GCAT WORKSHOPS 2009

NSF SUPPORTED DNA MICROARRAY WORKSHOPS

FOR

HISTORICALLY BLACK COLLEGES AND UNIVERSITIES,
HISPANIC-SERVING INSTITUTIONS AND TRIBAL COLLEGES
FACULTY

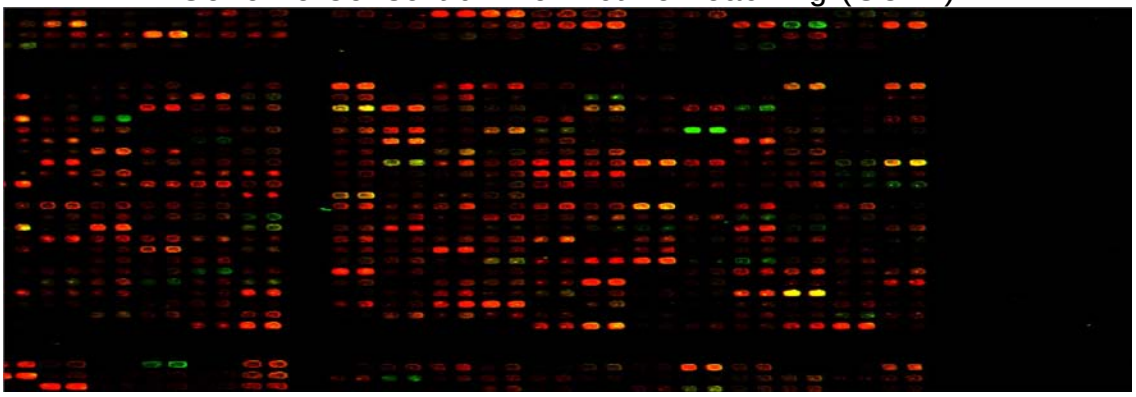
JULY 5 - July 11, 2009

HOST INSTITUTION: MOREHOUSE COLLEGE
BIOLOGY DEPARTMENT; ATLANTA, GA

INSTRUCTORS

Consuelo Alvarez, Malcolm Campbell, Todd Eckdahl,
Edison Fowlks, Charles Hauser, Laurie Heyer, and Anne Rosenwald

Genome Consortium for Active Teaching (GCAT)



NSF GRANT DBI-0520908 Awarded to Hampton University
PI: Edison R. Fowlks; Co-PIs: Mary Lee Ledbetter and Anne Rosenwald
and
NSF GRANT DBI-0520908 Awarded to Davidson College
PI: Malcolm Campbell; Co-PIs: Laurie Heyer and Todd Eckdahl
By the
DIRECTORATE FOR BIOLOGICAL SCIENCES
NATIONAL SCIENCE FOUNDATION
DR. SALLY O'CONNOR, MANAGERING PROGRAM OFFICER

NSF-SPONSORED GCAT

DNA MICROARRAY SUMMER WORKSHOPS FOR UNDERGRADUATE HBCU, HSI AND TRIBAL COLLEGE FACULTY

TAB

ORIGIN OF THIS NSF WORKSHOP

At a curriculum gathering during summer of 2004, Drs. Fowlks and Campbell presented their perspectives on the undergraduate educational implications of the NIH Roadmap. Fowlks addressed curricular changes in undergraduate biology, as well as recruitment and training for underrepresented populations. Campbell provided an overview of activities by GCAT, a non-profit educational consortium, to bring functional genomics methods into undergraduate courses and independent student research. After the presentations, an invited representative of NSF encouraged Dr. Fowlks and Dr. Campbell to submit a proposal that would provide a microarray workshop opportunity for undergraduate faculty at minority serving institutions (MSIs). The 2005 GCAT workshop at Morehouse College was an overwhelming success based on participant and instructor evaluations (see Outcomes of Previous Workshops below), as well as comments from two NSF Program Officers (POs) who attended part of the workshop. These POs encouraged Fowlks, Campbell, and their GCAT colleagues to submit a three-year proposal for DNA microarray workshops that would have a significant impact on the introduction of genomics in undergraduate curricula, especially in MSIs. On the advice of the POs, this collaborative proposal brings Hampton University as the lead institution with Davidson College as the collaborative institution. Fowlks and Heyer will co-chair the advisory board which will focus on participant recruitment (emphasizing faculty from MSIs). Fowlks also will oversee workshop housing, transportation, and on-site facilities logistics while Campbell will focus on wet- and dry-lab curriculum, recruitment of instructors, and organization of the workshop technical aspects. As described in the budget justification, the budget will be divided according to these two functional roles.

History of Genomics and GCAT Genomics first grabbed the headlines with the Human Genome Project. This monumental task was the beginning of a long process to understand how humans work. The production of any genome sequence is analogous to creating a periodic table for the chemical elements – a great start but ultimately only a parts list, failing to explain how the parts interact. Experimental data are needed to illuminate how all the genes work synergistically. One way to measure every gene's activity simultaneously is to use DNA microarrays. DNA microarrays are produced by printing one spot of DNA for each gene onto a specially coated glass microscope slide. Each spot contains a unique DNA sequence specific for only one gene. For example, a yeast DNA microarray is composed of ~ 6,500 spots – one for each of its 6,500 genes. The investigator isolates the mRNAs and couples one color fluorescent dye to the mRNAs for one growth condition. In parallel, the investigator isolates and labels with a different color the mRNAs isolated from cells grown under a different condition. The two populations of differently colored mRNAs are mixed and allowed to base pair with the appropriate genes among the 6,500 spots on the microarray. The fluorescence for each color on every spot is quantified to produce a ratio of mRNA binding from the two growth conditions. From this list of 6,500 ratios, the investigator can determine which genes are activated and which ones are repressed in the two growth conditions (see an animated version here: <www.bio.davidson.edu/Courses/genomics/chip/chip.html>). Measuring the genomic transcriptional response to different growth conditions is a critical step in understanding how each gene contributes to the genome's combinatorial functional capacity.

GCAT was conceived in December 1998 by Dr. A. Malcolm Campbell (Davidson College) and Dr. Mary Lee Ledbetter (College of the Holy Cross), who were inspired by Dr. Patrick Brown's talk at the 39th Annual meeting of the American Society of Cell Biology regarding his pioneering work on DNA microarrays. DNA microarray technology has revolutionized the investigation of gene expression at the genomic level, and holds great promise for understanding the complex, systems-level regulation in many species. Many job opportunities in academia and industry are available for trained college graduates. Furthermore, the best graduate programs are looking for students with practical experience. However, microarray technology is inaccessible to many undergraduates because of limitations in funding and faculty expertise. GCAT's mission is to make such experiments accessible to all undergraduates and we are succeeding: since 2000, about 5,000 undergraduates have performed experiments with DNA microarrays obtained through GCAT and analyzed their results with free software developed by co-PI Laurie Heyer and her undergraduate students (www.bio.davidson.edu/MAGIC). GCAT makes microarray experiments affordable through cost-sharing, provides a clearinghouse of information, raw data and analyzed results for use in teaching genomics, and represents a network of mutually supportive teachers using functional genomics.

WORKSHOP #1 PARTICIPANTS

Name	Institution	Email Address
Dixon, Emily	St. Lawrence Univ.	edixon@stlawu.edu
Essig, David	Geneva College	dessig@geneva.edu
Estevez, Ana	St. Lawrence Univ.	aestevez@stlawu.edu
Ghosh, Sibdas	Dominican Univ, CA	sghosh@dominican.edu
Hammonds-Odie, Latanya	Dillard Univ.	lhammonds@dillard.edu
Harrison, Benjamin	College U of Alaska	afbrh@uaa.alaska.edu
Holmes, Keinya	Hampton Univ.	keinya.holmes@hamptonu.edu
Horst, Cynthia	Carroll College	chorst@carrollu.edu
Johnson-Brousseau, Sheila	Dominican Univ, CA	sajb109@gmail.com
Lodhi, Muhammad	Fayetteville St. Uiv.	mlodhi@uncfsu.edu
Louie, Maggie	Dominican Univ, CA	maggie.louie@dominican.edu
Lu, Yuefeng	Tougaloo College	ylu@tougaloo.edu
McCray, Joseph	Morehouse	jmccray@morehouse.edu
Nagengast, Alexis	Widener	nagengast@pop1.science.widener.edu
Osgood, Robert	Rochester Inst Tech	rcoscl@rit.edu
Peng, Chuang	Morehouse College	cpeng@morehouse.edu
Reyna, Nathan	Ouachita Baptist U.	reynan@OBU.EDU
Santisteban, Maria	UNC-Pembroke	maria.santisteban@uncp.edu
Vanderpuye, Oluseyi	Albany State Univ.	vanderpuye@asurams.edu
Woriax, Velinda	UNC-Pembroke	velinda.woriax@uncp.edu
Wright, Stephen	Carson-Newman Coll	swright@cn.edu

WORKSHOP #2 PARTICIPANTS

Name	• Institution	• Email Address
Bayline, Ronald	Wash. & Jeff. College	rjbayline@washjeff.edu
Bellin, Robert	Coll. Of the Holy Cross	rbellin@holycross.edu
DeBerry, Candy	Wash. & Jeff. College	cdeberry@washjeff.edu
Dwyer, Kathleen	Univ. Scranton	kgd301@scranton.edu
Fahy, Michael	Chapman Univ.	fahy@chapman.edu
Fuselier, Linda	MN St. Univ.	fuselier@mnstate.edu
Gordon, Ethel	NC A&T Univ.	ejgordon@ncat.edu
Hammond, Charlotte	Quinnipiac Univ	charlotte.hammond@quinnipiac.edu
Holgado, Andrea	SW OK State Univ.	andrea.holgado@swosu.edu
Hum-Musser, Sue	Western Ill. Univ.	sm-hum-musser@wiu.edu
Jenik, Pablo	Franklin and Marshall	pjenik@fandm.edu
Lanni, Jennifer	Wheaton College MA	lanni_jennifer@wheatoncollege.edu
LaRiviere, Frederick	Washington and Lee U.	larivieref@wlu.edu
Lee, Alice	Wash. & Jeff. College	aglee@washjeff.edu
Mukhtar, Hamid	NC A&T Univ	hdismail@ncat.edu
Pavao, Maura	Worcester State Coll	mpavao@worcester.edu
Rhode, Jennifer	UNC-Ashville	jrhode@unca.edu
Roig-Lopez, Jose	Universidad del Este (UNE)	joroig@suagm.edu
Rowland-Goldsmith, Melissa	Chapman Univ.	rowlandg@chapman.edu
Schisa, Jennifer	Central Michigan Univ.	schis1j@cmich.edu
Singh, Minati	Univ. Iowa	minati-singh@uiowa.edu
Villafane, Robert	Alabama State Univ	drbob523@yahoo.com rvillafane@alasu.edu
Ward, Gregg	Winston-Salem St. Univ.	wardgr@wssu.edu
Watson, Fiona	Washington and Lee Univ.	watsonf@wlu.edu

DNA MICROARRAY WORKSHOP FACULTY
SUMMER 2009

A. Malcolm Campbell, Ph.D.; Professor; Department of Biology; Davidson College; Davidson, NC 28036; Email: macampbell@davidson.edu

Todd T. Eckdahl, Ph.D.; Professor; Department of Biology; Missouri Western State University; St. Joseph, MO 64507-2294;Email: eckdahl@missouriwestern.edu

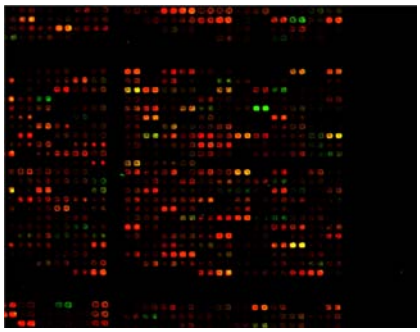
Laurie J. Heyer, Ph.D.; Associate Professor; Department of Mathematics; Davidson College; Davidson, N.C. 28035-6959; Email: laheyer@davidson.edu

Ann G. Rosenwald, Ph.D.; Assistant Professor; Department of Biology; NW, Georgetown University; Washington, DC 20057; Email: rosenwaa@georgetown.edu

Edison R. Fowlks, Ph.D.; Professor of Biology; Department of Biological Sciences; Hampton University, Hampton, VA 23669; Email: edison.fowlks@hamptonu.edu

Consuelo J. Alvarez, Ph.D.; Assistant Professor; Department of Biology; Longwood University; Farmville, VA; Email: alvarezcj@longwood.edu

Charles Hauser, Ph.D.; Associate Professor; Department of Biology; Saint Edwards College; Email: charlesh@stedwards.edu



2009 GCAT Microarray Workshops

Workshop #1 July 6 - 10

Morehouse College, Atlanta, GA

This will be the last GCAT Microarray Workshop

Workshop #1

[Link to Workshop #2]

Schedule Overview

	July 5	July 6	July 7	July 8	July 9	July 10	July 11
Day of Workshp	Pre-workshop	1	2	3	4	5	6
Workshop #1	arrive	Dry Lab	Wet am / Dry pm	Wet Lab	Dry Lab	Half day dry	gone
Workshop #2		arrive	Dry am / Wet pm	Dry Lab	Wet Lab	Dry Lab	Half day dry

2009 Instructors

Todd Eckdahl, Laurie Heyer, Anne Rosenwald, Consuelo Alvarez, Charles Hauser, Malcolm Campbell, Edison Fowlks

Day 1 (Monday July 6)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Last Minute Laptop Preparation	Laurie Heyer and Malcolm Campbell
8:30 - 9:00	Greetings and Overview	Edison Fowlks

9:00 - 10:15	Introduction to Microarrays, MAGIC Tool web, Raw Data, Online Tutorials Launch  	Malcolm Campbell (quick links to short and long animations) Laurie Heyer
10:15 - 10:30	Short Break	Dr. Pleez B. Prompt
10:30 - 11:15	Making a Project	Laurie Heyer & Consuelo Alvarez
11:15 - 11:35	Practice Making Projects	Laurie Heyer & Consuelo Alvarez
11:35 - 12:15	Slide Layout and Gene List	Laurie Heyer & Consuelo Alvarez
12:15 - 1:15	**Lunch**	
1:15 - 1:45	Practice Layout and Gene List	Laurie Heyer & Consuelo Alvarez
1:45 - 2:30	Addressing and Gridding	Laurie Heyer & Consuelo Alvarez
2:30 - 2:45	Short Break	
2:45 - 3:15	Practice Gridding	Laurie Heyer & Consuelo Alvarez
3:15 - 4:00	Segmentation	Laurie Heyer & Consuelo Alvarez
4:00 - 4:30	Practice Segmentation	Laurie Heyer & Consuelo Alvarez
4:30 - 5:00	Generate Multiple Ratio Expression Files	Laurie Heyer & Consuelo Alvarez
5:00 - 6:00	Practice Multiple Ratio Expression Files Quit MAGIC Tool	Laurie Heyer & Consuelo Alvarez
Goal for Day 1: Create 6 Columns of Data		
6:00 onwards	**Dinner**	

Schedule Overview

Day 2 (Tuesday July 7)

Time	*Activity*	*Lead Instructors*
8:30 - 11:00	cDNA prep with provided RNA (get started right away)	Anne Rosenwald & Todd Eckdahl
	Introductions and overviews (during incubation)	Anne Rosenwald & Todd Eckdahl
	get the slides ready for the first hybridization mix	Anne Rosenwald & Todd Eckdahl
11:00 - 11:30	RNA degradation using NaOH	Anne Rosenwald & Todd Eckdahl
11:30 - 12:00	cDNA concentration	Anne Rosenwald & Todd Eckdahl
12:00 - 12:30	Prepare hybridization mix 1 to go on arrays for overnight incubation	Anne Rosenwald & Todd Eckdahl
12:30 - 1:30	**Lunch**	
1:30 - 2:00	Log Transforming Data	Laurie Heyer & Consuelo Alvarez
2:00 - 2:15	Practice Log Transforming Data	Laurie Heyer & Consuelo Alvarez
2:15 - 2:45	Gene Information	Laurie Heyer & Consuelo Alvarez
2:45 - 3:00	Short Break and Practice Gene Information	
3:00 - 4:00	Explore Data	Laurie Heyer & Consuelo Alvarez
4:00 - 4:30	Practice Exploring	Laurie Heyer & Consuelo Alvarez
4:30 - 5:30	Clustering Overview (you practice later: online + PDF)	Laurie Heyer & Consuelo Alvarez

Goal for Day 2: Manipulate and Explore Data **Generate cDNA Probes & Hybe Them on Chips**		
6:00 - 7:30	**Dinner**	
7:30 - 8:30	TBA	

Schedule Overview

GCAT Wet Lab Protocols

Day 3 (Wednesday July 8)

Time	*Activity*	*Lead Instructors*
8:30 - 9:30	Washes, and set up 2nd hybridization mix	Anne Rosenwald & Todd Eckdahl
8:30 - 12:30	Second hyb begins	Anne Rosenwald & Todd Eckdahl
	prep RNA from frozen pellets with Ambion kit	Anne Rosenwald & Todd Eckdahl
	Run some on gels	Anne Rosenwald & Todd Eckdahl
12:30 - 1:30	**Lunch**	
12:30 - 5:30	Wash and scan chips	Anne Rosenwald & Todd Eckdahl
	Discussion of Quantifying mRNA Levels with RT PCR	Anne Rosenwald & Todd Eckdahl
	Discussion of Using Microarrays in Courses Spot Synthesizer	Anne Rosenwald & Todd Eckdahl
	Discussion of Using Microarrays in Research	Anne Rosenwald & Todd Eckdahl
	Discussion of Results	Anne Rosenwald & Todd Eckdahl
Goal for Day 3: Wash & Scan Chips **Discuss and Trouble Shoot Results**		
6:00 -	**Dinner**	

7:00		
7:30 - 8:30	TBA	

GCAT Wet Lab Protocols

Schedule Overview

Day 4 (Thursday July 9)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Refresh Memory	Laurie Heyer & Consuelo Alvarez
8:30 - 10:15	Generate Ratios from Your Microarray	You
10:15 - 10:30	Short Break	
10:30 - 12:15	1 Column Exploration	Laurie Heyer & Consuelo Alvarez
12:15 - 1:15	**Lunch**	
1:15 - 3:00	Make a Series from Separate 1 Columns	Laurie Heyer & Consuelo Alvarez
3:00 - 3:15	Short Break	
3:15 - 6:00	DeRisi Data Analaysis	Laurie Heyer & Consuelo Alvarez
Goal for Day 4: Know what to do after wet lab!		
6:00 - 7:00	**Dinner**	
7:30 - 8:30	TBA	

Schedule Overview

Day 5 (Friday July 10)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Summarize Past 4 Days	**GCAT** Instructors
8:30 - 10:15	Discuss Curriculum Ideas	**GCAT** Instructors
10:15 - 10:30	Short Break	Dr. Juan Las Time
10:30 - 12:15	Your Quesitons and Workshop Assessment	**GCAT** Instructors
12:15 - 1:15	**Lunch and depart**	Eat at Cafeteria
Goal for Day 5: Know how you can adapt and adopt GCAT resources.		

Schedule Overview

Email Edison Fowlks or Malcolm Campbell with Questions

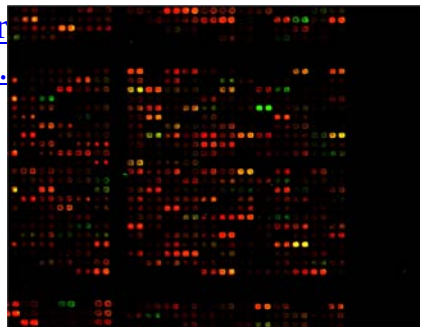
2009 Workshop Main Page

This material is based upon work supported by the National Science Foundation under Grant No. DBI-0627478. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

GCAT Home Page

Send comments, questions, and suggestions to: **macampbell@davidson.edu**

Click to see a larger version of this image.



