

RMA Probe Summarization

GeneSpring GX 7.3.1 And GeneSpring GX 9.0

Contents

Probe Summarization Algorithms	3
Definition and Applications	3
RMA	4
Analyzing Affymetrix Expression Data	5
GeneSpring GX 7.3.1	5
GeneSpring GX 9.0	12

Probe Summarization Algorithms

Definition and Applications

Probe summarization algorithms perform the following 3 key tasks:

- Background Correction
- Normalization
- Probe Summarization (i.e. conversion of probe level values to probeset expression values in a robust, i.e., outlier resistant manner)

The order of the last two steps could differ for different probe summarization algorithms.

For probe intesnsity measurements from Affymetrix Gene expression chips, one of the algorithms used in both GeneSpring GX 7.3.1 and GeneSpring GX 9.0 is RMA

RMA

RMA (Robust Multi-array Analysis) is a method for normalizing and summarizing probe-level intensity measurements from Affymetrix Gene Chips[®]. Starting with the probe-level data from a set of Gene Chips, the perfect-match (PM) values are background-corrected, normalized and finally summarized resulting in a set of expression measures. The three steps of the process are outlined below.

Background Correction

It has been argued that background correction is the most crucial step for probe level processing. The background correction used in RMA is a non-linear correction, done on a per-chip basis. It is is based on the distribution of PM values amongst probes on an Affymetrix array. PM values are a mixture of a background signal, caused by optical noise and non-specific binding, plus a signal, which is what we are trying to detect. The background is estimated as expectation of the signal (S) conditioned on observed PM values (O), using a kernel density estimation in both GeneSpring GX 7.3.1 and GeneSpring GX 9.0. However, however GeneSpring GX 7.3.1 uses direct convolution while GeneSpring GX 9.0 uses Fast Fourier Transformation.

Normalization

Normalization is necessary so that multiple chips can be compared to each other, and analyzed together. The normalization procedure is aimed at making the distributions identical across arrays. The normalization used in RMA is quantile normalization. This usually gives very sharp normalizations.

Both GeneSpring GX 7.3.1 and GeneSpring GX 9.0 use quantile normalization. Note that, in this procedure, all the arrays are used and no chip is discarded based on extreme value considerations.

Summarization

Once the probe-level PM values have been background-corrected and normalized, they need to be summarized into expression measures, so that the result is a single expression measure per probe-set, per chip. The summarization used is motivated by the assumption that observed log-transformed PM values follow a linear additive model containing a probe affinity effect, a gene specific effect (the expression level) and an error term. For RMA, the probe affinity effects are assumed to sum to zero, and the gene effect (expression level) is estimated using median polishing. Median polishing is a robust model fitting technique, that protects against outlier probes.

Both GeneSpring GX 7.3.1 and GeneSpring GX 9.0 use same methodology for summarization.

Analyzing Affymetrix Expression Data

GeneSpring GX 7.3.1

The following steps need to be performed in GeneSpring GX 7.3.1 to analyze Affymetrix gene expression chips :

Step 1 : Import Data

Select the data file you want to import in GeneSpring GX 7.3.1 using File > Import Data

ile	Edit	View	Experiments	Colorbar	
İm	port D	ata	N	©trl+0	
Im	port D	ata fror	n Database 🦷	2	•
Op	en Ge	nome	or Array		•
Vie	ew Pro	jects			•
Ge	nome	Manag	jer		
lm	port G	enome			
Ne	w Win	dow			
Ne	w Linł	ed Wir	ndow		
Lo	gin to '	Norkgr	oup Server		
Bu	ilk Upli	bad to N	Norkgroup Sen	ver	
Co	py Ge	nome f	rom Workgroup) Server	
Ne	w Pati	nway			•
Ne	w Scri	pt			
Ne	w Pro	gram			
Im	port G	eneSpr	ring GX Zip		
Lo	ad Bo	okmark	File		
Sa	ive Boo	kmark	L.,		
Pri	int Ima	ge			•
Sa	ive Ima	ige		8	•
Ch	ose			Ctrl+W	
QL	ıit			Ctrl+Q	

tariq_E4h		tariq_E24h_run1.CEL	tariq_E72h	
-	_run2.CEL	tariq_E24h_run2.CEL	🔟 tariq_E72h	_run2.CEL
🔤 tariq_E4h	_run3.CEL	ariq_E24h_run3.CEL		
🛅 tariq_E16	n_run1.CEL	🔤 tariq_E48h_run1.CEL		
國 tariq_E16	n_run2.CEL	ariq_E48h_run2.CEL		
🗟 tariq_E16	n_run3.CEL	ariq_E48h_run3.CEL		
	10			
File name:	tariq_E4h_	run1	N	Open

Step 2 : Choose File format and select the appropriate genome

GeneSpring GX 7.3.1 automatically recognizes the file format and displays it for standard Affymetrix expression; Agilent one color and two color; Illumina; and Codelink chips.

