
The labeling protocol of the extracted product with specific label is stored
Hybridization
Hybridization Protocol
The details about the hybridization protocol and equipment used is stored
Hybridization Solution
Various hybridization, blocking and washing solution composition is stored

Figure 8 Data Files Annotation files



Acknowledgement: I would like to acknowledge the financial support in the form of Indigenous Ph.D. Scholarship by Higher Education Commission, Pakistan.

REFERENCES

Anonymous, (2004). Minimum Information About a Microarray Experiment –MIAME for Plant Genomics (MIAME/Plant). http://www.mged.org/Workgroups/MIAME/miame.html.

Bao, W., J.E. Schmid, A.K. Goetz, H. Ren and D.J. Dix (2005). A database for trackin gtoxico-genomic samples and procedures. Reprod. Toxicol. 19:411-419.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron (2001). Minimum information about a microarrayexperiment (MIAME)-toward standards for microarraydata. Nat. Genet. 29:365-371.

Field, D., B. Tiwari and J. Snape (2005). Bioinformatics and data management support forenvironmental genomics. PLoS Biol. 3: e297.

Mukherjee, G., N. Abeygunawardena, H. Parkinson, S. Contrino, S. Durinck, A. Farne, E. Holloway, P. Lilja, Y. Moreau, A. Oezcimen, T. Rayner, A. Sharma, A. Brazma, U. Sarkans and M. Shojatalab (2005). Plant-based Microarray Data at the EBI. Introducing AtMIAM Express a Submission Tool for Arabidopsis Gene Expression Data to Array Express. Plant Physiol. 139:632-636

Rayner, T.F., P. Rocca-Serra, P.T. Spellman, H.C. Causton, A. Farne, E. Holloway, R.A. Irizarry, J. Liu, D.S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C.J. Stoeckert Jr., J. White, P.L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C.A. Ball and A. Brazma (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 7:489 doi: 10.1186/1471-2105-7-489.

Zimmermann, P., B. Schildknecht, D. Craigon, M. Garcia-Hernandez, W. Gruissem, S. May, G. Mukherjee, H. Parkinson, S. Rhee, U. Wagner and L. Hennig (2006). MIAME/Plant – adding value to plant microarray experiments. Plant Methods 2:1 doi: 10.1186/1746-4811-2-1.