2009 GCAT Microarray Workshops

Workshop #2 July 7 - 11

Morehouse College, Atlanta, GA

This will be the last GCAT Microarray Workshop

Workshop #2

Link to Workshop #1

Schedule Overview

	July 5	July 6	July 7	July 8	July 9	July 10	July 11
Day of Workshp	Pre-workshop	1	2	3	4	5	6
Workshop #1	arrive	Dry Lab	Wet am / Dry pm	Wet Lab	Dry Lab	Half day dry	gone
Workshop #2		arrive	Dry am / Wet pm	Dry Lab	Wet Lab	Dry Lab	Half day dry

2009 Instructors

Todd Eckdahl, Laurie Heyer, Anne Rosenwald, Consuelo Alvarez, Charles Hauser, Malcolm Campbell, Edison Fowlks

Day 1 (Tuesday July 7)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Last Minute Laptop Preparation	Laurie Heyer and Malcolm Campbell
8:30 - 9:00	Greetings and Overview	Edison Fowlks
9:00 - 10:15	Introduction to Microarrays, MAGIC Tool web, Raw Data, Online Tutorials Launch 	Malcolm Campbell (quick links to short and long animations) Laurie Heyer & Consuelo Alvarez
10:15 - 10:30	Short Break	
10:30 - 11:15	Making a Project	Laurie Heyer & Consuelo Alvarez
11:15 - 11:35	Practice Making Projects	Laurie Heyer & Consuelo Alvarez
11:35 - 12:15	Slide Layout and Gene List	Laurie Heyer & Consuelo Alvarez
12:15 - 1:30	**Lunch**	
1:30 - 4:00	cDNA prep with provided RNA (get started right away)	Charles Hauser & Anne Rosenwald
	Introductions and overviews (during incubation)	Charles Hauser & Anne Rosenwald
	get the slides ready for the first hybridization mix	Charles Hauser & Anne Rosenwald
4:00 - 4:30	RNA degradation using NaOH	Charles Hauser & Anne Rosenwald
4:30 - 5:00	cDNA concentration	Charles Hauser & Anne Rosenwald
5:00 - 5:30	Prepare hybridization mix 1 to go on arrays for overnight incubation	Charles Hauser & Anne Rosenwald
Goal for Day 1: Start MAGIC Tool Start Wet Lab		

6:00 - 7:30	**Dinner**	
7:30 - 8:30	TBA	

Schedule Overview

Day 2 (Wednesday July 8)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Practice Layout and Gene List	Laurie Heyer & Consuelo Alvarez
8:30 - 9:15	Addressing and Gridding	Laurie Heyer & Consuelo Alvarez
9:15 - 9:45	Practice Gridding and Break if you need one	Laurie Heyer & Consuelo Alvarez
9:45 - 10:30	Segmentation	Laurie Heyer & Consuelo Alvarez
10:30 - 11:00	Practice Segmentation	Laurie Heyer & Consuelo Alvarez
11:00 - 11:30	Generate Multiple Ratio Expression Files	Laurie Heyer & Consuelo Alvarez
11:30 - 12:30	Practice Multiple Ratio Expression Files	Laurie Heyer & Consuelo Alvarez
12:30 - 1:30	**Lunch**	
1:30 - 2:00	Log Transforming Data	Laurie Heyer & Consuelo Alvarez
2:00 - 2:15	Practice Log Transforming Data	Laurie Heyer & Consuelo Alvarez
2:15 - 2:45	Gene Information	Laurie Heyer & Consuelo Alvarez
2:45 - 3:00	Short Break and Practice Gene Information	

3:00 - 4:00	Explore Data	Laurie Heyer & Consuelo Alvarez
4:00 - 4:30	Practice Exploring	Laurie Heyer & Consuelo Alvarez
4:30 - 5:30	Clustering Overview (you practice later: online + PDF)	Laurie Heyer & Consuelo Alvarez
	Goal for Day 2: Manipulate and Explore Data	
6:00 - 7:00	**Dinner**	Eat at Cafeteria
7:30 - 8:30	TBA	

Schedule Overview

GCAT Wet Lab Protocols

Day 3 (Thurday July 9)

Time	*Activity*	*Lead Instructors*
8:30 - 9:30	Washes, and set up 2nd hybridization mix	Charles Hauser & Anne Rosenwald
8:30 - 12:30	Second hyb begins	Charles Hauser & Anne Rosenwald
	prep RNA from frozen pellets with Ambion kit	Charles Hauser & Anne Rosenwald
	Run some on gels	Charles Hauser & Anne Rosenwald
12:30 - 1:30	**Lunch**	
12:30 - 5:30	Wash and scan chips	Charles Hauser & Anne Rosenwald
	Discussion of Quantifying mRNA Levels with RT PCR	Charles Hauser & Anne Rosenwald
	Discussion of Using Microarrays in Courses Spot Synthesizer	Charles Hauser & Anne Rosenwald
	Discussion of Using Microarrays in	Charles Hauser & Anne

	Research	Rosenwald
	Discussion of Results	Charles Hauser & Anne Rosenwald
Goal for Day 3: Wash & Scan Chips **Discuss and Trouble Shoot Results**		
6:00 - 7:00	**Dinner**	
7:30 - 8:30	TBA	

GCAT Wet Lab Protocols

Schedule Overview

Day 4 (Friday July 10)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Refresh Memory	Laurie Heyer & Consuelo Alvarez
8:30 - 10:15	Generate Ratios from Your Microarray	You
10:15 - 10:30	Short Break	
10:30 - 12:15	1 Column Exploration	Laurie Heyer & Consuelo Alvarez
12:15 - 1:15	**Lunch**	
1:15 - 3:00	Make a Series from Separate 1 Columns	Laurie Heyer & Consuelo Alvarez
3:00 - 3:15	Short Break	
3:15 - 6:00	DeRisi Data Analaysis	Laurie Heyer & Consuelo Alvarez

Goal for Day 4: Know what to do after wet lab!		
6:00 - 7:00	**Dinner**	
7:30 - 8:30	TBA	

Schedule Overview

Day 5 (Saturday July 11)

Time	*Activity*	*Lead Instructors*
8:00 - 8:30	Summarize Past 4 Days	**GCAT** Instructors
8:30 - 10:15	Discuss Curriculum Ideas	**GCAT** Instructors
10:15 - 10:30	Short Break	Dr. Juan Las Time
10:30 - 12:15	Your Quesitons and Workshop Assessment	**GCAT** Instructors
12:15 - 1:15	**Lunch and depart**	Eat at Cafeteria
Goal for Day 5: Know how you can adapt and adopt GCAT resources.		

Schedule Overview

WORKSHOP ASSESSMENT

Anonymous Pre-Workshop Assessment for
MOREHOUSE GCAT Workshop, summer 2008

Please supply your responses to these questions. We are interested in your thoughts and concerns prior to the workshop so we can compare these with your responses after the workshop. **Circle the best response**. Space is provided at the bottom of the survey for additional comments.

1) I have performed experiments with DNA microarrays before this workshop.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

2) I have successfully generated usable data with DNA microarrays before this workshop.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

3) I have covered DNA microarrays in lecture prior to this workshop (about _____ minutes of class time.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

4) I have a general knowledge about DNA microarrays, but not enough to teach about them in any depth.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

5) I am uncertain how to analyze DNA microarray data.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

6) DNA microarrays are too expensive for me to use in my curriculum.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

7) I would be more likely to teach a new method (e.g. DNA microarrays) if I had a support network of colleagues.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

8) My institution would appreciate me bring a new, genomics method into the curriculum.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

9) I am concerned that I may not learn enough at this workshop to teach DNA microarrays.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

10) I was forced/pressured by a superior to attend this workshop.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

11) I doubt I have the computer power to perform DNA microarray analysis.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

12) The hardest part of DNA microarray experiments is data production.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

13) My students would benefit from learning more about DNA microarrays.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

==========

Rank these criteria in priority from most important to least important. You may use a number more than once (e.g. there may be two equally rated "most important" criteria). You may skip numbers too. If you have already used DNA microarrays, rate them according to reasons your used of DNA microarrays in your curriculum is limited.

1 = most important and 6 = least important.

____ Departmental budget is why I have not used DNA microarrays in my curriculum.
____ "I am too busy already." is why I have not used DNA microarrays in my curriculum.
____ Lack of training is why I have not used DNA microarrays in my curriculum.
____ My level interest is why I have not used DNA microarrays in my curriculum.
____ No space in the curriculum is why I have not used DNA microarrays in my curriculum.
____ Intimidating technology is why I have not used DNA microarrays in my curriculum.

Please share any additional comments or thoughts:

Anonymous Post-Workshop Assessment for
MOREHOUSE GCAT Workshop, summer 2009

Please supply your responses to these questions. We are interested in your thoughts and impressions after the workshop so we can evaluate the effectiveness of the workshop. **Circle the best response**. Space is provided at the bottom of the survey for additional comments.

I was in: workshop #1 (started Wed., July 27), or workshop #2 (started Thurs., July 28).

1) My lab group successfully generated usable data with DNA microarrays.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

2) I successfully analyzed DNA microarray data during the workshop.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

3) I know enough about DNA microarrays now that I could include them in my curriculum.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

4) With the help of GCAT, I could afford DNA microarrays in my curriculum.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

5) The workshop provided me with enough confidence to include DNA microarrays in my curriculum.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

6) I will include DNA microarrays in my curriculum for 2005 – 2006 academic year. (about how many students would be affected?_____)

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

7) I want to include DNA microarrays in my curriculum but cannot this coming year.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

8) It was helpful/comforting to have other faculty learning with me in the workshop.

Strongly agree	**Somewhat agree**	**Neutral opinion**	**Somewhat disagree**	**Strongly disagree**

9) Meeting faculty from other schools increased the probability of me adding DNA microarrays to my curriculum.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

10A) After the workshop, I am more likely to add a data analysis dry-lab to my curriculum.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

10B) After the workshop, I am more likely to add a data production wet-lab to my curriculum.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

11) I will try to create space in the curriculum to add DNA microarrays to my curriculum.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

12) A letter from GCAT might convince my administrators I need more computer power.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

13) The hardest part of DNA microarray experiments is data production.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

14) My students would benefit from learning more about DNA microarrays.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

15) The quality of the data analysis experience was very good.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

16) The quality of the wet lab experience was very good.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

17) If I analyzed data, I will use MAGIC Tool.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

18) I think more faculty would like to attend this type of workshop.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

19) I would recommend a GCAT workshop to my friends.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

20) It was good to meet the NSF program officers.

Strongly agree	Somewhat agree	Neutral opinion	Somewhat disagree	Strongly disagree

On a scale of 1 – 10, how would you rate the workshop logistics? (1 = great; 10 = terrible)

Application process	
Housing	
Food	
Wet lab facilities	
Computer lab facilities	

Comments or suggestions about the workshop (use back if necessary):

GCAT Microarray:
Culturing Yeast, Isolating RNA, 3DNA Protocol

Written by Anne Rosenwald and Todd Eckdahl (Spring 2008)
Thanks to Mary Lee Ledbetter and Dave Kushner who provided resource materials and Consuelo Alvarez and Chuck Hauser for comments

Note that all of the information here, as well as ideas for implementation in a variety of classes and formats can be found at:

http://www.bio.davidson.edu/projects/GCAT/GCAT.html

We also urge you to join the GCAT Listserv – this community of undergraduate teachers is a great source of information and ideas.

Supply List

S288C yeast strain
YPD media (recipe is given here on p. 2, but more details on GCAT web site)
RNA isolation kit (we're using the Ambion kit – www.ambion.com)
Agarose
1x Tris-Borate-EDTA (1x TBE), pH 8.3 (89 mM Tris-borate, 2 mM EDTA)
Ethidium bromide (10 mg/ml)
RNA standards
Genisphere Array 350 Ex kit (www.genisphere.com)
Reverse transcriptase (which can be ordered with the Genisphere kit)
100 mM DTT
Molecular Biology Grade Ethanol
3 M sodium acetate, pH 5.2
1 M Tris-HCl, pH 7.5
0.5 M NaOH / 50 mM EDTA
1x Tris/EDTA (1x TE), pH 8.0 (10 mM Tris / 1 mM EDTA)
Microcon YM-30 concentrators (www.millipore.com)
Yeast microarrays
Sonicated salmon sperm DNA
20x SSC
SDS (10% stock)

Equipment List

Shaking incubator to grow yeast cultures
Clinical centrifuge
Microcentrifuges
Vortex mixers
UV/visible spectrophotometer

Agarose gel electrophoresis apparatus and power supply
Hybridization chambers
Dry incubator or water bath
Heating blocks (not crucial if water baths and incubators are available, but handy)
Access to microarray scanner (i.e. *via* GCAT – see the web site for details)
Speedvac
Micropipettors, including ones that will measure out 1-2 µl volumes
Sterile, RNAse free tips appropriate for your micropipettors

Culturing Yeast

A strain of yeast commonly used by researchers is S288C. Most GCAT members also use this strain and share it with other members of the community. Information about S288C, as well as great information about other strains of yeast, yeast genes, yeast researchers, and other yeast resources, can be obtained at the Saccharomyces Genome Database (http://www.yeastgenome.org/straintable.shtml#S288C). Dave Kushner (Dickinson College) has a detailed protocol for growing yeast on the GCAT web site.

DeRisi *et al.* used S288C in collecting microarray data on the yeast diauxic shift from anaerobic to aerobic metabolism. Below is a procedure by which yeast can be cultured and harvested to measure the effect of this diauxic shift on gene expression. The goal is to collect yeast at a point early in the growth curve and at one or more later points.

1. Transfer a colony or loop of S288C yeast to 5 ml YPD (5 g yeast extract, 10 g peptone, 10 g dextrose in 500 ml, autoclave) and incubate at 30°C with shaking overnight.

2. Transfer 1 ml of overnight into 200 ml YPD in a 500 ml flask and incubate with shaking at 30°C.

3. Check the absorbance at 600 nm after about eight hours and determine if it is the value you want to have for the early time point. If so, collect volumes of yeast culture in separate tubes that correlate with the capacity of the RNA isolation kit. The table shows some estimates for this. For example, 167 ml of yeast at an A_{600} of 0.14 yields 3×10^8 cells, the capacity of the Ambion RiboPure Yeast kit described below.

A_{600}	**cells / ml**
0.14	1.8×10^6
0.46	6.3×10^6
0.80	1.1×10^7
1.8	2.3×10^7
3.7	4.8×10^7
6.9	1.0×10^8
7.3	1.2×10^8

4. Spin the yeast culture samples in a clinical centrifuge for 10 minutes at 4000 rpm.

5. Pour off the supernatant and either refrigerate the pellets for use soon or freeze them for use later.

6. Repeat steps 3-5 for one or more later time points. Reaching an A_{600} of 6.9 may take 12-16 hours.

Prepare Total Yeast RNA

RNA can be prepared in any number of ways. Check the GCAT web site for options used by instructors and other experienced microarrayers for ideas. Dr. Dave Kushner (Dickinson College) has a protocol on the GCAT web site that involves an extraction with hot phenol. There are also a number of kits available, including the Total RNASafeKit from QBiogene, the RNeasy kit from Qiagen, and the RiboPure Yeast kit from Ambion. The kit we're using is the one from Ambion.

Ambion RiboPure Yeast Protocol

Implementation note: This protocol, from yeast pellets, takes about an hour.

1. To pellets with no more that 3×10^8 cells, add the following and resuspend the cells by vigorous vortexing for 15 seconds.

   480 µl lysis buffer
   48 µl 10% SDS
   480 µl phenol/chloroform/IAA

2. Pour about 750 µl of cold Zirconia beads into a provided 1.5 ml screw cap tube, transfer the yeast cell mixture into the tube, and vortex for 10 minutes.

3. Spin 5 minutes at 16,000 x g to separate aqueous and organic phases.

4. Carefully transfer top aqueous phase to a 15 ml tube without disturbing the interface.

5. Preheat 60 µl elution solution for each isolation to about 95°C.

6. Add 1.9 ml binding buffer to the sample in the 15 ml tube and mix thoroughly.

7. Add 1.25 ml 100% ethanol to the sample in the 15 ml tube and mix thoroughly.

8. Add 700 µl of the sample to a filter cartridge placed onto a supplied collection tube.

9. Centrifuge for 30 sec at 14,000 x g.

10. Discard flow-through

11. Repeat steps 8 and 9 until the entire sample has been applied to the filter cartridge.

    (Note: steps 8-11 can also be accomplished by fitting the filter cartridge onto a 5 ml syringe barrel connected to a vacuum source.)

12. Wash the filter by adding 700 µl wash solution 1 and centrifuging for 1 minute at full speed. Discard flow-through.

13. Wash by adding 500 µl wash solution 2/3 and centrifuging at full speed. Discard flow-through. Repeat this step.

14. Centrifuge for 1 minute at full speed to dry filter.

15. Transfer filter to clean 2 ml collection tube.

16. Add 25 µl elution solution, preheated to about 95^{o}C, to the center of the filter.

17. Centrifuge at full speed for 1 minute.

18. Repeat steps 16 and 17. Isolated total RNA should be in a volume of about 50 µl.

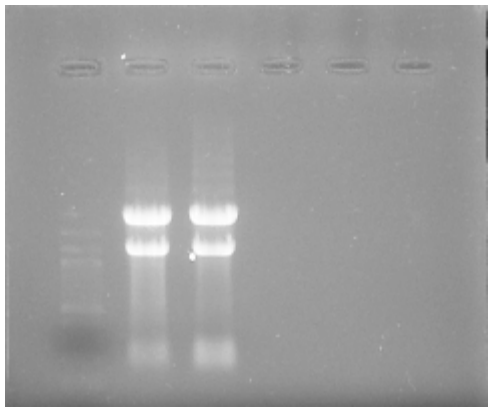
Checking Quality and Quantity of RNA

Important implementation note: Although the Genisphere protocol we describe below is fairly robust, even with less than ideal RNA samples, we strongly recommend checking your RNA by gel before proceeding. It's better and ultimately cheaper to make more RNA than to waste a microarray with bad samples.