	Choose File Format. AffyMetrix MG_U74Av2 CEL File 💌
Genom	ie
	Select the genome (set of genes on the array) for this data. If your genome does not appear on the list, you can create a new one by selecting Create a New Genome.
 Sel 	ect Genome
	- & MG_U74Bv2
-	- 🔁 MG_U74Cv2
2	—落 Mouse (011978)
-	- S Mouse Development 44K Oligo Microarray (012799)
9	Mouse Development Microarray4x44k (015062)
	- 🔁 Mouse Oligo Microarray (011978)
2	- S new_genome
	New_Human 1A (V2) Oligo Microarray (012097)
	Cher_MG_U74AV2
	RatGenome9999
001125-025	ate a New Genome
Cho	ose a Name: 🔤

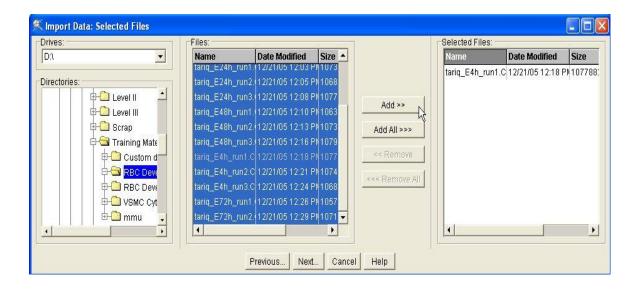
Step 3 : Choose the Preprocessor

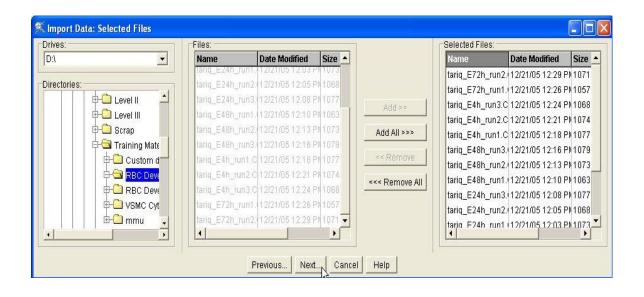
Select the appropriate preprocessing algorithm – '*RMA'* or '*GC RMA'*. You might be asked to define the location of the CDF file or Array Definition file.

Import Data: Preprocess Data Files	
Do you want to preprocess your data files before ir	nporting your samples?
RMA File Preprocessor	•
GC RMA File Preprocessor	
RMA File Preprocessor	

Step 4 : Choose more data files

This window allows you to add more files of the same type to add to your experiment during the import process.





Step 5 : Sample Attributes window

This window allows you to add sample attributes, which are required for MIAME compliance. This is an optional step and can be performed at a later stage as well.

			Please select value	s for sample attributes.		
	Sample Name					New Attribute
Attribute N		Array Design	Author	Experiment Type	Labeling Protocol	Edit Attribute Value
Attribute U	n					
Numeric		no	no	no	no	Delete Attribute
1	tariq_E16h_run1.txt	MG_U74Av2				
2	tariq_E16h_run2.txt	MG_U74Av2				Replace Text
3	tariq_E16h_run3.txt	MG_U74Av2				-
4	tariq_E24h_run1.txt	MG_U74Av2				Fill Down
5	tariq_E24h_run2.txt	MG_U74Av2				Fill Sequence Dow
6	tariq_E24h_run3.txt	MG_U74Av2				
7	tariq_E48h_run1.txt	MG_U74Av2		6		Sort
B	tariq_E48h_run2.txt	MG_U74Av2				
9	tariq_E48h_run3.txt	MG_U74Av2				
10	tariq_E4h_run1.txt	MG_U74Av2				
11	tariq_E4h_run2.txt	MG_U74Av2				
12	tariq_E4h_run3.txt	MG_U74Av2				
13	tariq_E72h_run1.txt	MG_U74Av2				
14	tariq_E72h_run2.txt	MG_U74Av2				
•						F

Step 6 : Experiment Creation

After the data files have been successfully imported and samples have been created, GeneSpring GX 7.3.1 prompts you to create an experiment from these samples.

	Create Experiment
14 new san	ples have been created. Would you
	create an experiment from these
samples?	

