

# **GCRMA Probe Summarization**

GeneSpring GX 7.3.1 And GeneSpring GX 9.0

# Contents

Probe Summarization Algorithms	3
Definition and Applications	3
GCRMA	4
Analyzing Affymetrix Expression Data	7
GeneSpring GX 7.3.1	7
GeneSpring GX 9.0	14

## **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 GCRMA.

## GCRMA

GCRMA is a method of converting *.CEL* files into expression set using the *Robust Multi-array Average* (*RMA*) with the help of probe sequence and with GC-content background correction.

It is a method for normalizing and summarizing probe-level intensity measurements from Affymetrix GeneChips. Starting with the probe-level data from a set of GeneChips, 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 given below.

#### **Background Correction:**

The background correction used in GCRMA is designed to account for background noise, as well as non-specific binding. Probe affinity is modeled as a sum of position-dependent base effects, and can thus be calculated for each PM and MM value, based on its corresponding sequence information.

The correction is motivated by the assumptions that observed PM and MM values consist of optical noise, non-specific binding noise, and signal. Optical noise is assumed to be normal, and logged non-specific binding noise from PM-MM pairs assumed to be bivariate normal. Using the data on a single array, the corresponding model parameters can be estimated.

The background adjustment in GCRMA consists of three sequential steps:

- 1. optical background correction
- 2. probe intensity adjustment through non-specific binding (NSB) utilizing affinity information and optical noise-adjusted MM intensities
- 3. probe intensity adjustment through gene-specific binding (GSB), where NSB-adjusted PM intensities are further corrected for the effect of PM probe affinities

Optical correction is important, as scanner measuring hybridization strength introduces optical noise. In GeneSpring GX 7.3.1 , background correction is still done using MM but these values are not adjusted for optical noise (i.e we still adjust through non specific binding using affinity and do probe intensity adjustments). Thus, implementation of optical correction in GeneSpring GX 9.0 is an important enhancement.

Each PM value is then adjusted by subtracting a shrunken MM value that has been corrected for its affinity. The affinity of a probe is described as sum of position dependent base affinities; this affinity of probe is calculated from their sequence. Each base at each position contributes to the total affinity of probe in an additive fashion. For a given type of base the positional effect is modeled.

#### Normalization

Normalization is necessary so that multiple chips can be compared to each other, and analyzed together. It is motivated by the assumption that all n chips should have approximately the same distribution of PM values. The normalization used in GCRMA is quantile normalization. This is a generalization of the idea behind quantile-quantile plots to more than two dimensions. The quantiles for each PM value are plotted in n dimensions, and projected onto the diagonal. The final result is that the PM values on each chip will have the same distribution.

#### 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 probeset, 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 GCRMA, 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. The analysis performed using GCRMA in GeneSpring GX 9.0 shows different results from that of GeneSpring GX 7.3.1. This is primarily due to the difference in the implementations of GCRMA algorithm in both the versions. GeneSpring GX 9.0 uses an additional component of 'Optical Correction' while performing GCRMA, where as GeneSpring GX 7.3.1 does not perform optical correction.

As an example, for the following data set both GeneSpring GX 7.3.1 and GeneSpring GX 9.0 produce different results, when processed using GCRMA.

#### Experimental design and data set:

Patients with cardiomyopathy have weakened heart pumps which can result in the heart not being able to pump enough blood to the body's other organs- a condition known as congestive heart failure (CHF). Patients with ischemic cardiomyopathy have weakened heart pumps due to insufficient blood and oxygen being delivered to the area. A patient with non-failing hearts was performed.