1. Measure RNA with UV spectrophotometer at 260 and 280 nm. The ratio of A_{260} to A_{280} should be between 1.8 and 2.2. This is an indication of the amount of protein contamination (including RNases) present in the samples. The RNA absorbs mostly at the lower wavelength while proteins absorb at the higher one. The concentration of RNA can also be estimated by the A_{260} using the following formula:

   RNA Concentration = A_{260} x dilution factor x 40 ng/µl

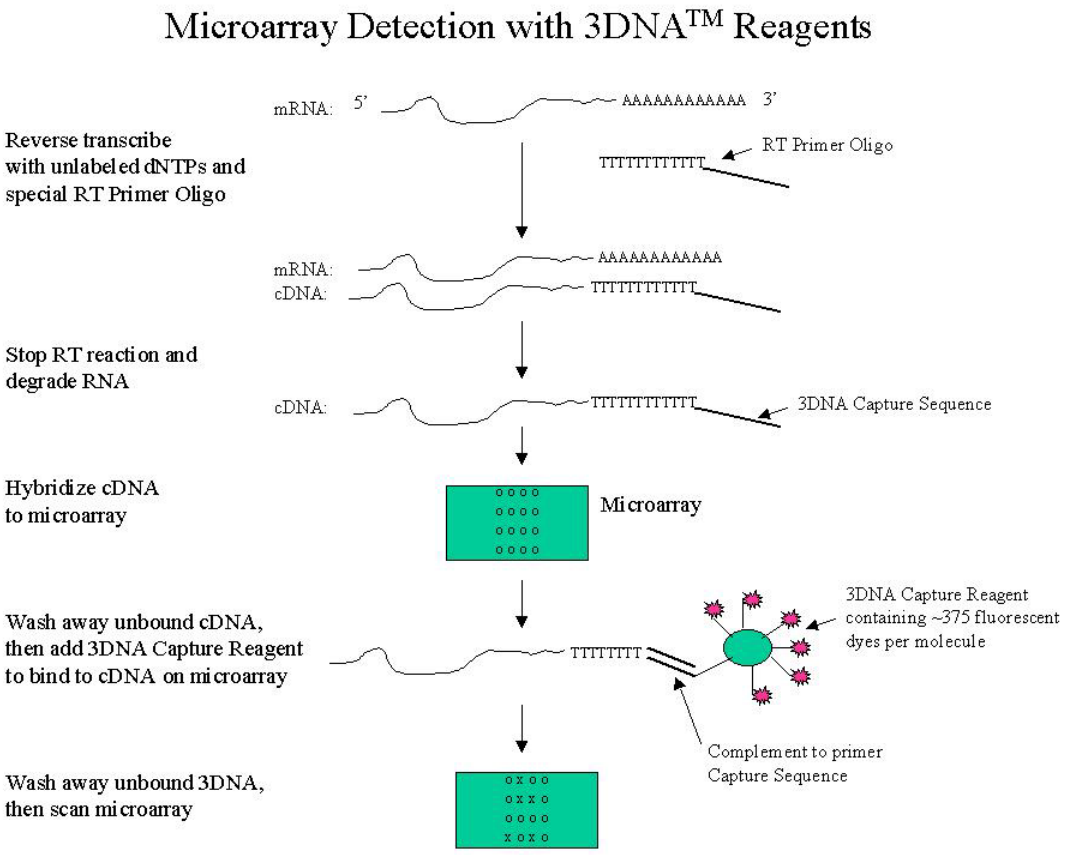
2. Check quality by running 1 µg on a 1% agarose gel (denaturing loading buffers may be used). Ethidium bromide may be added to gel and buffer at 0.5 µg/ml for staining. As seen at right, two bright rRNA bands should be visible on a background smear of mRNA with a minimum of small molecular weight degradation products.



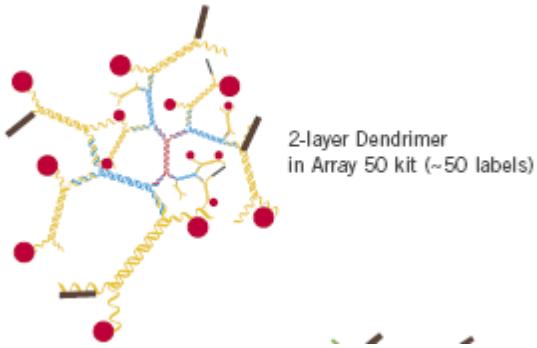
3. Precipitate 10 µg aliquots of RNA for use in labeling procedure using 1/10 volume 3M sodium acetate, pH 5.2, and 2 volumes ethanol. Centrifuge at high speed for 10 minutes and air dry pellets (or use speed vac).

Labeling of Microarrays with Genisphere Array 350 Kit

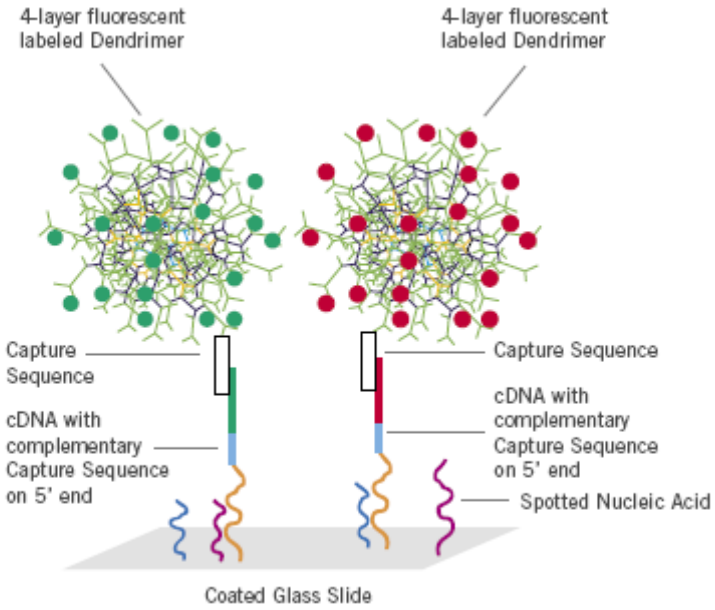
This protocol is based on the Array 350 kit – other kits are available, differing in number of fluorophores attached to the dendrimer. The general web address is www.genisphere.com. This site has good pictures – take a look!



The capture reagents are dendrimers, incorporating many fluorophores into one capture molecule.



Two different cDNA preparations will each have a capture tag – each capture tag will bind to a different dendrimer tag, so that you will be able to visualize the amount of red and green fluorescence of the two cDNAs in the two different conditions.



Important Note:

The two dyes we are using for this experiment are called Cy3 and Cy5

Cy3 absorbs blue (looks blue to your eye) but fluoresces red.
Cy5 absorbs pink (looks pink to your eye) but fluoresces green.

Prepare cDNA with capture tag

Implementation note: These steps (cDNA synthesis and RNA degradation) can easily be accomplished in a 3 hour lab period. After the RNA degradation and concentration steps, store the cDNAs at -80°C.

Your RNA samples will be provided as dry pellets containing 10 μg in microfuge tubes (separate tubes for the two samples)

cDNA synthesis

1. To each sample add (in this order):

   10 μl DEPC-treated water (vial 10 from the Genisphere kit)
   1 μl RT primer (vial 2)
      One gets the Cy5 ***capture*** primer (blue) the other Cy3 ***capture*** (red). *Be sure to write down which gets which here!*
   1 μl SuperaseIn (vial 4)

2. Mix by flicking the tube gently and flash spin.

3. Incubate **10 minutes at 75 – 80°C in the heating block**.

4. Incubate **2 minutes on ice.**

5. Prepare a "master mix" for Reverse Transcriptase
   For the pair of reactions add the following in order
      8 μl Reverse Transcriptase Buffer (extra vial)
      2 μl dNTP mix (vial 3)
      4 μl DTT (100 mM)
      2 μl Reverse Transcriptase Enzyme

6. Gently mix and flash spin, then aliquot 8 μl of the mix to each of your 2 tubes – the total volume/tube should now be 20 μl.

7. Incubate **2 h at 42°C.**

RNA degradation

1. To each tube: Add 3.5 μl 0.5 M NaOH / 50 mM EDTA to stop the reaction and degrade the RNA.

2. Incubate **10 minutes at 65°C in heating block**

3. Add 5 μl 1 M Tris-HCl, pH 7.5 to neutralize.

4. Mix the contents of both tubes together into one.

5. Rinse the now empty tube with 73 µl 1x Tris-EDTA, pH 8.0 (TE buffer) and combine with mixed cDNA samples. Total volume now should be 130 µl.

Alternative method to degrade RNA

1. Make sure contents of tubes are spun down.

2. Add 1 µl of RNase cocktail (RNase A at 4 mg/ml and RNase H at 1 unit/µl). (Note: the RNase H is fairly expensive and may be omitted without significant effect)

3. Incubate at 37^{o}C for 15-30 min

4. Mix the contents of both tubes together into one.

5. Rinse the now empty tube with 88 µl 1x Tris-EDTA, pH 8.0 (TE buffer) and combine with mixed cDNA samples. Total volume now should be 130 µl.

Concentrating the cDNA
(read these instructions carefully before proceeding so you know what to do!)

1. Use a Microcon YM-30

2. Prepare the concentrator by adding 100 µl 1x TE, pH 8.0 to the reservoir

3. Spin **3 minutes at 13K rpm**.

4. Transfer your sample (130 µl) to the reservoir.

5. Spin **9 minutes at 13K rpm.**

6. Remove the reservoir – <u>your cDNA is now concentrated *on the membrane*</u>

7. Add 5 µl of 1x TE, pH 8.0 to the reservoir *without* touching the membrane.

8. Carefully *invert* reservoir over a fresh tube and spin **2 minutes at 13K rpm.**

9. Carefully measure the amount of liquid recovered with a micropipettor (should be 3-10 µl).

10. Put the sample back in this tube (*Implementation note: stopping point if necessary – freeze at -80^{o}C*)

First Hybridization

Slide Preparation*

1. Incubate microarray slides for at least 60 minutes in **3x SSC, 0.1% SDS, and 0.1 mg/ml sonicated salmon sperm DNA** at room temperature (this serves to block the non-specific sites on the slides).

2. Carefully dip the slides into distilled water and spin dry in a 50 ml conical tube with a kim-wipe in the bottom to collect the drips. Put the slide label side down to avoid scratching the array with the kim-wipe.

Alternative procedure for slide preparation:

After the incubation, blow the slide dry with air (connect a piece of tubing with a filter tip on the end to house air or to a nitrogen tank). Don't blast the slide too hard or too long. Rather, the idea is to chase all the drops of water off the slide while it is held at an angle on a paper towel. If drops of water start to dry in place on the array, quickly immerse the slide back into water and start again. You are not trying to blow dry the slide, rather you are trying to push the liquid away from the spots. If you see streaks at this stage, rewet the slide. If you see dried-on streaks, you will have streaks in your final scan. Store dried slide in a conical tube until ready to hybridize.

3. Thaw the **2x formamide-based hybridization buffer** (vial 7). Incubate at 55^{o}C for 10 minutes and mix well to make sure all crystals dissolve, then spin for 1 minute at 13K rpm to remove any residual particles. ***Don't shake it up again!***

4. Add enough DEPC-treated water to your concentrated sample to make 25 µl total.

5. Add 25 µl vial 7

6. Mix gently by flicking the tube and flash spin.

7. Incubate at **10 minutes at 80^{o}**, then keep at 42^{o}C until ready to put on slide

8. Transfer the entire cDNA sample (50 µl) to the center of your slide, carefully making a line down the length of the slide (*don't touch the slide with pipet tip! don't introduce bubbles into the solution! – better to lose some of the solution than have bubbles*).

9. Place the short edge of the cover slip on the short edge of the slide. Gently lower the cover slip onto the liquid with a syringe needle, but don't let it fall all the way down. Pull the cover slip back up, then lower with the needle

again, and this time, gently let the cover slip fall into place. Be very careful to *avoid bubbles* at this stage, too.

Fig. 1. Using needle to carefully lower Lifter Slip onto microarray using mock slide as guide (pictures from the GCAT web site courtesy of Dr. David Kushner [Dickenson College])

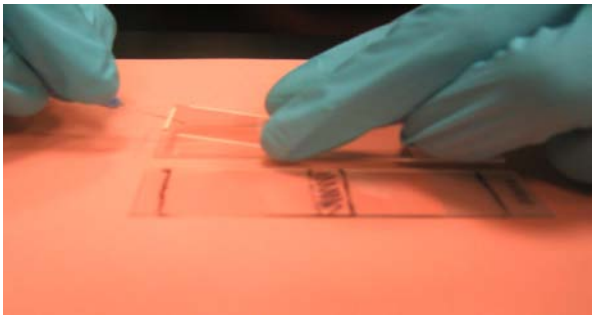


Fig. 2. First Lifter Slip in place on array next to mock slide.

10. Place the slide into either a slide hybridization chamber or a 50 ml conical tube. Be sure to include some water in either case to keep the hybridization mix from evaporating (the chambers have wells that hold ~20 µl of water; for the conical tube, just add ~50 µl to the tube).

11. Incubate the slide **at 37^{o}C overnight.** (*Implementation note: this step can go for several days if necessary, but check once in awhile to make sure there is still some water in the chamber).*

<u>Second hybridization</u>

Tubes have been prepared for you containing the wash solutions. Read the labels ***<u>carefully</u>*** as they all contain similar ingredients, but at different concentrations.

Part I – washing the first hybridization mix off the slide

1. Carefully transfer your slide, coverslip side down, into a tube containing **Room Temperature 2x SSC + 0.2% SDS**. Slosh gently until the coverslip comes off – it may need a little nudge; carefully remove your slide and leave the coverslip behind. (You're trying to avoid scratching the array with the cover slip)

2. Transfer the slide into **55°C 2x SSC + 0.2% SDS** and **incubate for 15 min at 55°C.**

3. Transfer the slide into **2x SSC** and **incubate for 15 min at room temperature** (shake gently periodically)

4. Transfer the slide into **0.2x SSC** and **incubate for 15 min at room temperature** (shake gently periodically).

5. Place the slide *label end first* in a 50 ml tube with a kim-wipe at the bottom and **spin 1 minute at 500 rpm** (alternatively use the blow-dry method)

Part II – preparing the 2nd hybridization mix

1. Thaw 2x-formamide based hybridization buffer (vial 7) and incubate **10 minutes at 55°C.** Then spin 1 min.

LIGHT SENSITIVE STEPS FROM NOW ON!

2. Thaw both capture reagents (both vials 1, cover with foil!!!) and the antifade reagent (vial 8).

3. Mix 100 µl hybridization buffer (vial 7) and 1 µl antifade reagent (vial 8). This is your antifade-treated hybridization mix.

4. Vortex the vials 1 for 3 seconds then flash spin.

5. Mix:
   25 µl antifade-treated hyb mix (your mix of vial 7 and vial 8)
   20 µl DEPC-water (vial 10)
   2.5 µl Cy3 capture reagent (vial 1)
   2.5 µl Cy5 capture reagent (alternate vial 1)

6. Incubate **10 min at 75°C.** Then pipet 50 µl onto the washed and dried slide using the coverslip protocol from yesterday.

7. Incubate your slide at **37°C for 2.5h. (***Implementation note: this incubation can go overnight if necessary).*

Second Wash and Scan

These solutions should have 1 mM DTT added to them to prevent oxidation of the fluorescent dyes. ***Remember to keep your slide in the dark as much as possible!!!*** Keep your tubes covered with foil during the incubations to help.

Washing the second hybridization mix off the slide

1. Carefully transfer your slide, coverslip side down, into a tube containing **Room Temperature 2x SSC + 0.2% SDS + 1 mM DTT**. Slosh gently until the coverslip comes off – it may need a little nudge; carefully remove your slide and leave the coverslip behind. (You're trying to avoid scratching the array with the cover slip)

2. Transfer the slide into **55°C 2x SSC + 0.2% SDS + 1 mM DTT** and **incubate for 15 min at 55°C.**

3. Transfer the slide into **2x SSC + 1 mM DTT** and **incubate for 15 min at room temperature** (shake periodically)

4. Transfer the slide into **0.2x SSC + 1 mM DTT** and **incubate for 15 min at room temperature** (shake periodically).

5. Place the slide *label end first* in a 50 ml tube with a kim-wipe at the bottom and **spin 1 minute at 500 rpm** (or use the blow dry method).

Store washed and dried slides in a conical tube covered with foil until ready to scan.

*Slide Pretreatment

Some slides need a pretreatment, others, such as those prepared for GCAT at Washington University, including these yeast arrays, do not. This process is used to redistribute the DNA on the slides, and helps with spot morphology and hybridization.

1. Steam the DNA side of the slide over boiling dH_2O. Do not allow visible droplets to form on the slide.

2. Immediately place the slide (DNA side up) on a heat block or hot plate set to 100°C or slightly less to snap dry. Take off after 5 seconds.

3. Repeat steam step, followed by drying step. Allow the slide to sit on the heat block for 1 minute this time. Allow slide to cool to room temperature.

Alternate Slide Pretreatment Protocol (from Charles Hauser, St. Edwards University)

1. Create a humidifier by placing slide (DNA side down) over warm water. Cover to create a humidified compartment (see Figure 3 below).

2. Place the slide (DNA side up) on a heat block or hot plate to 100^{o}C or slightly less to snap dry. Take off after 5 seconds.

3. Repeat humidified step, followed by drying step. Allow the slide to sit on the heat block for 1 minute. Allow slide to cool to room temperature.

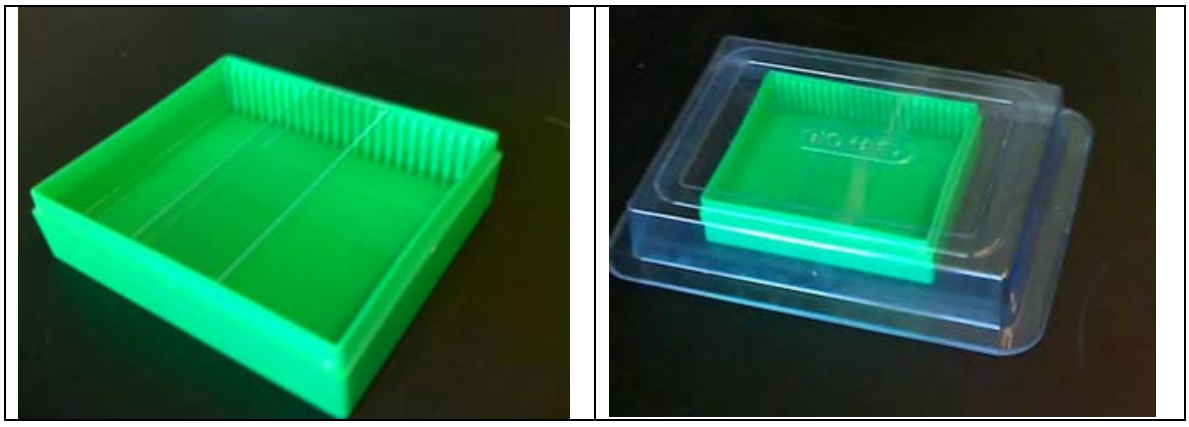


Figure 3: H

Installing Java

If you intend to install MagicTool on your computer, you will need the Sun version of Java.

a. On a PC, go to http://java.com, and click on "Get It Now". The correct version of the code should download and begin to install automatically. Complete installation instructions, with screen shots, are at http://java.com/en/download/help/win_auto.jsp .
b. On Mac, Java is already included. (OS X 10.2.6 or later is required.)

Installing MagicTool

1. Create a folder called magictool on your computer.
   a. On a PC, we recommend that you place this folder directly under the C: drive, or a folder within the C: drive that has no spaces in its name. This specifically excludes, for example, the Desktop (which is under "Documents and Settings") or "Program Files."
   b. On a Mac, we recommend that you place this folder in any convenient place, but **not** under a folder that contains spaces in its name.

2. Go to www.bio.davidson.edu/magic, and follow the links to download the software, or go directly to http://www.bio.davidson.edu/projects/magic/agreement.html.

3. Download the file called MagicTool.jar to the magictool folder you created in step (1).

4. On most operating systems, you can now run MagicTool by simply double-clicking on the file MagicTool.jar. You will know it is starting when the magic wand waves across the MagicTool logo on the screen. If the jar file does not start the program, it may be because Java is not installed properly on your machine. Be sure you have completed the instructions for Installing Java before proceeding to step (5).

5. This step sets up a script so you can run MagicTool with extra memory. This is important for working with large data files.
   a. On a PC:
      i. Download the file called Magic_launch.bat to the magictool folder.
      ii. Double click on Magic_launch.bat. This should cause a DOS command window to open, then MagicTool should start.
   b. On a Mac:
      i. Download the file called Magic_Launch.txt
      ii. Open Script Editor (this program is under the Applications, AppleScript folder).

iii. In Script Editor, go to Open Script… under the File menu. Select the file Magic_Launch.txt. The text from the file should now appear in the lower window in Script Editor.
iv. In Script Editor, go to Save As under the File menu, remove the .txt extension from the filename, select Application under the Format menu, and select the magictool folder as the destination.
v. Open a finder window and open the magictool folder, but do not use favorites, shortcuts or the multicolumn finder display. (In other words, use View as Icons or View as List.) Double click on Magic_Launch, which should start MagicTool.