Provide an appropriate name for the New Experiment

Name	New Experiment			
Folder Project				Change Project(s).
Notes				
0- 6 1 E	xperiments		Sample Name	
		1	tariq_E16h_run1.txt	
		2	tariq_E16h_run2.txt	
		3	tariq_E16h_run3.txt	
		4	tariq_E24h_run1.txt	
		5	tariq_E24h_run2.txt	
		6	tariq_E24h_run3.txt	
		7	tariq_E48h_run1.txt	
		8	tariq_E48h_run2.txt	
		9	tariq_E48h_run3.txt	
		10	tariq_E4h_run1.txt	
		11	tariq_E4h_run2.txt	
		12	tariq_E4h_run3.txt	
		13	tariq_E72h_run1.txt	
		14	tariq_E72h_run2.txt	

Step 7 : New Experiment Checklist

After the Experiment is created, you get the option to define Experiment Normalizations, Parameters, interpretation and Cross Gene Error Model.

should se choose yo	Imost finished creating your experim t up its normalizations, experimental our default experiment interpretation. s below. Alternatively, you may find th	parameters, and error model, and You may reach these windows using
New Experimer	it Checklist	
	Define Normalizations	Normalizations
	Define Parameters	Parameters
	Define the Default Interpretation	Experiment Interpretation
	Define the Error Model	Error Model

Step 8 : Experiment Normalizations

This window allows you to define what normalization(s) need to be performed on your data.

Total Names Teachers Office and		Order of Normalizations to Perform	Inspect
Add Normalization Step >>	1	Data Transformation: Set measurements less than 0.01 to 0.01	Delete ,
Choose a Normalization Step	2	Per Chip: Normalize to 50th percentile	Delete
Start with pre-normalized values	3	Per Gene: Normalize to median	Move Up
Data Transformation: SAGE transform			Move Down
Data Transformation: Real Time PCR transform			
Data Transformation: Subtract background based c			Use Defaults
Data Transformation: Set measurements less thar			
Data Transformation: Transform from log to linear v			
Data Transformation: Dye swap			
Per Spot: Divide by control channel			
Data Transformation: Reserve control channel			
Per Spot and Per Chip: Intensity dependent (Lowes			
Per Chip: Normalize to a median or percentile			
Per Chip: Normalize to positive control genes	L Llee	Recommended Order Get Text Description	
Per Chip: Normalize to a constant value	1 036		
Per Gene: Normalize to specific samples	Use	a Saved Scenario Save As Scenario	
Per Gene: Normalize to median	Warn	ings	
Per Chip and Per Gene: Median polishing		No warnings.	

For affymetrix data preprocessed using RMA or GC RMA preprocessor, 'data transformation' and 'per chip' normalization needs to be deleted at this step, as these normalization steps have already been performed during preprocessing.

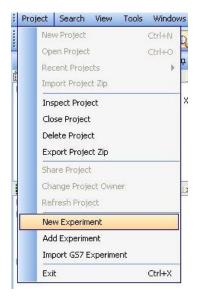
Experiment Normalizations: New Experiment		
the state of the s	Order of Normalizations to Perform	Inspect
Add Normalization Step >>	1 Per Gene: Normalize to median	Delete
Choose a Normalization Step 📃		
Start with pre-normalized values		Move Up
Data Transformation: SAGE transform		Move Down
Data Transformation: Real Time PCR transform		
Data Transformation: Subtract background based on nega		Use Defaults
Data Transformation: Set measurements less than 0.01 to		
Data Transformation: Transform from log to linear values		
Data Transformation: Dye swap		
Per Spot: Divide by control channel		
Data Transformation: Reserve control channel		
Per Spot and Per Chip: Intensity dependent (Lowess) norr		
Per Chip: Normalize to a median or percentile		1
Per Chip: Normalize to positive control genes	Use Recommended Order Get Text Description	Ī
Per Chip: Normalize to a constant value		
Per Gene: Normalize to specific samples	Use a Saved Scenario Save As Scenario	
Per Gene: Normalize to median	Warnings	
Per Chip and Per Gene: Median polishing	No per chip normalization has been applied to any sample.	
	OK Cancel Help	

GeneSpring GX 9.0

The following steps need to be performed in GeneSpring GX 9.0 to analyze Affymetrix gene expression chips :

Step 1 : Create New Experiment

Create a new experiment using Project > New Experiment



Step 2 : Experiment Description

Provide an appropriate Name and Experiment type (or, chip type) for the new experiment

Experiment description	
	select the appropriate experiment type, and choose the I take you through experiment creation and analysis, while he full set of analysis tools.
Experiment name	New Experiment
Experiment type	Affymetrix Expression
Workflow type	Illumina Single Color Affymetrix Exon Expression
Experiment notes	Generic Two Color Agilent Two Color
	Agilent Single Color
	Generic Single Color Affymetrix Expression
	Construction of the Constr
	Affymetrix Expression

You can also define the Workflow type - 'Guided Workflow' or 'Advanced Analysis'

Guided workflow is designed to assist the user throughout the creation and analysis of an experiment with a set of default parameters, while in the *Advanced Analysis*, the parameters can be changed to suit individual requirements.