We have four replicates each for *ischemic cardiopathy* and *Normal Ischemic* : PAS\_3.cel, PAS\_6.cel, PAS\_7.cel, PAS\_8.cel *Normal* : PA-N\_249.cel, PA-N\_300.cel, PA-N\_322.cel, PA-N\_326.cel

#### Welch T-test Result Matrix

Total Number of genes = 54675

Application	P value less than 0.05 without MTC
R without optical correction	11862
R with optical correction	33706
GeneSpring GX 7.3.1	12867
GeneSpringGX 9.0	33964

The above data set can be downloaded from the following ftp site <a href="http://cg2.med.harvard.edu/pub/proj1/expts/Hs/Affy/">http://cg2.med.harvard.edu/pub/proj1/expts/Hs/Affy/</a>

## **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 I
Im	oort D	ata	N	Ctrl+O
Im	oort D	ata fron	n Database 🛛 <sup>K</sup>	\$,
Op	en Ge	nome	or Array	۲
Vie	w Pro	jects		•
Ge	nome	Manag	jer	2
Im	oort G	enome		
Ne	w Win	dow		
Ne	w Linł	ked Wir	ndow	
Lo	gin to '	Workgr	oup Server	
Bu	lk Upli	oad to \	Norkgroup Serv	/er
Co	py Ge	nome f	rom Workgroup	) Server
Ne	w Patł	nway		۲
Ne	w Scri	pt		
Ne	w Pro	gram		•
Im	oort G	eneSpr	ing GX Zip	
Lo	ad Bo	okmark	File	
Sa	ve Boo	okmark	la la	
Pri	nt Ima	ge		•
Sa	ve Ima	age		۱.
Clo	se			Ctrl+W
Ou	it			Ctrl+Q

Select one d	ata file 🔹 🖓 🔀
Look in: ଢ	Pam's GCRMA test data 💽 🔶 🖻 👘 🖽 -
PA-N_249 PA-N_300 PA-N_322 PA-N_326 PA-S_3 PAS_3 PAS_6	▶ PAS_7 ▶ PAS_8
File name:	PA-N_249 Open
Files of type:	All Files (*.*)

### 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.

🍭 Import Data: Define File Format and Genome 📃 🗖 🔀
File Format
Choose File Format: AffyMetrix HG-U133_Plus_2 CEL File
Genome
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.
<ul> <li>Select Genome</li> </ul>
- 🔁 Chicken 🔤
- S Chicken_new
- S Custom_Cotton
Custom Agilent WHole Human Genome (012391_copy)
HG_Focus
-월 <mark>HG-U133_Plus_2</mark>
- 🔁 Human 1A (V2) Oligo Microarray (012097)
Human 1AA
Human 1AA
Human 1A Oligo Microarray (011521)
Create a New Genome
Choose a Name:
Next Cancel Help

#### 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.



## 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.

🌾 Import Data: Selected Files						
Drives:	Files:				Selected Files	3:
CA 🔹	Name	Date Modified	Size		Name	Date Modified Size
	PA-N_249.cel	12/14/03 11:00 P	326576:		PA-N_249.ce	el 12/14/03 11:00 PN 326576:
Directories:	PA-N_300.cel	12/14/03 11:00 P	326830			
🕀 🖾 .ssh 🔶	PA-N_322.cel	12/14/03 11:00 PI	327368	ee bbA		
🕀 🗀 Application Data	PA-N 326.cel	12/14/03 11:00 PI	329002			
🕀 🗀 Cookies	PAS 3.cel	5/31/04 11:00 PM	135516	Add All >>>		
🕂 🖨 Desktop	PAS 6.cel	5/31/04 11:00 PM	135511			
🕀 🛄 Agilent	PAS 7.cel	5/31/04 11:00 PM	135514	<< Remove		
GCRMAtestd	PAS_8 cel	5/31/04 11:00 PM	135506	Less Romous MI		
Phillip_test_(				SSS REMOVE AIL		
🕂 🕀 🗁 Travel						
🕂 🕀 🗐 🗊 🕂						
	•		•		•	• •
		Previous Next	. Cance	I Help		