MicroArray Genome Imaging and Clustering Tool

MAGIC Tool
User Guide

MAGIC Tool v2.1

July 27, 2007

The Goal of MAGIC Tool

The purpose of MAGIC Tool is to allow the user to begin with DNA microarray tiff files and end with biologically meaningful information. Comparative hybridization data (glass chips) and Affymetrix data are compatible with MAGIC Tool. You can start with tiff files or expression files (spreadsheet of ratios or absolute expression levels).

MAGIC Tool was created with the novice in mind but it is not a "dumbed down" program. In fact, MAGIC Tool is designed to illuminate the algorithms being used rather than be a black box that produces results with little input from the user. MAGIC Tool allows the user to change parameters for clustering, data quantification etc. This User's Guide will teach you how to use the software but leaves the theoretical explanations to the Instructor's Guide.

Users are also encouraged to visit related sites:

MAGIC web site: http://www.bio.davidson.edu/MAGIC
Online MAGIC Tool lab: http://gcat.davidson.edu/GCAT/workshop2/derisi_lab.html
Tutorial for Clustering: http://gcat.davidson.edu/DGPB/clust/home.htm
GCAT: http://www.bio.davidson.edu/GCAT/
Genomics Course: http://www.bio.davidson.edu/genomics

MAGIC Tool support is provided by the authors and student assistants (with NSF support). Email magictool.help@gmail.com or laheyer@davidson.edu for assistance. You can also email the GCAT listserv for help, as there are many MAGIC Tool users on this list. See http://www.bio.davidson.edu/projects/gcat/GCAT-L.html for more information about the GCAT listserv.

Release Information

The following features were added in MAGIC Tool 2.1:

- Users can move multiple grids at once with the shift key.
- Users can combine multiple grid files.
- Spot flagging is significantly faster
- In Segmentation, users can visualize MA and RI plots.
- In Segmentation, users can choose whether their automatic flagging criteria should be combined with a Boolean AND (all) or OR (any).
- Raw data for all genes, including blanks/empties, is now printed in the raw file, if the user chooses to create a raw file.
- In Explore, users can create box plots of expression files and groups to see the five-number summary (minimum, lower quartile, median, upper quartile, and maximum).
- In Explore, the user can now find genes greater than or less than a maximum, minimum, or average absolute value.
- Loading of projects is significantly faster.

Installing MAGIC Tool

MAGIC Tool is distributed freely by Davidson College under the GNU public license. New versions of MAGIC Tool can be downloaded from the MAGIC Tool web page: http://www.bio.davidson.edu/MAGIC

Beginning with version 1.5, the MAGIC Tool download consist of a single zip archive file, called MAGIC_Tool_x-y.zip, which you must decompress to see the MAGIC Tool folder, called MAGIC_Tool_x-y. The contents of the folder are described in the following table.

File Name	Description
Magic_launch.bat	Launcher for Windows (Executable)
MAGIC_launch	Launcher for Mac OS X (Executable)
MagicTool.jar	MAGIC Tool code (called by launcher)
MAGIC Users Guide v2-1.pdf	Users guide (this file)
Installation_guide.pdf	Detailed instructions for installing and running MAGIC Tool
MAGIC Instructor's Guide.pdf	Instructors guide with additional algorithmic details
Plugins	Necessary files for Java TreeView

After you unzip the downloaded file, navigate into the MAGIC_Tool_x-y folder and double click on the appropriate launcher file for your operating system. After a few seconds, the MAGIC Tool "splash screen" logo should appear, and in a few more seconds the program should be open. If the launcher does not properly start the MAGIC Tool program, see the MAGIC installation guide for detailed instructions.

Sample files and source code for MAGIC Tool are also available at the MAGIC Tool Website at http://www.bio.davidson.edu/magic/.

System Requirements

- Windows 2000 or later OR Mac OS X 10.4 or later OR Unix/Linux
- Java JRE 1.5 (5.0) or later
- 512 MB RAM required for full size arrays; 1 GB of RAM recommended.
- Several hundred MB of hard drive space available, depending on the files you work with and what type of analyses you perform

Vocabulary

Addressing is the short process of telling MAGIC Tool the layout of the spots and grids in the tiff file as viewed within MAGIC.

Chip is a synonym for a microarray.

Feature is a synonym for a single spot on a microarray.

Flag is a verb that means you mark a particular spot to indicate its data are not reliable. This may be due to high background in the area, a dust bunny sitting on the spot, etc.

Grid is a compact arrangement of spots with even spacing.

Gridding is the process that MAGIC uses to find the spots on your tiff files

Metagrid is a higher order level of organization. A set of grids are organized into groups called metagrids. For a more complete description, see this web page <www.bio.davidson.edu/projects/GCAT/Griding.html>.

Segmentation is the process of finding the signal and distinguishing it from the background. There are three methods in MAGIC. Fixed circle is the fastest, and recommended for most purposes. Adaptive circle and seeded region growing are also provided.

Tiff files (e.g. file_name.tif) are the raw image data that are produced when a DNA microarray is scanned. One tiff file is produced for each color on each chip.

Getting Started

Overview of Steps

If you start with two tiff files, you will need to perform the following steps in order to produce clusters or explore your data.

1) Start a Project
2) Add files to project (recommended)
3) Load tiff files
4) Load gene list
5) Locate spots (Gridding and Addressing)
6) Distinguish signal from background and generate expression file (Segmentation)
7) Repeat steps 1-6 for all experimental conditions, appending to previous data and forming an expression file with several columns
8) Log-transform ratios
9) Add gene info to expression file (optional)
10) Explore data (recommended)
11) Filter data (recommended)

The following steps can only be performed if you have three or more columns in your expression file:

12) Calculate dissimilarity (e.g. 1 – correlation)
13) Cluster genes

(1) Start a Project

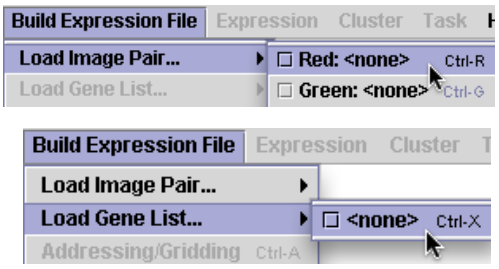
Under the Project menu, create a new Project. You can save this project in a convenient location on your hard drive. We recommend that you NOT use the MAGIC Tool software folder, since you may want to open this project with a newer version of MAGIC Tool in the future. Project files are automatically given a name that ends with the suffix ".gprj" and stored in a folder by the same name, automatically created by MAGIC Tool.

(2) Add Files to Project

We recommend that you copy files into your project, either through the Project menu options, or by dragging the files into the project folder and then selecting "Update Project" under the Project menu. Adding files to your project organizes your files for you into default folders, and simplifies future steps in the analysis.

(3) Load Tiff Files (Control R and Control G)

Under the Build Expression File menu, load the red and green tiff image pairs. Remember that red is a longer wavelength than green, so if your files are identified by the wavelengths, you should still be able to determine which color is which.

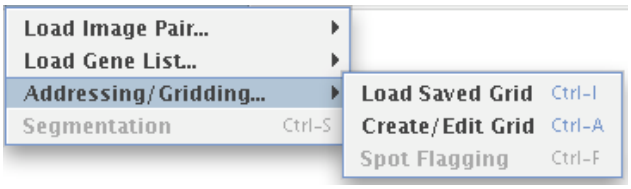


(4) Load Gene List (Control X)

Load the gene list, also under the Build Expression File menu. This should be a text file with suffix of ".txt" and be in MAGIC Tool format. (See full instructions below.)

(5) Locate Spots

Under the Build Expression File, select Addressing/Gridding option.
There are several distinct steps in Addressing and Gridding, which we will walk through one by one in the following paragraphs (a) – (j).

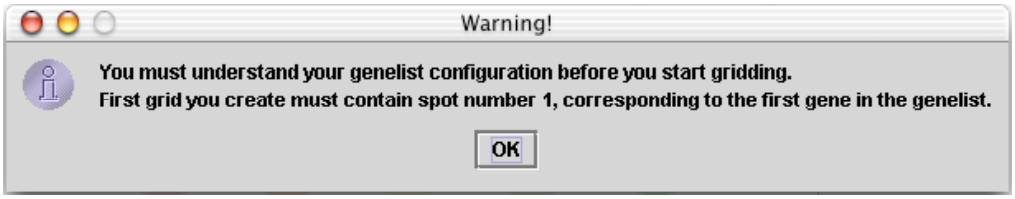


***(a)* Decide whether you want to create a new grid or load a saved grid.**

Unless you have done this before, you will need to create a new grid. If you have a previously-created grid that is appropriate for this image, you can simply load it by choosing "Load Saved Grid" from the Addressing/Gridding submenu, or by pressing Control+W, and proceed directly

to Step 6, segmentation. To create a new grid, choose "Create/Edit Grid" from the Addressing/Gridding submenu, shown above, or press Control+A.

When you create a new grid, you will get a warning window that is normal and intentional. The warning is a reminder that you MUST understand how your spots are arranged on your microarray. For more information about this step, consult http://gcat.davidson.edu/GCAT/workshop2/addressing_MT.html

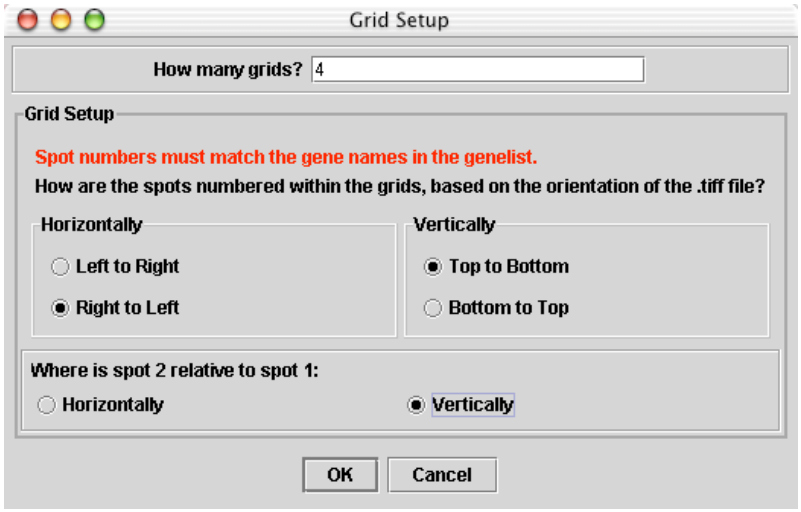


Do not proceed any further if you do not understand the organization of your microarray.

Failure to perform Addressing and Gridding correctly will result in features being incorrectly identified.

You should see two windows. One will show your merged tiff files and the other will permit you to address the tiff file. The smaller (moveable) window will ask you information about how your microarray is organized; this is called addressing.

***(b)* Answer the four questions in the Grid Setup window.**

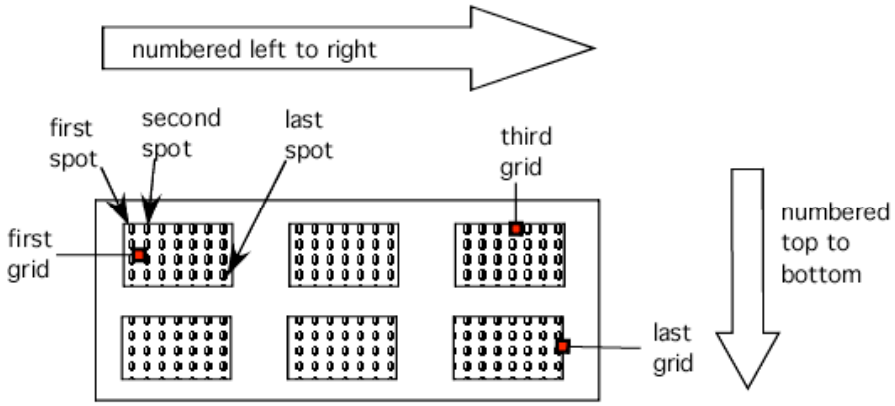


First, enter the total number of grids on the tiff file.

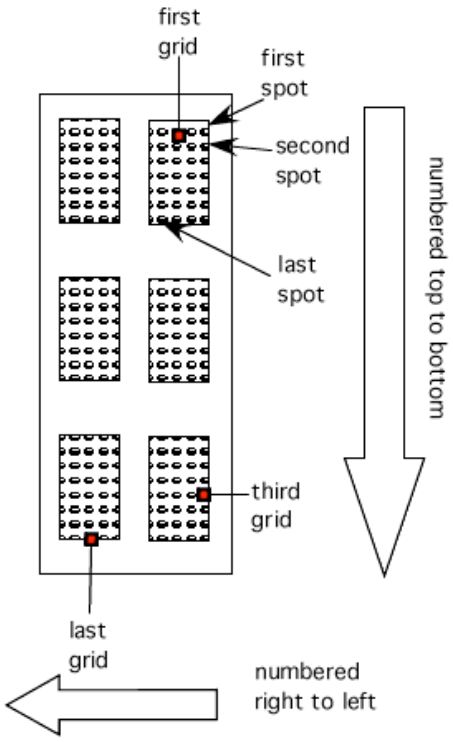
Answering the remaining three questions is the easiest step to make a disastrous mistake. Answer the three questions based on **the way you are seeing your microarray at this time**. Here is an example to illustrate the point. Suppose the image has been rotated 90 degrees clockwise compared to the way you normally think about your chip, but your gene list is not altered to account for the rotation. Then the way you are seeing your tiff file will not match what you think of as your microarray

organization. The following two images show the layout of the microarray before and after rotation.

Before rotation, the spots would be described as being numbered from top to bottom and from left to right, with the second spot horizontal of the first spot (just like you would read a book). These are the default options. However, it is important that you keep track of the spots if the chip is rotated.

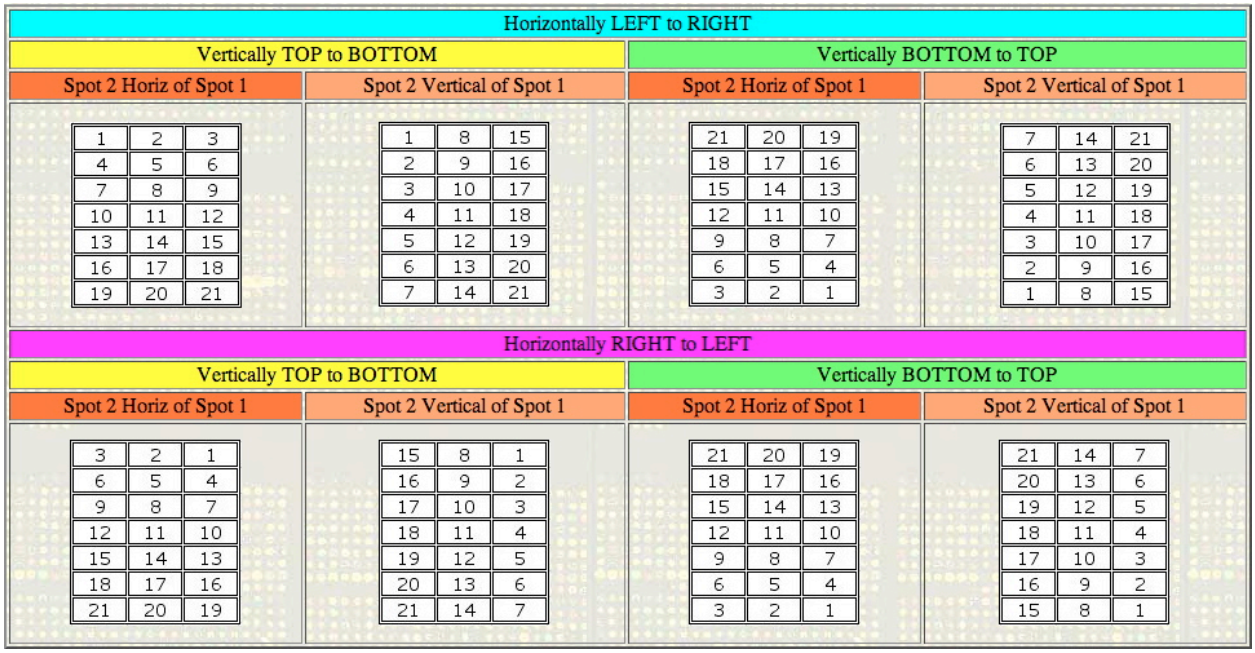


After rotation, the spots are numbered top to bottom, *right to left*, and the second spot is now *vertical* from (below) the first spot. Study the before and after rotation images, to understand how the spots have moved and why the new orientation resulted in the addressing provided in the figure above. Then study all the other options for numbering spots in the table below.



Use the pattern of missing spots and the comments in your gene list to help you become reoriented if necessary. The layout and number of grids is an easy way to orient yourself as well.

If you make a mistake, you can change your answers to these addressing problems by selecting “Grid properties…” under the file menu of the gridding window.



Horizontally LEFT to RIGHT

Vertically TOP to BOTTOM

Spot 2 Horiz of Spot 1

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21

Spot 2 Vertical of Spot 1

1	8	15
2	9	16
3	10	17
4	11	18
5	12	19
6	13	20
7	14	21

Vertically BOTTOM to TOP

Spot 2 Horiz of Spot 1

21	20	19
18	17	16
15	14	13
12	11	10
9	8	7
6	5	4
3	2	1

Spot 2 Vertical of Spot 1

7	14	21
6	13	20
5	12	19
4	11	18
3	10	17
2	9	16
1	8	15

Horizontally RIGHT to LEFT

Vertically TOP to BOTTOM

Spot 2 Horiz of Spot 1

3	2	1
6	5	4
9	8	7
12	11	10
15	14	13
18	17	16
21	20	19

Spot 2 Vertical of Spot 1

15	8	1
16	9	2
17	10	3
18	11	4
19	12	5
20	13	6
21	14	7

Vertically BOTTOM to TOP

Spot 2 Horiz of Spot 1

21	20	19
18	17	16
15	14	13
12	11	10
9	8	7
6	5	4
3	2	1

Spot 2 Vertical of Spot 1

21	14	7
20	13	6
19	12	5
18	11	4
17	10	3
16	9	2
15	8	1

***(c)* Begin gridding.**

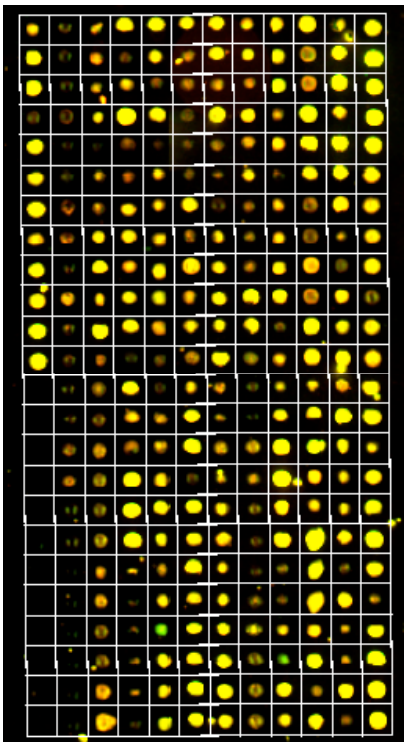
The goal of gridding is to tell MAGIC where the spots within each grid are located. This feature is one of the best innovations in MAGIC Tool. Before you begin, you may want to adjust the contrast to help illuminate faint spots. To do this, slide the indicator that is currently pointing to 100% contrast near the top of this window. Adjusting contrast does NOT affect the raw data; it only allows you to see spots better for this step.