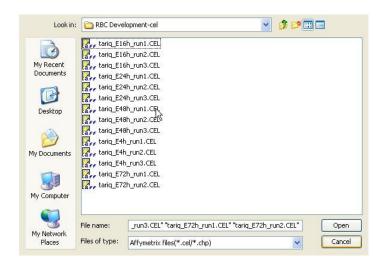
🕅 New Experiment		
Experiment description		
	select the appropriate experiment type, and choose the I take you through experiment creation and analysis, while he full set of analysis tools.	
Experiment name	New Experiment	
Experiment type	Affymetrix Expression	~
Workflow type	Advanced Analysis	~
Experiment notes	Guided Workflow - Find Differentially Expressed Genes Advanced Analysis	
Help	OK Can	cel

Step 3 : Load Data

An experiment can be created using either the data files or else using samples. Upon loading data files, GeneSpring GX associates the files with the technology (see below) and creates samples. These samples are stored in the system and can be used to create another experiment via the *Choose Samples* option. For selecting data files and creating an experiment, click on the *Choose File(s)* button.

Load	Data	
You can choose data files, previously used samples or both to use in this experiment. Once a data file has been imported and used as a sample, it will be available for use in any future experiment.		
Туре	Selected files and samples	
	Choose Files Choose Samples Reorder Remove	
Не	lp << Back Next >> Finish Cancel	

Navigate to the appropriate folder



Select the files of interest and select *Open* to proceed.

Look in:	C RBC Deve	elopment-cel	*] 🤌 🕬 🛄 🖬	
My Recent Documents	Image: The second sec	6h_run2.CEL 6h_run3.CEL 4h_run1.CEL 4h_run2.CEL			
Desktop	tariq_E2 tariq_E4 tariq_E4 tariq_E4 tariq_E4 tariq_E4 tariq_E4	8h_run1.CEL 8h_run2.CEL 8h_run3.CEL			
My Documents	tariq_E4	h_run3.CEL 2h_run1.CEL			
My Computer	File name: Files of type:	_run3.CEL" "tariq_E72h_ Affymetrix files(*.cel/*.ct		run2.CEL"	Open Cancel

Туре	Selected files and samples	
	tariq_E16h_run1.CEL	~
	tariq_E16h_run2.CEL	
	tariq_E16h_run3.CEL	
	tariq_E24h_run1.CEL	
	tariq_E24h_run2.CEL	
	tariq_E24h_run3.CEL	
	tariq_E48h_run1.CEL	
	tariq_E48h_run2.CEL	
	tariq_E48h_run3.CEL	
	tariq_E4h_run1.CEL	
	tariq_E4h_run2.CEL	
	tariq_E4h_run3.CEL	
	tariq_E72h_run1.CEL	~

There are two things to be noted here. Upon creating an experiment of a specific chip type for the first time, the tool asks to download the technology from the GeneSpring GX update server. If an experiment has been created previously with the same technology, GeneSpring GX then directly proceeds with experiment creation.

Step 4 : Select ARR files

ARR files are Affymetrix files that hold annotation information for each sample CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample.

	ute files (.ARR files) associated with chosen samples. The ARR files amples based upon the sample name. These will be imported as sle.
Select ARR files	
Select ARR files	
	Choose file(s) Remove file(s)
Help	<pre><< Back Next >>> Finish Cancel</pre>

Step 5 : Select Probe Summarization and Normalization options

Select RMA as the Probe Summarization algorithm from the drop down list.

Subsequent to probe set summarization, baseline Transformation of the data can be performed. The baseline options include:

- Do not perform baseline
- Baseline to median of all samples
- Baseline to median of control samples

Note : `Baseline Transformation' in GeneSpring GX 9.0 is equivalent to `per gene normalization' in GeneSpring GX 7.3.1

Summarization Algorithm:	RMA	
aseline Transformation	RMA	
	PLIER16	he .
🔵 Do not perform baseline transformal	t GCRMA	•
Baseline to median of all samples	LiWong	
) Baseline to median of control sample	MASS	
Choose samples:		
Available samples	Control samples	
tariq_E16h_run1.CEL		
tariq_E16h_run2.CEL		
tariq_E16h_run3.CEL		
tarig_E24h_run1.CEL		
tarig_E24h_run2.CEL		
tariq_E24h_run3.CEL	~	
Esua Edge wist / El		

Clicking Finish creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to View in Toolbar.