🕵 Import Data: Selected Files							
Drives:	Files:				Selected Files:		
C:1 💌	Name	Date Modified	Size		Name	Date Modified	Size
	PA-N_249.cel	12/14/03 11:00 PI	326576:		PAS_8.cel	5/31/04 11:00 PM	135506
Directories:	PA-N_300.cel	12/14/03 11:00 PI	326830		PAS_7.cel	5/31/04 11:00 PM	135514
tan an a	PA-N_322.cel	12/14/03 11:00 PI	327368	<< bbA	PAS_6.cel	5/31/04 11:00 PM	1355111
🕀 🛄 Application Data	PA-N_326.cel	12/14/03 11:00 PI	329002		PAS_3.cel	5/31/04 11:00 PM	135516
🕀 🕮 Cookies	PAS_3.cel	5/31/04 11:00 PM	135516	Add All >>>	PA-N 326.cel	12/14/03 11:00 P	329002
📄 📄 🔁 Desktop	PAS_6.cel	5/31/04 11:00 PM	1355111		PA-N 322.cel	12/14/03 11:00 P	327368
🕀 💼 🖾 Agilent	PAS 7.cel	5/31/04 11:00 PM	135514	<< Remove	PA-N_300.cel	12/14/03 11:00 P	326830
GCRMAtestd	PAS 8.cel	5/31/04 11:00 PM	135506	and Romovo All	PA-N 249 cel	12/14/03 11:00 P	326576
🖶 📴 Phillip_test_(							1
Travel							
🕂 🕀 🗐 bike 🖓							
	•		F		•		► I
		Previous Next	Cance	I Help			

## 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.

Import Da	ta: Sample Attributes							
	Please select values for sample attributes.							
	Sample Name					New Attribute		
Attribute Nam		Array Design	Author	Experiment Type	Labeling Protocol	Edit Attribute Value		
Attribute Units						Euit Auripute Value		
Numeric		no	no	no	no	Delete Attribute		
1	PA-N_249.bt	HG-U133_Plus_2						
2	PA-N_300.txt	HG-U133_Plus_2				Replace Text		
3	PA-N_322.txt	HG-U133_Plus_2				Troplace Text		
4	PA-N_326.txt	HG-U133_Plus_2				Fill Down		
5	PAS_3.txt	HG-U133_Plus_2				Fill Sequence Down		
6	PAS_6.txt	HG-U133_Plus_2				0.4		
7	PAS_7.txt	HG-U133_Plus_2				son		
8	PAS_8.txt	HG-U133_Plus_2						
•								
		F	Previous Next Canc	el Help				

## 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.

🧟 Import Data: Create Experiment 🛛 🔀
8 new samples have been created. Would you also like to create an experiment from these samples?
Yes No

Provide an appropriate name for the New Experiment

🕵 Save	New Experiment				×
Name	CHF				
Folder					
Project				Change Project(s)	
Notes					< >
в-🔄 Б	periments		Sample Name		4
		1	PA-N_249.txt		
		2	PA-N_300.txt		
		3	PA-N_322.txt		
		4	PA-N_326.bt		
		5	PAS_3.txt		
		6	PAS_6.txt		
		7	PAS_7.txt		
		8	PAS_8.txt		
					-
			Save Cancel		

## 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.

You are a should se choose yo the button	Imost finished creating your experim t up its normalizations, experimental our default experiment interpretation. s below. Alternatively, you may find th	ent. Before you begin analysis, you   parameters, and error model, and You may reach these windows using nem in the Experiments menu.
New Experimer	it Checklist	
	Define Normalizations	Normalizations
	Define Parameters	Parameters
	Define the Default Interpretation	Experiment Interpretation
п	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 Normanzation 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			- more bomin
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	Percemmended Order Get Text Description	
Per Chip: Normalize to a constant value	1 030		
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. Applying 'Per Gene' normalization is optional.

K Experiment Normalizations: CHF					
teld Marmalization Stan 22		Order of Normalization	ons to Perform	<u> </u>	Inspect
Add Normalization Step >>		No normalizations ha	ve been added.		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 negative					Use Defaults
Data Transformation: Set measurements less than 0.01 to 0.					
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) normali					
Per Chip: Normalize to a median or percentile				·	
Per Chip: Normalize to positive control genes	🗌 Use F	Recommended Order	Get Text Description		
Per Chip: Normalize to a constant value	Lise :	a Saved Scenario	Save As Scenario	1	
Per Gene: Normalize to specific samples	−Marni	ings	ourorio oconuno	_	
Per Gene: Normalize to median	**um	No per chip po	rmalization has been applied	to any sample, and no per sp	nt
Per Chip and Per Gene: Median polishing		or per gene no	rmalization has been applied	to any sample.	-
<b></b>					
		OK Cancel He	elp		