The number one tab should be selected as the default when you begin gridding. The tab numbers on the microarray correspond to the grid numbers. Selecting tab #1 indicates you are working with grid #1 (based on the gene list order). You may begin with a different grid if you wish, but be sure to keep straight where each grid is on the microarray. Again, if you do not follow this procedure of matching grid numbers with tab numbers, you will cause the features to be incorrectly identified. Grid #1 is the grid that contains spot #1, corresponding to gene #1 in the gene list.

***(d)* Center current grid in gridding window.**

Scroll and zoom the image until you can see the first grid as defined by the gene list. To zoom in, click on the "Zoom In" button and then click on the grid where you want the zoom to center. Remember that spots and genes do not change their numbers with image rotation. In the example above where the image is rotated 90 degrees clockwise, the first grid would be the grid in the top right corner.

***(e)* Enter grid location information using "3-click" mouse method.**



a. Click on the button that says "Set Top Left Spot" and then click on the center of the top left spot of the grid.

b. Click on the button that says "Set Top Right Spot" and then click on the center of the top right spot.

c. Click on the button that says "Set Bottom Row" and then click on the center of any spot in the bottom row. Choose a big round spot to make this step easier.

d. Enter the number of rows and columns. This is to be answered based on the way you are currently viewing the tiff file. In this example, there are 24 rows and 12 columns.

e. Click the "Update" button. At this time, you should see all the spots in the first grid surrounded by boxes as shown in the figure.

At any time in the gridding process, you can mouse over a spot and identify its location (x and y coordinates in pixels, row, column and spot number) as well as its identity from the gene list. This information is displayed in the bottom left corner and is especially useful for navigating during segmentation.

***(f)* Adjust the grid to center spots.**

At this time, see if the spots look centered in the boxes. If not, then adjust the position of the boxes either by clicking on the appropriate button and then the correct spot, by manually typing in numbers to adjust the boxes, or by adjusting the grid with the mouse. If you click anywhere inside the grid, you can drag the entire grid to a new location. The grid can be resized from a corner by clicking on one of the gray dots and dragging the mouse. As you drag, the new size and position of the grid will be displayed. Finally, if you click one of the rotation buttons, the entire grid will rotate around its center, allowing you to adjust for slightly tilted grids on your images. If you decide to manually adjust the grid by changing the values in the boxes, note that the position of the mouse is displayed in the bottom left corner of the window so you can determine if the numbers should be bigger or smaller to shift the boxes in the correct direction. Gridding takes a bit of practice, but it is MUCH easier than most other methods for gridding.

(g) **Define the next grid.**

If you only have one grid, skip to step (i). If you have more than one grid, continue. Once the first grid is properly gridded (surrounded with boxes with the spots in the centers), it is time to repeat this process for grid #2. Be sure you know whether grid #2 is left, right, above or below grid #1.

Press and hold the Control (Ctrl) key on the keyboard, then click on the middle of the top left spot of grid #2. The same grid, translated to the location specified by your mouse click, will appear as grid #2, and all the numbers in the boxes on the left will be filled in automatically. If you release the Control key, you can adjust the grid just as you did in step *f*. Repeat this process for all grids.

(h) **Continue gridding.**

Continue step (g) for each remaining grid on the microarray, so that all the grids on the microarray are boxed with the spots in the center of the boxes. At any time, you can change your answers to the four addressing problems by selecting "Grid properties…" under the file menu of the gridding window.

If you need to move multiple grids at once, press and hold the Shift key, then click on each grid that you want to move. As the grids are selected, they will turn blue. Once all the grids you want to move have turned blue, click and drag inside any one of the grids to move all of the grids at once. You can also rotate multiple grids at once by selecting them the same way and clicking the one of the rotation buttons.

You may stop at any time and save your work so far, using the "Save Current Grid As…" under the file menu of the gridding window. Next time you begin Addressing/Gridding, you can simply open this saved grid file.

If you create two different grid files, you can combine them using the "Combine and Load Grid Files" option on the Build Expression File menu. When you choose this menu option, you'll be prompted to pick the first grid file. From this file, MAGIC Tool will take the grid orientation details that you determined in step *(b)* above, in addition to taking all the grids in this file. Once you select the first grid file, you'll then be prompted to select the second grid file, and then the new filename for the combined grid file. MAGIC Tool takes the grids from the first file as the first *n* grids in the new file followed by the grids from the second file as the remainder of the grids. You should make sure that the grids are combined in the right order. Once the grid file has been created, MAGIC Tool will automatically load the combined grid file, and you can edit the grid by choosing "Create/Edit Grid," or continue straight on to Step 6 (Segmentation).

You can also save a snapshot of the combined tiff images at any time before or during the gridding process. You can save the image as tiff, jpg or gif. Tiff format works on all drawing and word processing programs so it is a universal format. Jpeg is good for images such as this that have many shades, like a photograph. Gif is the simplest format but may lose some of the

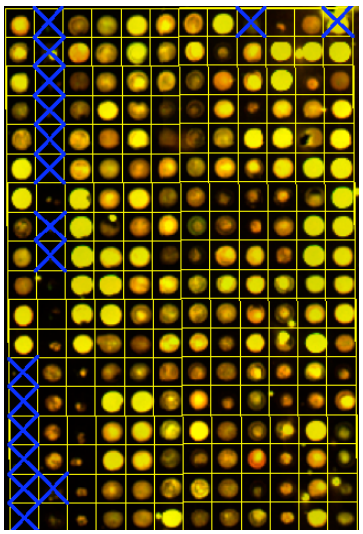
subtlety of your original file. This saved merged image is useful if you want to take a picture of the overall grid and can be used for publishing or teaching.

***(i)* Complete the gridding process.**

When you have finished gridding all your grids, click on the "Done!" button. If you have not already saved your grid, you will be prompted to do so before moving on to the next step. A grid file should be saved in your project folder and automatically given a suffix of ".grid" (so you do not need to type .grid yourself).

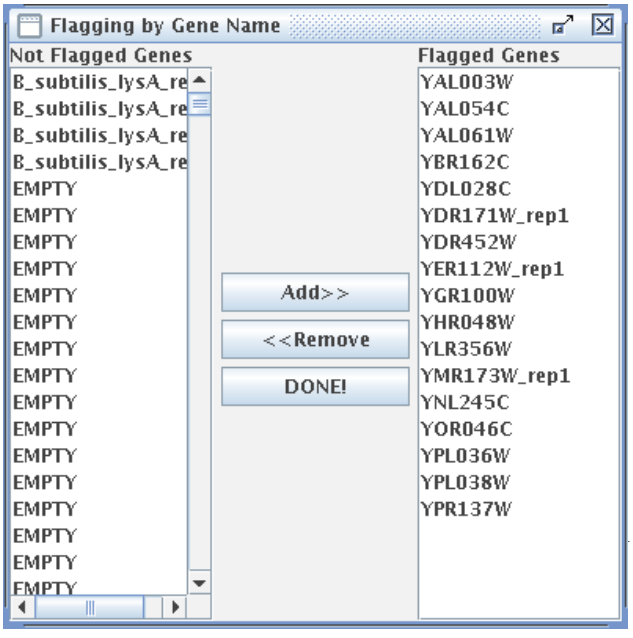
If the number of genes in your gene list and the number of spots you gridded do not match, you will get an error message. You must have exactly one grid square for each line (gene or gene replicate) in the gene list. If not, you probably will make an error identifying the spots later so you are required to fix this problem now. If your gene list and the number of gridded spots match, then you will be informed of the total number of spots and allowed to save the grid file for further use.

(j)* Flag problematic spots *(optional)



If there are spots on your grid that you do not wish to be used in your data analysis, you can choose to exclude the data at this stage, before the creation of the expression file. To do this, choose "Spot Flagging" from the Addressing/Gridding submenu, or press Control+F.

Just as in the gridding window, you can zoom in and out, and fit the image to the screen. Also like the gridding window, when you hover the mouse pointer over a spot, the status bar at the bottom of the window will display information about the gene. If you see a spot that you do not want included in your calculations, click on it. A blue "X" will appear on top of the spot marking it as "flagged" to be ignored by segmentation.

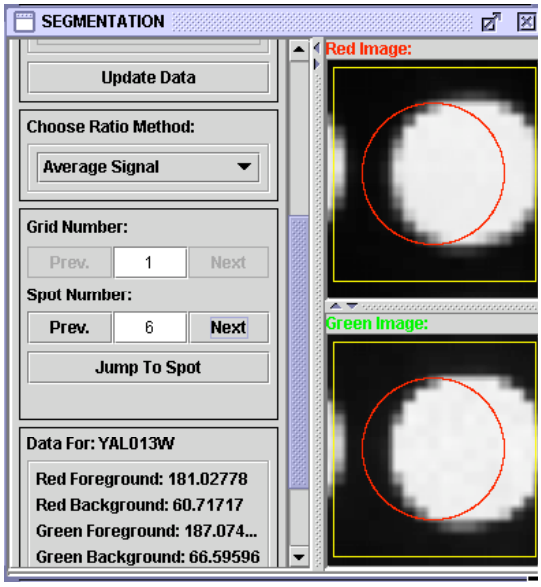


To see what genes have been flagged, or to choose genes to be flagged or not be flagged by their gene name, choose "Flagging by Gene Name" from the Flagging menu. In the dialog that appears, the unflagged genes (the genes that will be used) are on the left, and the flagged genes appear on the right. To flag a gene, click its entry in the list on the left, then click "Add >>." To unflag a gene, click its entry in the list on the right, then click "<< Remove." You can

select multiple items on the list by pressing and holding the Control key, then clicking on each item, or, to select a range of items, click the first, press and hold the Shift key, then click the last. Once you press the Add or Remove button, the changes become visible on the image behind the Flagging by Gene Name window. Note that genes with names "empty," "missing," "none," or "blank" are automatically excluded from the expression file, so they need not be flagged by name. When you're finished flagging by gene name, click "DONE!"

From the main Flagging window, you can also choose to save or load flag files. These files have the extension ".flag" and are stored in the "flags" subfolder of the project folder. The saving process works like the grid file saving described in paragraphs *(h)* and *(i)* above, but you are not automatically prompted to save a flag file. To load a flag file, open the Spot Flagging window, then choose "Load Saved Flags…" from the File menu. From that window, you can choose the flag file to load. Note that the number of grids and number of spots per grid must match the current grid to be able to load a flag file.

(6) Distinguish signal from background and generate expression file (Segmentation; Control S)

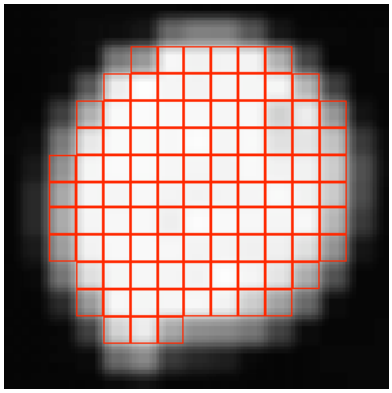


We will break this step into three parts, described in paragraphs (a) – (d).

(a) Select a method for distinguishing signal from background.

Fixed Circle: The most common way is to simply place a circle in the middle of the squares you drew for gridding. This is called *fixed circle*, though you can adjust the radius of this circle as shown in the figure to the right. Note that even if the circle is bigger than the box, only signal inside the box is used for measuring signal.

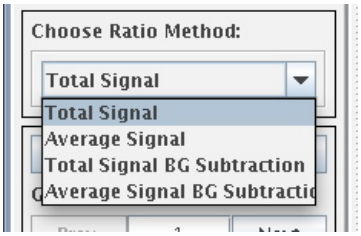
Adaptive Circle: The second method to choose from is the *adaptive circle*. The size and the location of the circle changes, depending of the size on the feature on the microarray. See the instructors guide for more details on this algorithm.



Seeded Region Growing: Seeded region growing is designed to find the signal for each spot based on the distribution of the signal. This method for segmentation looks for the brightest pixel near the center of the grid

square, and then connects all pixels adjacent to this pixel and connects them into one shape. The algorithm simultaneously connects pixels to background and foreground regions, continuing until all pixels are in one of the regions. A user-specified threshold determines which pixels can be used to "seed" the regions. This is the slowest method since each pixel is processed individually. The bigger the threshold, typically the bigger the spot will be defined.

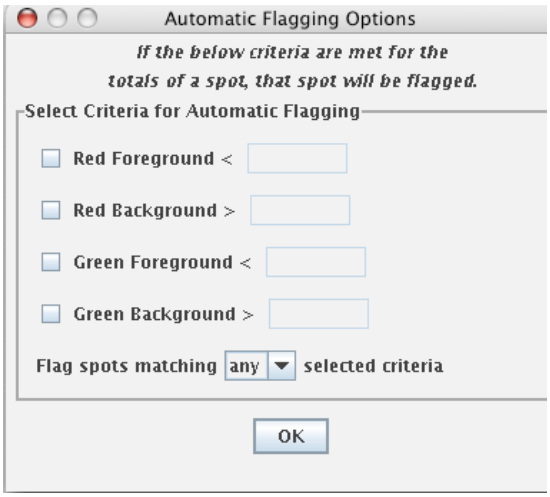
(b) Choose a Ratio Method



The final product of segmentation is a list of gene expression ratios. There are four choices for how to combine the four numerical values in segmentation (red foreground, red background, green foreground, green background) to determine a ratio for each feature on the microarray. Total signal adds the values in all the pixels designated as signal, and divides the red total by the green total. Average signal averages the values of signal pixels. The remaining two options subtract the background (total or average, respectively) before dividing the red by the green to get the ratio. Background subtraction introduces the possibility of a negative value (if background is greater than foreground). MAGIC Tool sets a negative value to 0. If background is greater than or equal to foreground in the green signal, this results in dividing by 0. In this case, MAGIC Tool resets the ratio to 998 or 999 (depending on whether the numerator of red foreground minus red background was also 0, or was greater than 0).

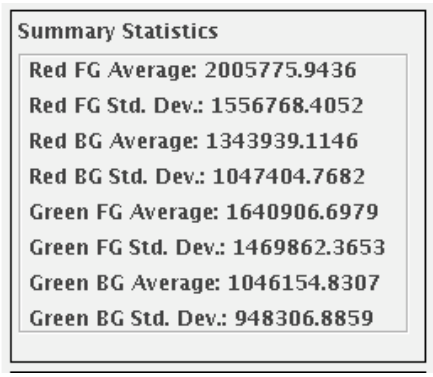
You can navigate around the spots, noting the summary of each spot's data below, to visually verify that the gridding and segmentation were performed adequately. This inspection gives you a chance to note any features you think should not be considered during subsequent data analysis.

(c) Automatically Flag Spots *(optional)*



Once you have chosen your segmentation method and ratio method, you can set criteria such that if any spot fails to meet the criteria, its ratio will not be included in the expression file. To do so, click on the "Automatic Flagging Options" button. Here, you can enter threshold values for the automatic flagging criteria, and choose whether to flag a spot if any (Boolean OR) or all (Boolean AND) of the criteria are met for that spot. When you click OK, you will be prompted whether or not to do calculations to find the flagging status of the spots. In the process, MAGIC Tool also computes the average and standard deviation for each of the four data points used in calculations (even if you leave all the thresholds blank). You can then use this data to refine your

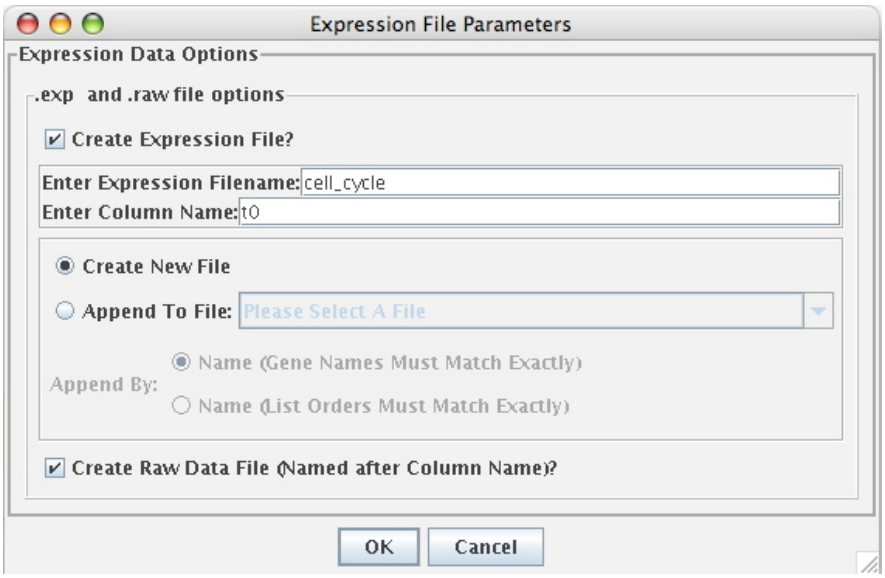
automatic flagging criteria. For example, you might wish to flag genes whose total red foreground or total green foreground is less than two standard deviations below the mean.



To see on a grid what spots have been flagged, open the Spot Flagging window from the Addressing/Gridding submenu. All spots that have been automatically flagged will be marked with an orange "X." These flags can only be changed by adjusting the automatic flagging criteria, but you can add or remove manual flags at this stage as well. If a spot is both manually and automatically flagged, a blue "X" will be shown superimposed on the spot instead of the orange "X." If you unflag manually flagged spot that is also automatically flagged, the "X" will turn orange and the spot will remain flagged.
If you adjust the automatic flagging options, you must recalculate the data to have the revised automatic flags appear on the Spot Flagging display. When you've finished adjusting the options to your satisfaction, continue to generate the expression file.

(d) Generate expression file

Click on "Create Expression File" when you are satisfied with the segmentation process. This will generate an expression file, which was the goal of all the previous steps. An expression file contains the ratios for each spot (red ÷ green), according to the method chosen. MAGIC will ignore certain entries in the gene name column ("blank", "EMPTY", "missing" and "none"; case insensitive). The ratios will be used for all subsequent data analysis. You do not need the tiff files any more.



Unless you have already created an expression file for this microarray, you should check the box next to "Create Expression File?", and name the expression file and the column (e.g. time point, treatment, etc.). You can append this column to an existing file or create a new expression file consisting of this column only. MAGIC Tool will never erase one of your files, so if you append this column to an existing expression file, that file

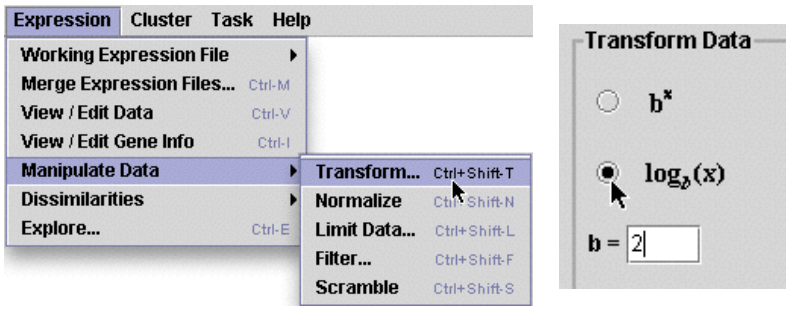
will remain as it was on your computer, and a new file will be created with the current column appended to the right of the columns in the existing file.

In the Expression File Parameters dialog box, you can also choose whether to save the "raw" data that was used to compute the expression ratios. If you check the box next to "Create Raw Data File," a tab-delimited text file will be created that contains 9 columns. The first column is the gene name. The next four columns are the pixel totals for red foreground, red background, green foreground, and green background. The final four columns are the pixel averages for these same four values. The raw data file will have the same name as your column label, with the extension ".raw". Your computer may think this is an image file, but it is just text. You can open the raw file from inside Excel (you may have to force it to look at files of all types for it to open). In future versions of MAGIC Tool, you will be able to use the raw data to filter your expression data, for example when signals are too weak to be reliable. In the meantime, this type of filtering must be done outside of MAGIC Tool.

