DEFINITIONS AND ILLUSTRATIONS OF DATA FROM EXPRESSION ANALYSIS USING AFFYMETRIX GENECHIPS

AFFYMETRIX PLATFORM

Terminology

In this document, we follow conventional terminology that relates microarray experiments to reverse Northerns. The Probe is the oligonucleotide immobilized on the GeneChip. The Target is the cRNA sample that is labeled and hybridized to the GeneChip.

GeneChip Design

Affymetrix GeneChips contain probe sets of short oligonucleotides (25mers). In general, each probe set represents a different transcript or gene, although some duplicates are present. Each probe set contains 11-20 pairs of perfect match (PM) oligos and mismatch (MM) oligos that differ by a single central nucleotide. The array oligos are hybridized to a single biotin-labeled target and stained with a florescent dye conjugated to strepavidin. Thus, the primary data are black and white images of each hybridization.

Data Extraction

Expression values are calculated based on the difference of the PM signal and MM signal at each of the probe pairs using the Affymetrix Microarray Suite Software version 5.0. This software employs statistical algorithms to calculate qualitative and quantitative expression values for each transcript. It also generates a statistical p-value, which suggests the level of confidence in each of the qualitative measurements.

The Affymetrix software generates Absolute Data for each hybridization (i.e. individual sample). It can also be used to calculate Comparison Data between two hybridizations (e.g. control vs. treated). As shown below, similar values are reported in both Absolute and Comparison Data.

Terminology for Data Analysis

Value Category	Absolute Values for Individual Samples	Comparison Values for Two Samples
Quantitative	Signal	Log Ratio (Base 2) indicating raw fold change
Qualitative	Detection (e.g. Absent or Present)	Change (e.g. Increase or Decrease)
Statistical	Detection P-Value	Change P-Value

Data Normalization

Scanned images may have differences in overall brightness due to non-biological factors such as the amount of target hybridized to the array and the amount of stain applied. To minimize such non-biological differences, we normalized the arrays to achieve comparable overall intensity between arrays. The Signal values of any hybridization are multiplied by a Scaling Factor to make their median intensity equal to 500. The Scaling Factor is unique to each hybridization.

ABSOLUTE DATA FILES

Winzip Files on CD

Affy image data = affymetrix image files (DAT)
Affy standard output = EXP, CEL, CHP, RPT affymetrix files
EA generated output = *.txt, *.xls data files

Affymetrix Data Files

Each sample hybridized to an Affymetrix GeneChip generates five Absolute Analysis files, which can be distinguished by their file extensions:

EXP The experimental file stores experimental information, such as sample name and GeneChip array used. It is readable in a text editor.

DAT The data file contains the raw image of the scanned GeneChip array.

CEL The cell intensity file assigns x,y coordinates to each cell (i.e. probe) on the array and calculates the representative intensity of each cell. This file can be used to re-analyze data with different expression algorithm parameters. It is readable in a text editor.

CHP The chip file is generated using information in the CEL file to determine the presence or absence of each transcript and its relative expression level. Text versions of this file are imported into many microarray analysis programs.

RPT This text file gives you quick access to each hybridization's quality control information .

Unfortunately, some Affymetrix files can only be read using Affymetrix software. For your convenience, we have made text (tab delimited) files for the "Pivot" worksheet information of each CHP file. Be assured that all the information you need is included in the text files. We use the Affymetrix software to collect your data, but it is not required for further data analysis. If requested, we can send you the black and white images of each hybridization.

Probe Detection Report

Probe Detection Report files are text-formatted versions of the Pivot worksheet within the Affy CHP files with one line per transcript. The files are labeled with project number_genomeID_chiptype_sample name. Probe Detection Report files contain the Signal, Detection Call and p-value for each transcript on the array. They also list the number of Stat Pairs and Stat Pairs Used for each transcript. These columns indicate how many probe pairs were on the array for each transcript and how many were used in calculating the expression values.

Summary Detection Report

For each batch of samples, we also create a Summary Detection Report, labeled with the Duke0061_2002_11_01_Summary.xls. The Summary Detection Report contains the Signal, Detection Call and P-Value for every sample in each experiment in an easy-to-read Excel format. It also includes the Scaling Factor for each hybridization (described in QC Report).

Illustration of

PROBE DETECTION REPORT

Format = Text; File Name = **project number_genome number_chip name_sample name**.txt
i.e. 0138_2929_H133A_123.*

This report provides standard information that is available from Affymetrix for each gene probe set in a sample. There is one file per sample per chip.

The data are provided in a text file format with one line per gene. An example from a report is shown below.