## 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

oject	Search	View	Tools	Window
Ne	w Project			Ctrl+N
Open Project				Ctrl+0
Re	cent Proje	cts		
Imp	oort Proje	t Zip		
Ins	pect Proje	ect		
Clo	ise Project	:		
De	lete Projec	:t		
Ex	Export Project Zip			
Sha	are Project	R.		
Ch	ange Proje	ect Own	er	
Rel	fresh Proje	ect		
Ne	w Experim	ent		
Ad	d Experim	ent		
Im	port GS7 E	xperime	ent	
Exi	t			Ctrl+X
	iject Nev Op Rev Im Ins Clo De Ex Ch Rel Sha Ch Rel <b>Ne</b> Ad	iject Search New Project Open Project Import Project Inspect Project Close Project Delete Project Share Project Share Project Change Project Change Project Refresh Project New Experim Import GS7 E Exit	jject Search View New Project Open Project Recent Projects Import Project Zip Inspect Project Close Project Delete Project Export Project Zip Share Project Change Project Own Refresh Project New Experiment Add Experiment Import GS7 Experime Exit	jject Search View Tools New Project Open Project Recent Project Zip Inspect Project Close Project Close Project Delete Project Export Project Zip Share Project Change Project Owner Refresh Project New Experiment Add Experiment Import GS7 Experiment Exit

## Step 2 : Experiment Description

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

Experiment description	
Enter a name for the new experiment, desired workflow. Guided workflows wil advanced analysis will allow access to t	select the appropriate experiment type, and choose the II 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

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		
Enter a name for the new experiment, desired workflow. Guided workflows wil advanced analysis will allow access to t	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	
He	Ip <<< Back Next >> Finish Cancel	

#### Navigate to the appropriate folder



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

Look jn	a: 🧰 GCRMAtestdata 💉 🦻 📁 🖽 📾
My Recent Documents	2022         PA-N_249.cel           2022         PA-N_300.cel           2022         PA-N_322.cel           2022         PA-N_326.cel           2022         PAS_3.cel           2022         PAS_6.cel           2022         PAS_7.cel           2022         PAS_7.cel           2022         PAS_8.cel
My Documents	
My Computer	
My Network Places	File name:     _326.cell" "PA5_3.cell" "PA5_6.cell" "PA5_7.cell" "PA5_8.cell"     Open       Files of type:     Affymetrix files(*.cel/*.chp)     Cancel

🕂 New	Experiment (Step 1 of 4)		
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		
	PA-N_249.cel		
	PA-N_300.cel		
	PA-N_322.cel		
	PA-N_326.cel		
	PAS_3.cel		
	PAS_6.cel		
	PAS_7.cel		
	PAS_8.cel		
	Choose Files Choose Samples Reorder Remove		
He	p <<< Back Next >> Einish Cancel		

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.

Select ARR Files Select the sample attribu	ute files (.ARR files) associated with chosen samples. The ARR files
will be associated with s annotations to the samp	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	<< Back Next >> Finish Cancel

### Step 5 : Select Probe Summarization and Normalization options

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

As an optional step, 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

K New Experiment (Step 3 of 4	)	×
Summarization Algorithm Select a summarization algorithm fr create new experiment with norma	om the dropdown list and the baseline transformation to lized expression values.	
Summarization Algorithm:	GCRMA	~
Baseline Transformation	RMA PLIER16	
<ul> <li>Do not perform baseline transforma</li> </ul>	GCRMA	
O Baseline to median of all samples	LiWong	
O Baseline to median of control sample	MAS5 es	
Choose samples:		
Available samples	Control samples	
PA-N_249.cel		
PA-N_300.cel		
PA-N_326.cel		
PAS_3.cel		
PAS_6.cel	~	
		- )
Help	<< Back Mext >> Einish Cancel	]

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.