(7) Repeat steps 1-6 for all experimental conditions

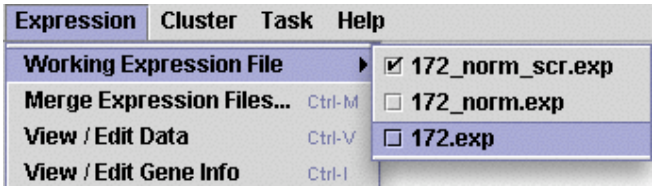
If you have multiple time points or experimental conditions in your study, you should repeat steps 1-6 for each condition before continuing to the data manipulations of step 8. Once you have all data in one file, continue with the remaining steps. If you have only one condition, there will only be one column of data in your file, and you can do steps 8-11.

(8) Manipulate Data



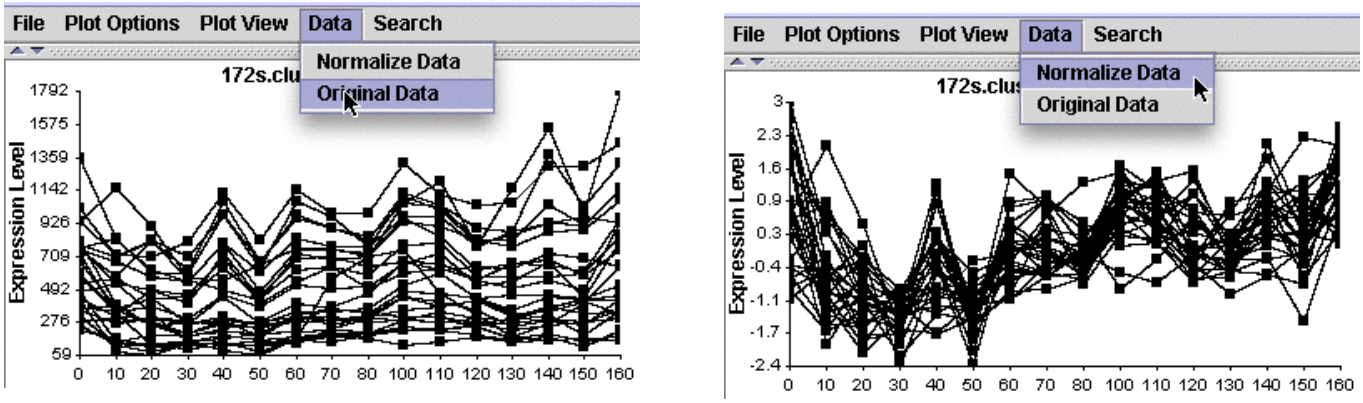
Although this step sounds like a point and click way to conduct scientific fraud, it is actually a beneficial step to consider (see Instructor's Guide). You can: transform or normalize your data; temporarily restrict your data analysis to a subset of experimental conditions (e.g. certain time points, or dye reversals); filter out some features that don't meet certain criteria; or generate a random set of data to use as a comparison.

If you manipulate your data, you will generate a collection of new expression files with names that match the manipulation. MAGIC Tool will never erase your data, so the result of any of these data manipulations is stored in a new file, and the original file still exists as it was before the manipulation. Be sure to verify which expression file you are working with in subsequent steps. It is easy to get confused. The current file is checked on the list under "Working Expression File."



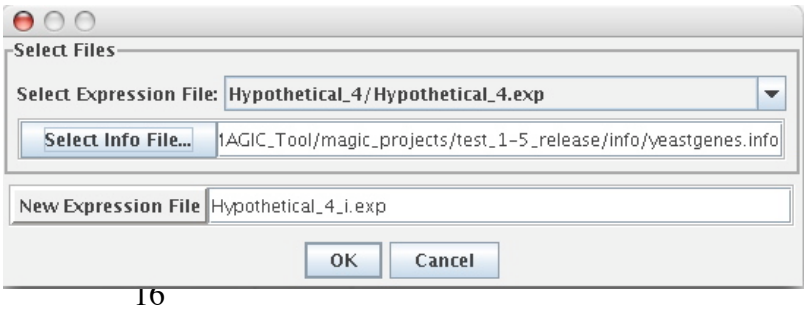
If you are working with ratio data, you should log transform your data. This will convert your ratios into values that are on the same numerical scale so that a gene that is 4 fold induced (+2) has the same numerical value as a gene that is 4 fold repressed (-2 instead of 0.25). Typically, this is done using a $\log_2$ transformation to indicate the number of two fold changes in gene expression (thus 4 fold changes resulted in numerical values of 2).

If you are working with absolute expression values (e.g. Affymetrix data) you may want to normalize your ratios. Normalizing in this case is on a gene-by-gene basis. For each gene, the mean value across the columns is subtracted from each value, resulting in an expression profile with a mean of 0. Then each value is divided by the standard deviation across the columns, resulting in an expression profile with a standard deviation of 1. This type of normalization is especially useful for viewing groups of genes on the same scale, so similarities are more easily seen when absolute expression levels vary greatly from gene to gene. Later, when you plot the various groups or clusters of genes, you can view the data in as normalized or original values, as shown in the following figure.



(9) Add gene info to expression file

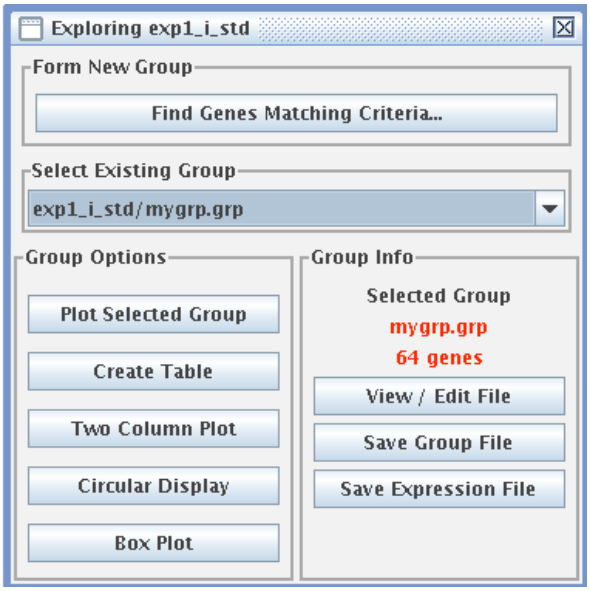
Now is the best time to add gene annotations to your expression file, so the annotations will be visible when you explore your data. Under the Expression menu, choose



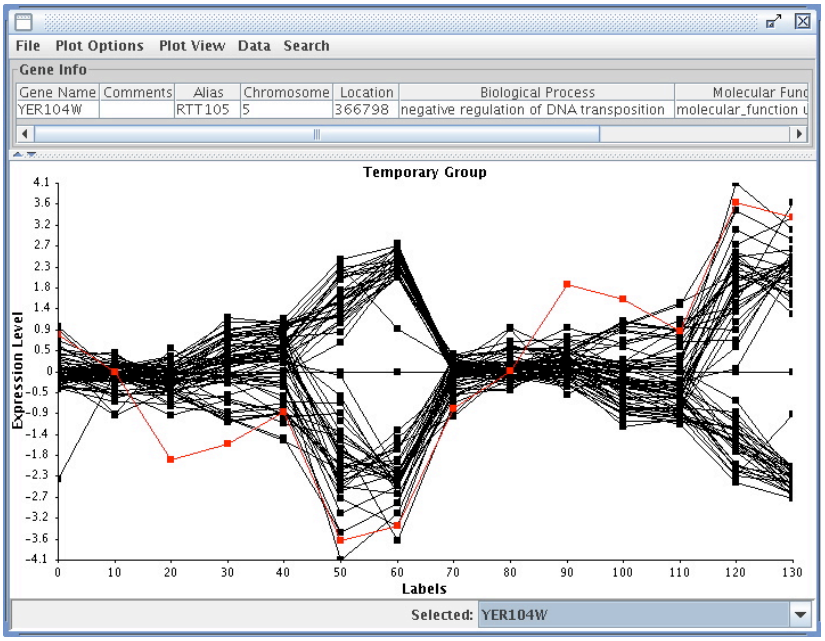
"Import Gene Info…" Select the expression file to which you wish to add annotations, and select the file containing the annotations. A file containing such annotations for yeast is included in the sample files. A similar file can be formed for any organism by creating a tab delimited text file with the appropriate columns (alias, chromosome, location on chromosome, biological process, molecular function, and cellular component).

(10) Explore data

Data exploration is a way to get familiar with your data, and find important functional relationships that may not be apparent from clustering. For example, you can find all genes that were upregulated after a certain time point, or all genes that increased their fold repression four times or greater at any time point. Once you have identified such genes, you can display them in a number of dynamic ways and save these images for publishing or teaching.



If you have not explored the current expression file before and saved group files, the only existing group is the entire expression file. You can create a temporary group by clicking "Find Genes Matching Criteria…" and filling out the form to find the genes and expression patterns you are interested in. If you want a group to be available the next time you explore your data, and the next time you open this project, you need to save the group file (which will automatically be given an extension of ".grp". A group file is just a text file that lists the names of the genes in the group. Any saved groups will then be listed under "Select Existing Group." A group of genes can also be saved as an expression file, which saves all columns of ratios or log-ratios along with the gene names.



Each of the displays on the left hand side of the Exploring window gives

you a different visualization of your data. The "Plot Selected Group" display is shown here, with gene YER104W highlighted. Note that the annotations of this gene can be revealed above the plot of the group. This group was formed by finding all genes whose minimum value was less than -2. Interestingly, the group seems to consist of two distinct sub-groups: genes that are upregulated early and downregulated later, and genes that have the opposite profile.

(11) Filter data

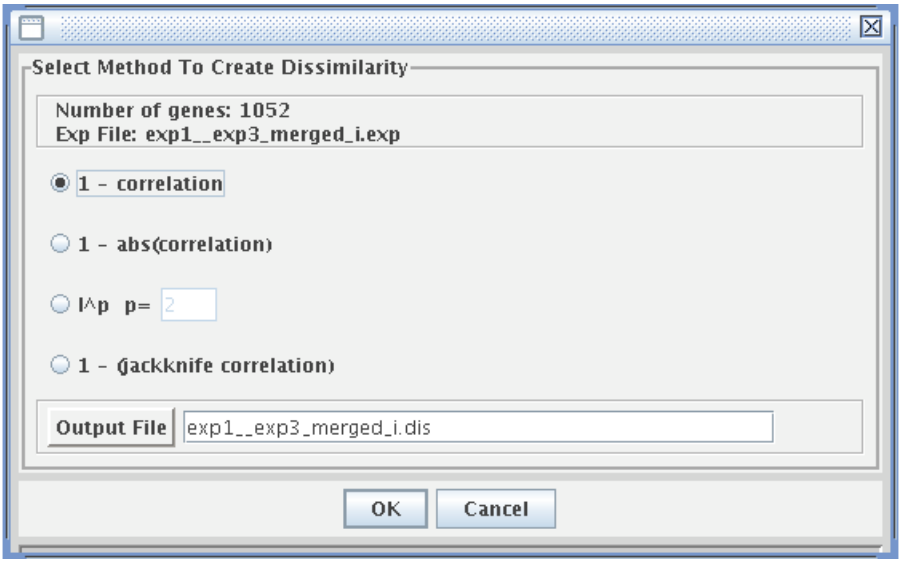
You should filter your data to remove uninteresting genes before proceeding to the next steps. For example, you might keep only those genes whose expression pattern has a sufficiently large standard deviation across the columns (in other words, whose expression is not constant). Or you might remove genes with unreliable ratios (including those involving a division by 0). It is important that your expression file be as small as possible, without losing important information, before beginning the clustering process. You can filter by saving the results of queries in the "Exploring" window as an expression file.

(12) Calculate dissimilarities

To form clusters of similar genes you need a way to compare the expression profiles of different genes. In this step, you will generate a huge table of "dissimilarities," measuring the difference between every pair of gene expression patterns. This step can take a very long time for a large number of genes. Be sure you have filtered your data suitably, and that you know you will learn something from the clustering process before you begin this step.

Under the Expression menu, choose "Dissimilarities" and then "compute". When you do this, a window will appear where you have to choose from three choices. This is another decision that will affect the data analysis.

The most common method is the default, which is 1 – correlation. The second method, 1 – abs(correlation), or 1 – |correlation|, is similar to the 1 – correlation method, but the absolute value of the correlation coefficient is taken before that number is subtracted from 1.



This method can give you a measure of how closely related genes appear to be without regard for if the correlation is positive or negative. The other two methods are described in the Instructor's Guide. When this step is complete, MAGIC Tool generates a dissimilarity file, which you can name in the output file box. The file will automatically be given the suffix ".dis". Click on OK to begin the computation process. The progress is monitored in a popup scale bar (not shown here). You can calculate dissimilarities on any expression file (.exp) but you should use your transformed ratios rather than non-transformed ratios. You can also use transformed and normalized expression files containing absolute expression values. Because correlation and distance calculations have no meaning if you have fewer than three columns, you will not be able to calculate dissimilarities if you have two or fewer columns.

(13) Cluster genes

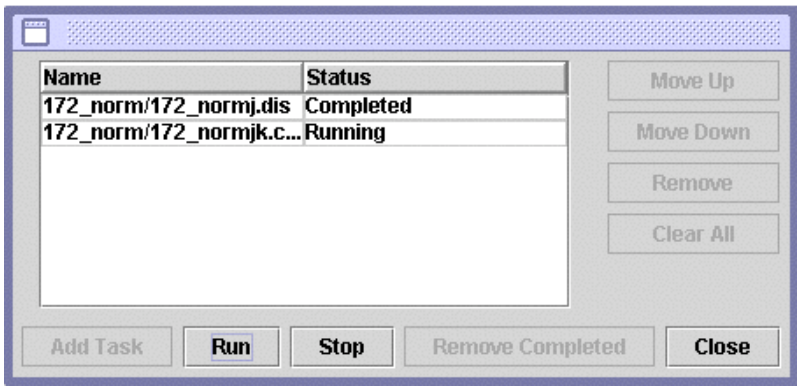
At this point, you can generate a series of clusters using four different methods. Clustering is a very popular process for DNA microarrays, so we will describe this first, but remember that exploration is equally valid and may tell you more about your genes and experimental conditions than clustering can. Exploring your data can be performed any time after segmentation. All you need to explore are expression files (*.exp).

With MAGIC Tool, there are four ways to cluster genes. You can cluster from any dissimilarity file. First you have to calculate the clusters and then you can display them in a variety of ways. The most common way to cluster is called hierarchical clustering, which you can do with MAGIC. However, we prefer Q-T clustering (see Instructor's Guide for details), but Hierarchical Clustering is the only format currently compatible with the data visualization program Java TreeView. You can also cluster by k-means or supervised clustering.

Once you have clustered the genes, you can display the results in several ways. MAGIC allows you to view these clusters in a variety of dynamic displays. Each display can be saved as an image file for publishing or teaching. Display options are addressed in more detail later in this manual.

Automating Tasks

As your datasets get bigger, the time it will take to make all the necessary calculations will increase rapidly. Therefore, MAGIC Tool allows you to establish a list of tasks to be performed in sequence. You can tell MAGIC Tool to begin a series of steps and then walk away from your computer. MAGIC Tool will perform this sequence of tasks while you do other things. For example, you can establish a list of tasks to perform and go home for the night. When you return the next morning, MAGIC will have completed the series of tasks.



Closing Comments

This section was intended as a way to get you launched into the MAGIC Tool way of working with DNA microarrays. MAGIC allows you to compare the consequences of different choices for quantifying, comparing and clustering the same raw dataset. This capacity to compare methods is a powerful way to understand better the assumptions and implications inherent in data analysis as published each week. MAGIC allows you to explore data and data analysis during the early days of DNA microarrays when the research community has not settled upon standards for comparing results. MAGIC was designed to empower the user and make DNA microarrays more approachable for a wider audience. In the following section, every option available in MAGIC Tool will be spelled out so you can utilize the full potential of MAGIC Tool.

Complete List of MAGIC Tool Options

Project Menu

New Project (Control N)
This begins a new project. A project is a way of organizing all related MAGIC Tool work in a folder. The name you give to the project is the name of the folder, and the folder is automatically created by MAGIC Tool. Each project name should be unique and descriptive. Within the folder created by MAGIC Tool will be a file that ends with the suffix ".gprj". All subsequent steps and files will be stored automatically in this project folder, until you start another new project. The .gprj file is a text file that is essentially a table of contents of your project.

Load Project (Control L)
This allows you to reopen a previous project. Navigate to the location of the project on your hard drive, and select the .gprj file within the project folder.

Close Project (Control P)
Allows you to stop a project without quitting MAGIC Tool completely. You can also stop a project by opening a new project and confirming that you wish to close the currently open project.

Add File….
Allows you to add files (e.g. tiff files, gene lists, info files, existing expression files) from other projects to your current project. You will be directed to a window from which you can click your way through the hard drive in search of the files you want to add. Holding down Shift and clicking allows you to select a consecutive range of files. (On Windows, you can hold down the control key and click on multiple files to select them.)

Add Directory…..
Alows you to add all files in the selected folder to your current project.

Remove File….
Lets you remove unwanted files from your current project folder. You can also delete files by writing over the older version (you will be prompted to verify you want to write over the existing file with the same name).

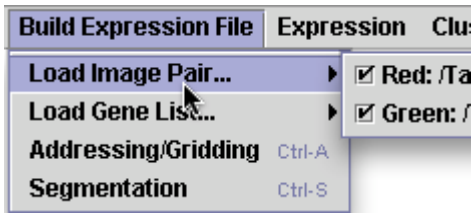
Update Project….
Allows you to drag files into existing folders and then update the currently active project. This allows the user to quickly move tiff, grid, expression, dissimilarity, and cluster files around and then utilize them in different projects.

Project Properties…
Allows you to modify the default properties and configure the behavior of MAGIC Tool. There are three tabs, each containing properties of different types.

> **Data Handling:** Currently the only data handling option is how to handle missing data. You can choose to *remove* or *ignore* any genes in your current project that are missing data. When a DNA microarray is printed, some features will be missing and therefore you cannot collect data for this gene. If you choose to *remove* all genes missing data, then genes missing any data from one or more columns will not be used for calculating dissimilarities. If you choose to *ignore*, you will be prompted for what percentage of possible data (in percent) must be available for a gene to be included in your data analysis. This allows you to work with genes that are missing data from less than that percentage of columns. Genes missing more than the input threshold percentage of columns will be removed.
>
> **Image Saving:** Controls maximum image size saved from MAGIC Tool
>
> **Group Files:** There are two options under this tab. The first, "New Expression Files Carry Group Files When:" controls how group files go along with expression files. This option comes into play whenever you create a new expression file from an existing expression file, for example by log-transforming, adding information, filtering, normalizing or limiting data. Since a group file is simply a list of genes, you may wish groups that you selected based on values in an earlier version of the expression data to be accessible after you do one of the above processes to create a new expression file. The default setting is Always, meaning all group files are copied to the folder containing the new expression file. You can also choose to never copy group files, or to only copy the group files when the expression data itself was not changed (e.g. when adding info to the expression file).

Exit (Control Q)
This quits MAGIC Tool. All completed steps and files will be saved in your project folder. Steps only partially completed will be lost. Open tiff files will not be reopened when the project is opened next.

Build Expression File Menu



Load Image Pair…. (Control R and Control G)

This allows you to browse your hard drive to find the tiff files for the two colors. You can load the two tiff files in either order. If you have added files to the project, or moved files into the project folder and updated the project, all tiff files will be located in the Images folder of the project. Otherwise, you can navigate to the location of the files on your hard drive. Just be sure to match the colors and the files. Remember that red is a longer wavelength than green.