Probe Set ID	Stat Pairs	Stat Pairs Used	Signal	De-tection	P-Value	Descriptions
92570_at	16	16	64.2	A	0.378184	Cluster Incl AW122482:UI-M-BH2.2-ao...
92571_at	16	16	2116.0	P	0.000266	Cluster Incl D85904:Mouse mRNA for ...
92572_at	16	16	183.0	P	0.021866	Cluster Incl AI509617:vx14h07.y1 ...
92573_at	16	16	4422.7	P	0.000266	Cluster Incl AB021743:Mus musculus ...
92574_at	16	16	1928.7	P	0.000219	Cluster Incl AI851046:UI-M-BH0-ajv-..

Probe Set ID – A unique identifier defined by Affymetrix that associates a set of Probes with a given gene.

Stat Pairs – The number of probe pairs that are associated with a given probe set.

Stat Pairs Used – The number of probe pairs that were used as a base for the calculation of signal. For example, this number would reflect any probe pairs that were masked.

Signal – An overall estimate of gene expression for the particular sample. Signal values are roughly proportional to the amount of transcript present.

Detection – A "call" of the presence or absence of an expression of a gene. This call indicates whether one can distinguish the signal against the background variability.

- P implies Present. The gene is expressing itself in a statistically significant way.
- M implies Marginal. The gene may be expressing itself but the evidence is not conclusive.
- A implies Absence. There is insufficient evidence that the gene is expressing itself.

P-Value – An indicator of statistical significance. The lower the value, the better from a statistical significance viewpoint.

Descriptions - Descriptor and nucleotide positions from Affymetrix

Illustration of

SUMMARY DETECTION REPORT

Format = Excel; File Name = Duke0080_2002_11_01_Summary.xls

This report assimilates the standard information that is available for each sample in a common study. There is one file for all samples in a study.

The results are provided in an Excel spreadsheet with one row per gene. Each sample information is provided in subsequent columns. An excerpt from a report is shown below.

Probe Set ID – A unique identifier defined by Affymetrix that associates a set of Probes with a given gene.

Gene Descriptor - If the gene has a useful name, then this name is provided. Otherwise, the description is summarized from information (source: Affymetrix) concerning the targeted transcript.

Signal – An overall estimate of gene expression for the particular sample. Signal values are roughly proportional to the amount of transcript present (according to Affymetrix).

Detect – A "call" of the presence or absence of an expression of a gene. This call indicates whether one can distinguish the signal against the background variability.

- P implies Present. The gene is expressing itself in a statistically significant way.
- M implies Marginal. The gene may be expressing itself but the evidence is not conclusive.
- A implies Absence. There is insufficient evidence that the gene is expressing itself.

P-Value – An indicator of statistical significance. The lower the value, the better from a statistical significance viewpoint.

Scaling Factor – A multiplying factor that was applied to each raw signal in an array (sample) so that the mean intensity of all signals for a given array (sample) is consistent across samples. Each array sample will have its own scale factor.

Functional gene information is included in the end of the table:

Unigene ID, Locus Link ID (active web link), OMIM, Gene Symbol, sequence derived from (Genbank), GO ontology functional classifications

Illustration of

EXPRESSION ANALYSIS QUALITY CONTROLS REPORT

The quality controls for your samples are presented on a hard-copy spreadsheet enclosed in your data packet. Your samples are measured at five points during the target development and hybridization process to ensure reliable data. The sample checkpoints and the data gathered are detailed below:

Client = Name
EA ID# = Duke project ID, unique to each PI.
Batch = Date of submission to EA

Assess quality of RNA via Agilent 2100 Bioanalyzer

28S/18S - Ratio of 28S peak to 18S peak. Quality total RNA samples have 28S/18S ratios around 2.0.
% Area - Percentage of the 18S and 28S RNA in the total RNA population.

Quantify biotin-labeled cRNA yield via Spectrophotometer

260	Absorption at 260 nm.	**ug/ul**	Concentration of the cRNA product.
280	Absorption at 280 nm.	**Total**	Total yield (ug) of cRNA.
260/280	Ratio of absorption at 260 to 280.	**Volume**	Volume (ul) of cRNA used to make final target.

Verify proper fragmentation via Agilent 2100 Bioanalyzer

Size < 200 Indicates if cRNA has been properly fragmented to a size between 25 and 200nt in length.

Hybridization Controls

Samples are generally hybridized to an Affymetrix Test3 chip before proceeding to the Species GeneChip hybridization. QC values from both hybridizations are recorded.

Noise - A measure of the variability of the background signal.

Scaling - A signal intensity multiplier that brings the average intensity to some preset level (default 500).

Background - A measure of the intensity of the overall background signal. The background value is, in effect, removed from calculations.

% Present - Percentage of genes called “Present” by Affymetrix.

Actin - Ratio of the 3’ end of the Actin target measured to the 5’ end. A high number indicates that full-length transcripts may not have been obtained during target development.

GAPDH - Ratio of the 3’ end of GAPDH target measured to the 5’ end.