Load Gene List… (Control X)

Reads a file that associates each feature on the microarray with a gene name. MAGIC Tool requires you to have this file, called the gene list, in a particular format. Gene lists in MAGIC Tool format are available for downloading from the GCAT and MAGIC Tool web sites, and are included in the Sample Files, downloadable from the MAGIC Tool Website.

Often, non-MAGIC Tool formatted gene lists have additional information such as which features did not print, alternative names for the gene, etc. You can open your gene list to see what information it contains. If it contains information about the plates and wells for each gene, this is not useful information for MAGIC but was used to help the people who printed the chips to keep track of what they were doing during the manufacturing of the chips.

If you have a gene list that is not in MAGIC Tool format, you can use these instructions, and examples at http://www.bio.davidson.edu/people/macampbell/ACS_MAGIC/genelists.html to create a gene list with the proper format. First, open your gene list from inside Excel. Find the column that contains ORF names such as YBL023c or YAR002W, etc. Copy this ORF column and paste it in the first column (you may have to create a new column to hold this information). Next, remove all header rows, so that the first row in your file is the first gene in the list. Save the modified file as a tab-delimited text file, with a new name that ends with the suffix ".txt". This file is now a valid MAGIC Tool gene list. Although it takes a bit of manual labor to create this MAGIC gene list, it allows the user to quickly adapt to different microarray production styles. Later, you will learn how to import additional information about genes from commonly studied organisms.

Load Saved Grid (Control W)

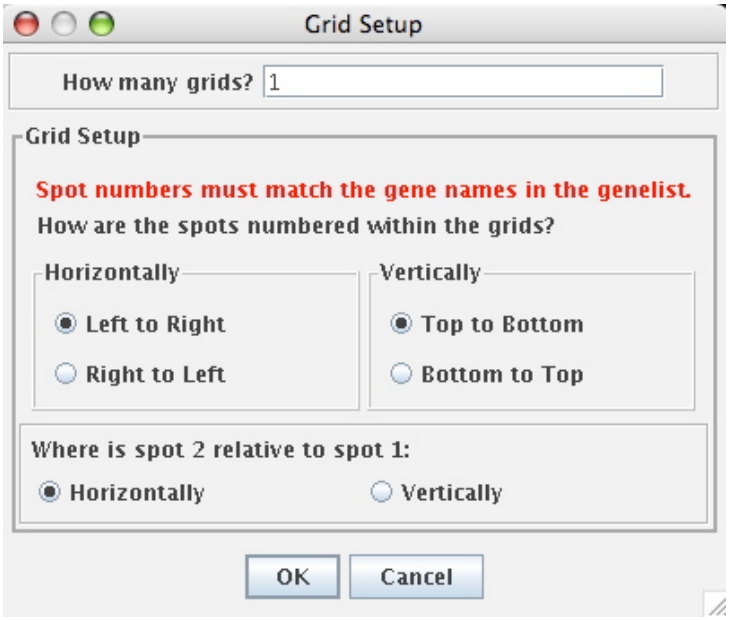
This menu option allows you to load a MAGIC Tool grid file that you've previously saved, which was created through the Create/Edit Grid option below.

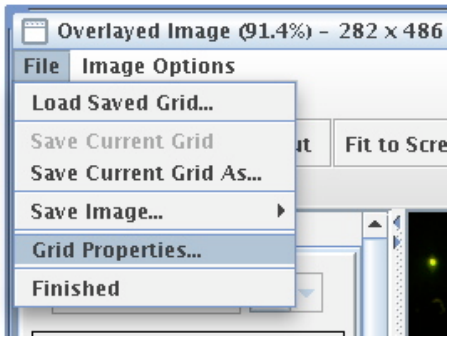
Combine and Load Grid Files

If you create two different grid files, you can combine them using this option. When you choose this menu option, you'll be prompted to pick the first grid file. From this file, MAGIC Tool will take the grid orientation details that you determined in step *(b)* above, in addition to taking all the grids in this file. Once you select the first grid file, you'll then be prompted to select the second grid file, and then the new filename for the combined grid file. MAGIC Tool takes the grids from the first file as the first *n* grids in the new file followed by the grids from the second file as the remainder of the grids. You should make sure that grids are combined in the right order. Once the grid file has been created, MAGIC Tool will automatically load the combined grid file, and you can edit the grid by choosing "Create/Edit Grid," or continue straight on to Segmentation.

Create/Edit Grid (Control A)

When you begin the addressing and gridding process, you should first see a merged image of your red and green tiff files, and where red and green are superimposed, you should see a shade of yellow. Then you will be asked four questions that tell MAGIC Tool how the spots are numbered, shown in the snapshot below. This step, called *Addressing,* is the easiest one to make a mistake on, so be very careful when answering the four questions as they appear in the window. It is vital you understand how your spots are organized on the microarray and in the gene list. All questions should be answered according to the way you see the merged image of your microarray in the viewing window. Are the genes printed in duplicate? If so, are the duplicate spots horizontal or vertical? You will need to know how many grids there are as well as the order of the spots in your gene list compared to the image in MAGIC Tool. The default answers to the Grid Setup questions correspond to the way you would read a book: left to right, top to bottom, with the second spot horizontal of the first one. It cannot be overemphasized how critical this step is. If you get this part wrong, you will not know the correct identity of any of the spots. Once you press OK, you have finished the Addressing step, but you can always choose File, Grid Properties in the Gridding window to get another chance to answer the four questions.





Gridding is much easier. The purpose of gridding is to draw little boxes around each feature so the spots are in the center of the boxes. You may find it helpful to zoom in on the first grid of spots. To zoom in, click on the "Zoom In" button and then click where you want the zoom to center. The number one tab should be selected as the default.

Navigate the image until you can see the first grid as the one you know to be the first grid in the original layout of your microarray. If you want, you can adjust the contrast to help illuminate faint spots. To do this, slide the indicator that is currently pointing to 100% contrast near the top of this window. If the maximum value of the slider is still not enough contrast, you can adjust further by typing the percentage contrast you want in the box next to the slider. Adjusting contrast does NOT affect the raw data; it only allows you to see spots better for this step.

To grid, you simply click on three spots. First, click on the button that says "Set Top Left Spot" and then click on the center of the top left spot. Second, click on the button that says "Set Top Right Spot" and then click on the center of the top right spot. Third, click on the button that says "Set Bottom Row" and then click on the center of any spot in the bottom row. Choose a good spot to make this step easier. Enter the information for the number of rows and columns. Rows and columns are defined based on the way you are currently viewing the tiff file. To finish this grid, click on "Update" button. At this time, you should see all the spots in the first grid surrounded by boxes as shown to the right. (You may need to zoom out to see the full grid.)

At this time, see if the spots look centered in the boxes. If not, then adjust the position of the boxes either by clicking on the appropriate button and then the correct spot, by manually typing

in numbers to adjust the boxes, or by adjusting the grid with the mouse. If you click anywhere inside the grid, you can drag the entire grid to a new location. The grid can be resized from a corner by clicking on one of the gray dots and dragging the mouse. As you drag, the new size and position of the grid will be displayed. Finally, if you click one of the rotation buttons, the entire grid will rotate around its center, allowing you to adjust for slightly tilted grids on your images. If you decide to manually tune the grid by changing the values in the boxes, note that the position of the mouse is displayed in the bottom left corner of the window so you can determine if the numbers should be bigger or smaller to shift the boxes in the correct direction. This step takes a bit of practice, but it is MUCH easier than most other methods for manual gridding, gives you more control and understanding of the process.

Once the first grid is properly gridded, it is time to repeat this process for grid number two. Press and hold the Control (Ctrl) key on the keyboard, then click on the middle of the top left spot of grid #2. The same grid, translated to the location specified by your mouse click, will appear as grid #2, and all the numbers in the boxes on the left will be filled in automatically. If you release the Control key, you can adjust the grid just as you did above. Repeat this process for all grids. Each time you click while holding down the Control key, you will automatically place the next lowest number grid that has not already been defined. Continue this process until all the grids are surrounded with the boxes.

If you need to move multiple grids at once, press and hold the Shift key, then click on each grid that you want to move. As the grids are selected, they will turn blue. Once all the grids you want to move have turned blue, click and drag inside any one of the grids to move all of the grids at once. You can also rotate multiple grids at once by selecting them the same way and clicking the one of the rotation buttons.

You can save your current grid at any time, using File, Save Current Grid (or Save Current Grid As… to save under a different name). Grid files are automatically given a suffix of ".grid". You can close the gridding window without saving, and the current grid will automatically be restored the next time you open the gridding window (without asking the four questions again). If you close the project, however, you must save your grid before you close the project, and choose the option Load Saved Grid when you begin gridding next time. This lets you pick back up where you left off with gridding.

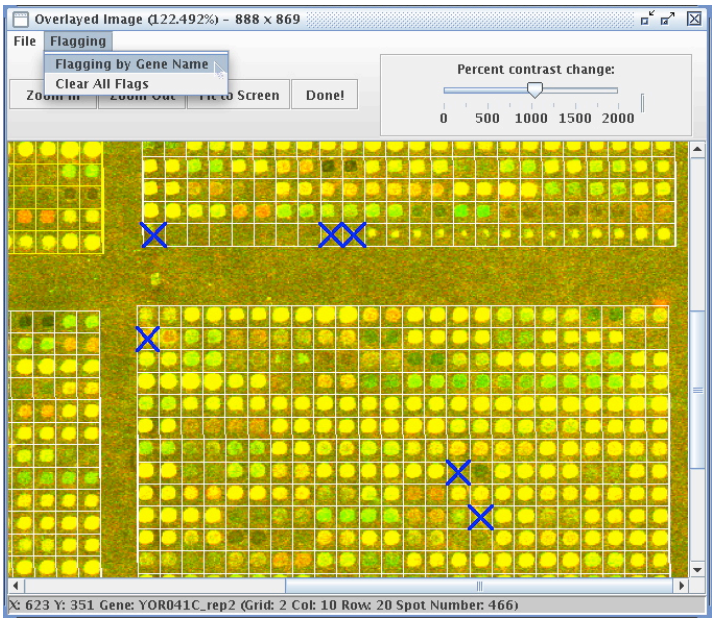
When you have finished gridding all the grids on the microarray, click on the "Done!" button. If you have not already saved your grid, you will be prompted to do so before moving on to the next step. If the number of genes in your gene list and the number of spots you gridded do not match, you will get an error message. You must have exactly one grid square for each line (gene or gene replicate) in the gene list. If not, you probably will make an error identifying the spots later so you are required to fix this problem now. If your gene list and the number of gridded

spots match, then you will be informed of the total number of spots and allowed to save the grid file for further use.

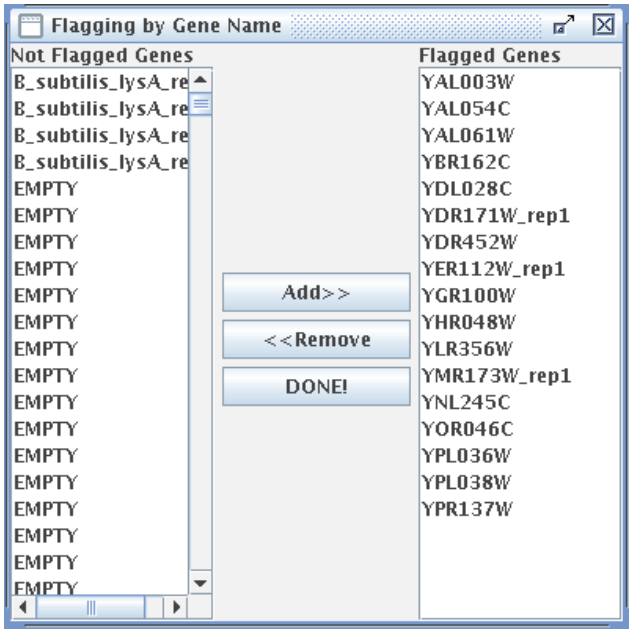
You can take a snapshot of the combined tiff images, before, after, or during the gridding process. You can save an image of whatever is currently showing inside the gridding window, in tiff, jpg or gif format. (Tiff format works on all drawing and word processing programs so it is a universal format. Jpeg is good for images such as this that have many shades, like a photograph. Gif is the simplest format but may lose some of the subtlety of your original file.) This saved merged image is useful if you want to document your gridding process and can be used for publishing or teaching.

Spot Flagging (Control F)

This menu option is used if you want to exclude certain spots from consideration and have their ratios left out of the expression file.



As in the gridding window, you can zoom in and out, and fit the image to the screen. Also like the gridding window, when you hover the mouse pointer over a spot, the status bar at the bottom of the window will display information about the gene. If you see a spot that you do not want included in your calculations, click on it. A blue "X" will appear on top of the spot marking it as "flagged" to be ignored by segmentation.



If you have set automatic flagging options and calculated data for the spots, orange "X"s will appear on top of the automatically flagged spots. These automatic flags can only be altered by changing the automatic flagging options in the Segmentation window.

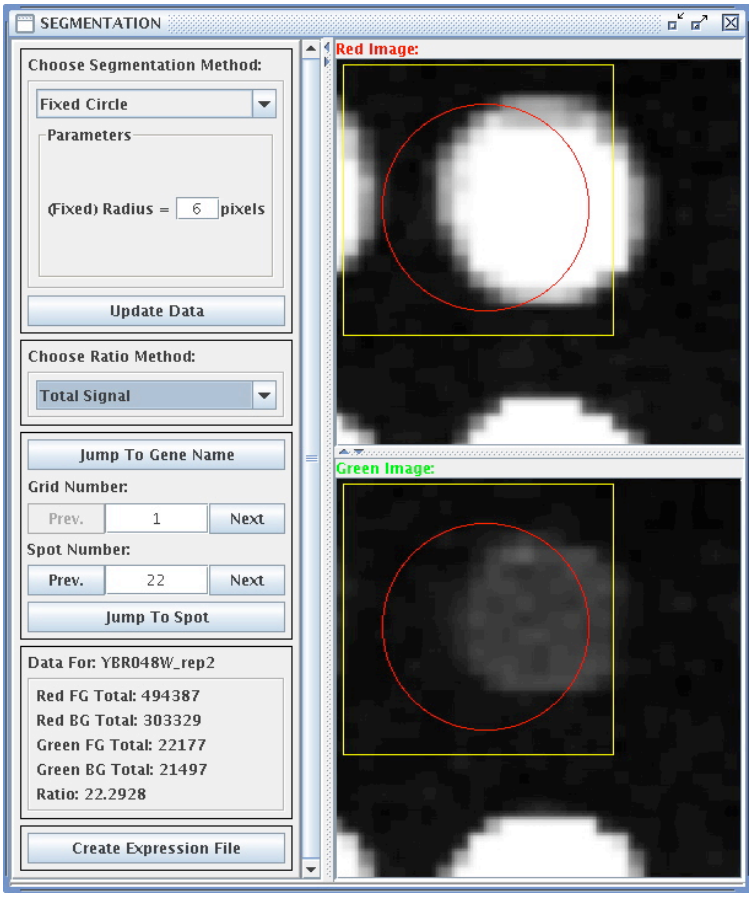
To see what genes have been flagged, or to choose genes to be flagged or not be flagged by their gene name, choose "Flagging by Gene Name" from the Flagging menu. In the dialog

that appears, the unflagged genes (the genes that will be used) are on the left, and the flagged genes appear on the right. To flag a gene, click its entry in the list on the left, then click "Add >>." To unflag a gene, click its entry in the list on the right, then click "<< Remove." You can select multiple items on the list by pressing and holding the Control key, then clicking on each item, or, to select a range of items, click the first, press and hold the Shift key, then click the last. Once you press the Add or Remove button, the changes become visible on the image behind the Flagging by Gene Name window. Genes with names "empty," "missing," "none," or "blank" are automatically excluded from the expression file, so they need not be flagged by name. When you're finished flagging by gene name, click "DONE!"

From the main Flagging window, you can also choose to save or load flag files. These files have the extension ".flag" and are stored in the "flags" subfolder of the project folder. The saving process works like the grid file, but you are not automatically prompted to save a flag file. To load a flag file, open the Spot Flagging window, then choose "Load Saved Flags…" from the File menu. From that window, you can choose the flag file to load. Note that the number of grids and number of spots per grid must match the current grid to be able to load a flag file.

Segmentation (Control S)

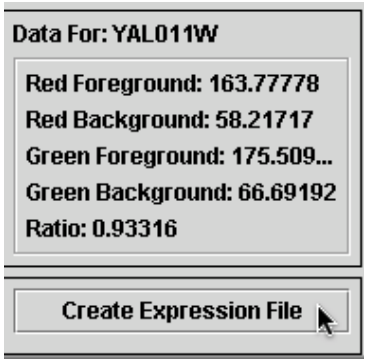
Segmentation is the process of distinguishing signal from background. There are three methods available for this process. During segmentation, you will have the opportunity to view each feature on the entire microarray, one at a time. In this step, the two tiff files are separated again, with the red image on top and the green image on bottom. There are three algorithms available in MAGIC Tool for finding the foreground (signal) and background (noise) in each channel (red and green) separately. In addition, there are four choices for how to combine these four numerical values to determine the ratio.



You might want to experiment with the different algorithms and choices before settling on the best method. By browsing from spot to spot, or jumping to potential problem spots you noticed while gridding, you can see how these choices will affect the final answer. When you are satisfied with your choices, hit the "Create Expression

File" button, and you will be prompted for a file in which MAGIC Tool will save all the ratios, one for each feature on the microarray. When you save the whole list, all values are recomputed, so it does not matter if you have browsed two spots or two hundred. In addition to saving the list of ratios, you will be given the opportunity to save "raw data," i.e. all foreground and background values in the red and green channels.

Fixed Circle

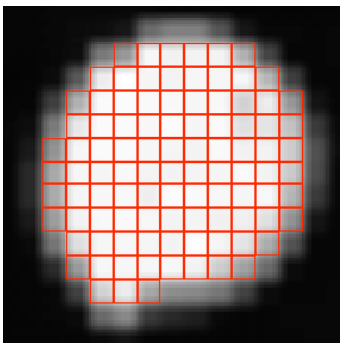


Fixed circle simply places a circle in the middle of the box. All pixels inside the circle (that are also inside the box) will be considered signal and pixels outside the circle (but still inside the box) will be background. You can set the radius of the circle in pixel units. In the above figure, you can see the features are in the box, but they are not centered. The foreground and background values of spots that are off center and spots that are bigger or smaller than the selected fixed radius will not be exactly right. However, the ratio between the red and green values should still be fairly accurate. Fixed circle is the most common method for segmentation, and is the fastest of the three segmentation methods.

Adaptive Circle

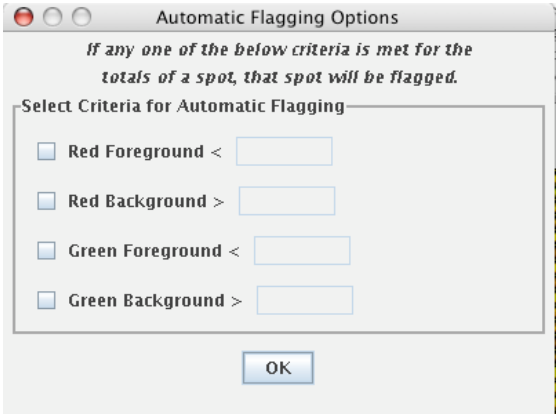
This method changes the center and radius of the circle to fit the size and location of each feature. The algorithm considers all pixels above a user-specified threshold to be "on," and finds the circle with the highest percentage of pixels that are on. The radius can range between a user-specified lower and upper bound; the center can be anywhere inside the grid box. This method is slower than Fixed Circle, but generally covers the actual spot better.

Seeded Region Growing

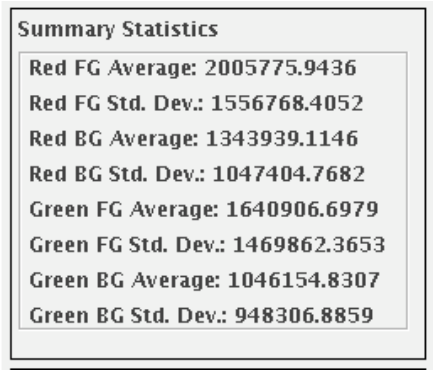


This method for segmentation is designed to find the signal for each spot based on the distribution of the signal. Seeded region growing looks for the brightest pixel and then connects all pixels adjacent to this pixel into one shape. The algorithm simultaneously connects pixels to background and foreground regions, continuing until all pixels are in one of the regions. A user-specified threshold determines which pixels can be used to "seed" the regions. This is the slowest method since each pixel is processed individually.

Regardless which method you choose, you can visually inspect the features to verify the gridding and segmentation were performed adequately. This inspection gives you a chance to flag any features you think should not be considered during subsequent data analysis.



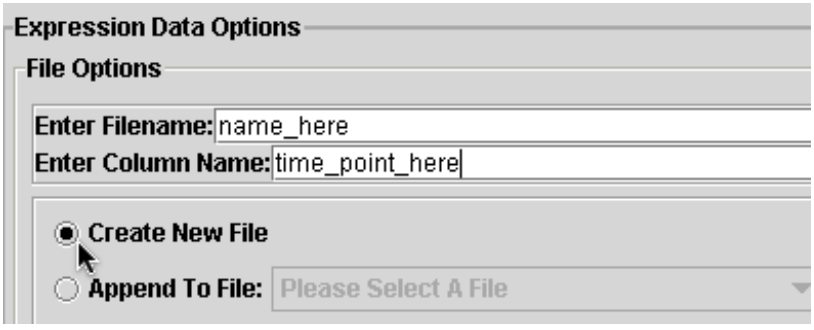
Once you have chosen your segmentation method and ratio method, you can set criteria such that if any spot fails to meet the criteria, its ratio will not be included in the expression file. To do so, click on the "Automatic Flagging Options" button. Here, you can enter threshold values for the automatic flagging criteria and choose whether to flag a spot if any (Boolean OR) or all (Boolean AND) of the criteria are met for that spot. When you click OK (even if you leave all the thresholds blank), you will be prompted whether or not to do calculations to find the flagging status of the spots. In the process, MAGIC Tool also computes the average and standard deviation for each of the four data points used in calculations. You can then use this data to refine your automatic flagging criteria. For example, you might wish to flag genes whose total red foreground or total green foreground is less than two standard deviations below the mean.



To see on a grid what spots have been flagged, open the Spot Flagging window from the Addressing/Gridding submenu. All spots that have been automatically flagged will be marked with an orange "X." These flags can only be changed by adjusting the automatic flagging criteria, but you can add or remove manual flags at this stage as well. If a spot is both manually and automatically flagged, a blue "X" will be shown superimposed on the spot instead of the orange "X." If you unflag manually flagged spot that is also automatically flagged, the "X" will turn orange and the spot will remain flagged. If you adjust the automatic flagging options, you must recalculate the data to have the revised automatic flags appear on the Spot Flagging display.

You can also create MA plots and RI (ratio-intensity) plots. These plots can help you visualize how uniform the printing and hybridization on your chip was, and can also help you determine if you need to perform some normalization outside of MAGIC Tool.
Note: In this context, $M = \log_2 R/G$, $A = \log_2 \sqrt{RG}$.

When you complete segmentation, you will produce an expression file. Click on "Create Expression File" when you are satisfied with the segmentation process. This will generate an expression file, which was the goal of the first half of MAGIC Tool. An expression

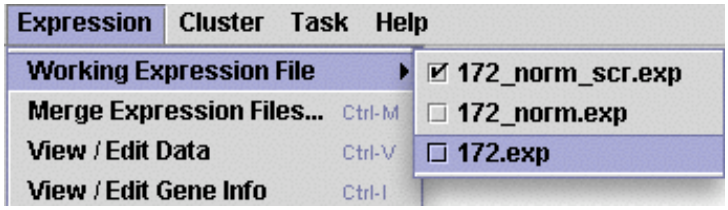


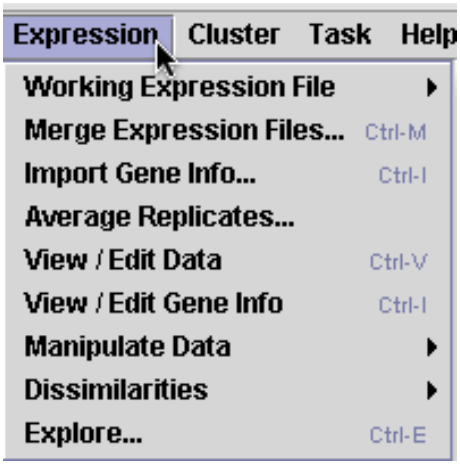
file contains the ratios for each spot (red ÷ green). A ratio of 999 means that a divide-by-zero would have occurred, meaning the green intensity was zero or negative; a ratio of 998 means that a zero-over-zero would have occurred, meaning that both red and green intensities were zero or less than zero. MAGIC will ignore certain entries in the gene name column ("blank", "EMPTY", "missing" and "none"; case insensitive), and will omit any flagged spots from the expression file entirely. This means that, in order to merge files properly, you may need to flag the same genes in all the expression files you wish to merge. The ratios will be used for all subsequent data analysis. You do not need the tiff files any more.

You will need to name the expression file and the column (e.g. time point, treatment, etc.). You can append this to an existing file or create a new one. You can also save raw signal and background intensity levels.

Expression Menu

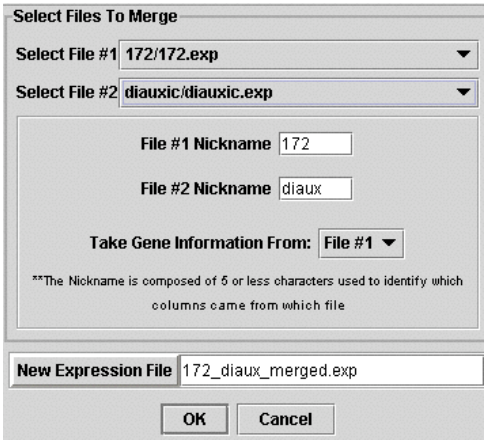
Working Expression File





This option allows you to choose from a range of expression files within a single project. As you can see from the image on the left, you can choose which one is active simply by clicking on it.

Merge Expression Files… (Control M)



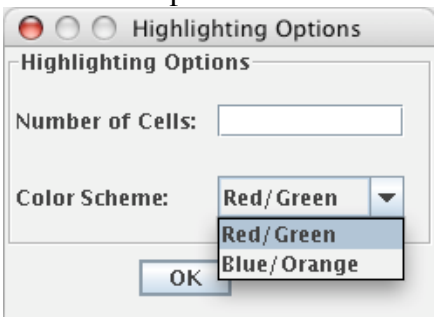
Merging expression files allows you to combine data from multiple chips so you can evaluate time course data, or other related data sets. You merge files one at a time and provide nicknames to assist MAGIC in keeping track of the soon to be combined data. Also, you can select one set of gene annotations as the one that is retained with the merged data set. A new file will be created, so your two original files are not lost.

Import Gene Info… (Control I)

This allows you to compile more complete information about your ORFs. For example, we have created a text file that describes the chromosomal location, the three categories of gene ontology

annotation, and synonym for all yeast genes. This permits you to search by each of these fields to help detect trends and meaningful information. **Average Replicates**
MAGIC Tool treats every spot as a unique feature and does not average for replicate genes automatically. This preserves all your original ratio data. If a set of feature names are identical in the gene list, MAGIC Tool will give each replicate a unique name by appending _rep1, _rep2, etc. After you have created expression files, you may choose to average replicate spots as defined by ORF name. When you average replicates, all features with identical names (disregarding _rep#) then the data will be averaged.

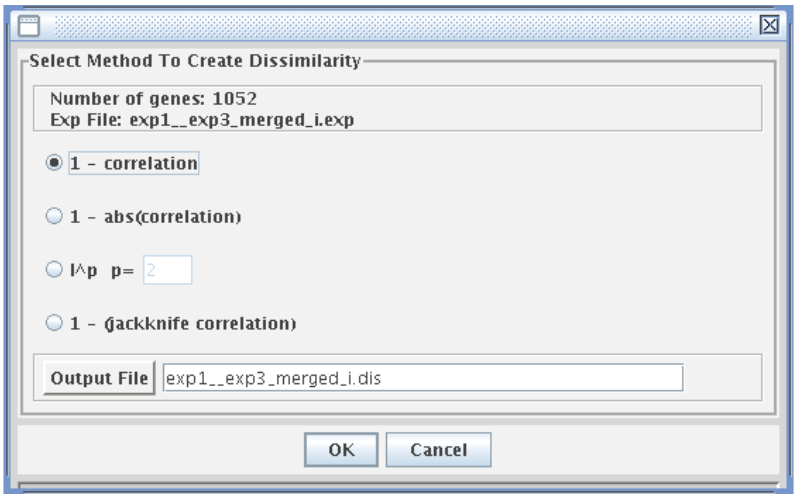
View/Edit Data (Control V)
After an expression file is created or merged, you can view and edit the data. This option should not be used often, but we did want you to have access to the ratio data if you deem it necessary. It is helpful if you want to verify steps or pick up a project after an extended period of time. From this table, you can choose to highlight the top and bottom *n* ratios in each column of your expression file. To do so, choose "Highlight Top and Bottom Ratios" from the Edit menu. In the options dialog that appears, enter how many high/low ratio cells you want to be highlighted, choose the color scheme, and click OK. For example, if you enter "10" in the box and choose red/green as your color scheme, the ten highest cells in each column will be highlighted in red, and the ten lowest cells in each column will be highlighted in green. This feature is useful for checking reproducibility between experiments.



View/Edit Gene Info (Control I)
This option allows you to view and modify the gene annotations. Of course, you can view and edit this file outside MAGIC Tool, but this option provides you an opportunity to do so within MAGIC. Perhaps you will want to perform a search on the gene function. Viewing the list can allow you to select appropriate terms for searching.

Replace Names With Aliases
If you have imported gene info into your active expression file, it likely contains aliases, or common names, for the genes. You can see these aliases by choosing View/Edit Gene Info. For example, YBR167C's alias is POP7. The "Replace Names With Aliases" option allows you to replace the gene names as defined in the gene list with their alias in the info file, creating a new expression file in the process. The old gene name will be the alias in the new expression file. If the alias appears more than once, each appearance will be appended with "_repX" where X is a number from 1 to the number of times that the alias occurs. If a gene does not have an alias, its name will not change.

Dissimilarities (Control D)

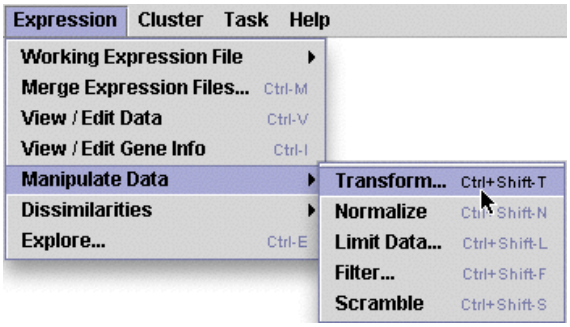


Calculating dissimilarities allows you to compare different genes to one another. When you do this, a window will appear where you have to choose from three options. The most common method is the default 1 – correlation (see Instructor's Guide for a detailed explanation of this and the other methods). When this step is complete, MAGIC generates a dissimilarity file which you can name in the output file box, automatically given the extension ".dis". Click on OK to begin this process. The progress is monitored in a popup scale bar (not shown here). Since correlation and distance calculations make no sense unless there are at least three columns, you will not be allowed to calculate dissimilarities if you have two or fewer columns.

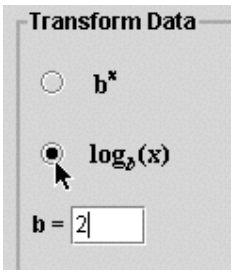
Manipulate Data

Manipulating data is not as bad as it sounds. This option allows you to choose from five options. These options do NOT alter your original data, they simply allow you to process the data further prior to clustering or exploring your data.

Transform (Control Shift T)



A standard process you should perform is transforming your data before performing any analysis (exploring or calculating dissimilarities and clustering). You want to log-transform your ratios so you eliminate any fractions. It is important to get all ratios on the same scale of magnitude. For example, if a gene is repressed 16 fold, the ratio will be 0.0625 while a gene that is induced 16 fold will have a ratio of 16.0. Before analyzing your data, you should log-transform your data. After transformation (typically $\log_2$), the two genes would be altered (-4 vs. +4) with equal magnitude but in opposite directions. See Instructor's Guide for more information. You should explore after transforming, but may or many not want to normalize before exploring (see below). If you want to "un-transform" your transformed data, you can use the exponent function b^x.



Normalize (Control Shift N)

This process takes your (transformed) ratios and corrects for the magnitude of a gene's ratios and the variation among each gene's ratios. Normalization is not appropriate for ratio data, but is useful for absolute expression values. See Instructor's Guide for more details.

Reorder/Delete Columns (Control Shift L)

If you have merged data from many microarrays (e.g. a time course experiment), you may want to study only certain portions of your merged data independently. Limiting data allows you to select column headings and retain these selected data for analysis in a "limited data set". Your original merged file is left unaltered and a new file is created. The new expression file will terminate with the name "x_limted.exp" where x would be the original expression file name.

Filter (Control Shift F)

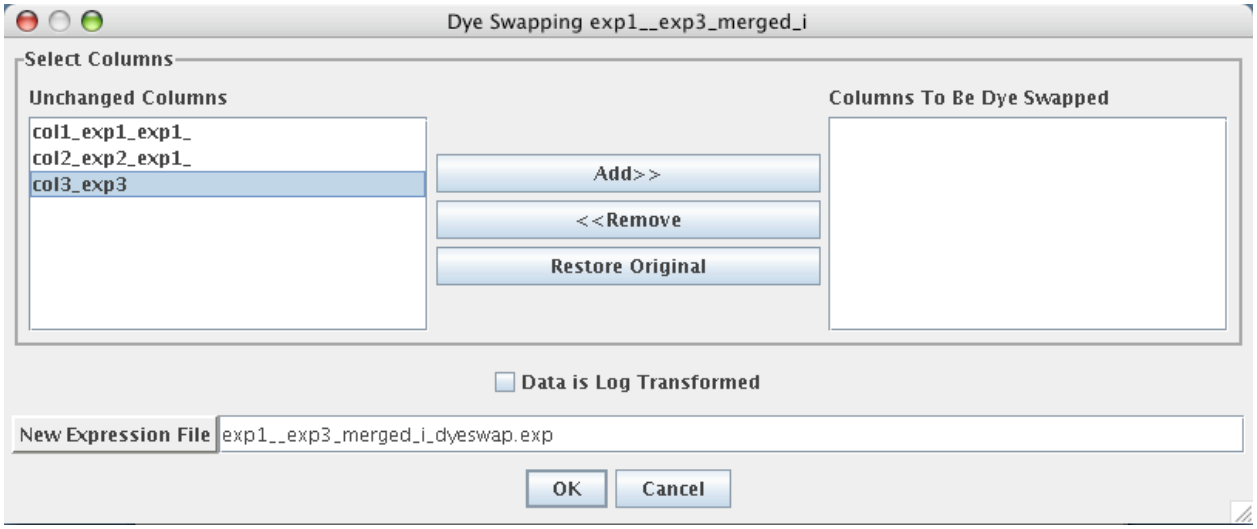
Filtering allows you to remove from further consideration genes that do or do not meet user-defined criteria. Filtering can be performed in this menu, or by saving query results as expression files from the Exploring window (see below).

Scramble

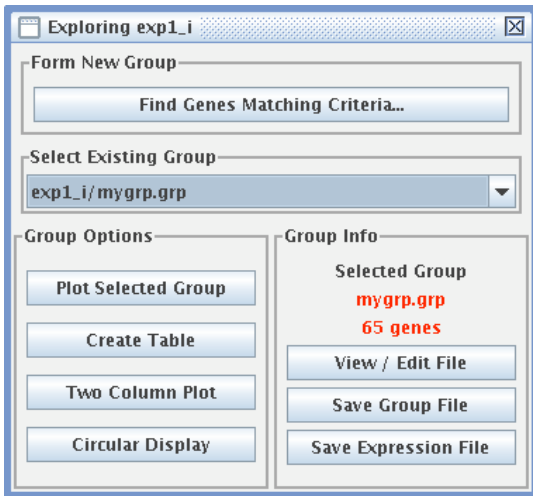
Gives three different methods for creating a gene expression file with the same exact numbers as your current file, but in random order. Randomization can help indicate whether the patterns found through exploration and clustering are real effects of the experimental conditions.

Dye Swap Data Manipulation (Control Shift D)

If you swapped the red and green images while building your expression file, you can swap the ratios after segmentation by choosing "Dye Swap Data Manipulation" from the "Expression" menu. From this window, you can choose columns of the working expression file to be dye swapped. If the "Data is Log Transformed" checkbox is unchecked, the ratios of the selected columns will be reciprocated to achieve the new values. If the "Data is Log Transformed" checkbox is checked, the data will be negated to achieve the new values.

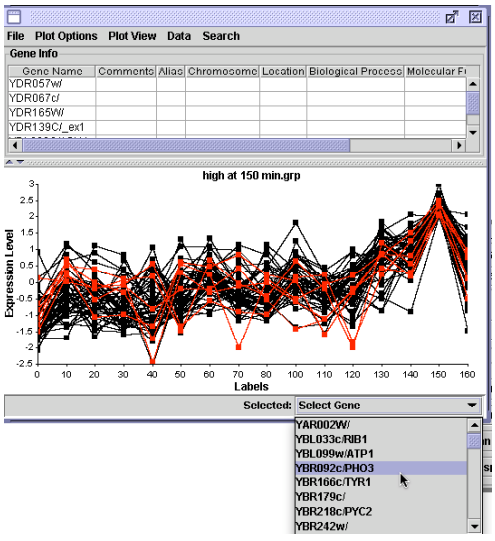


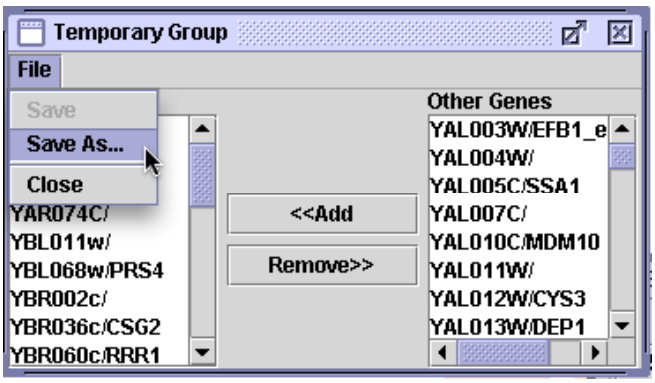
Explore (Control E)



After you have transformed your data, you can explore it in a number of ways. The default group of genes is the entire list in the expression file. You can select a subset of genes via the Form New Group button called "Find Genes Matching Criteria…" You can search for criteria similar to those shown for the filter set on the previous page. When you have identified genes of interest, the window changes as shown to the right in red text. To save this new group of genes, click on the "View/Edit File" button just below the red text, or click the "Save Group File" button just below that. You can also save any open group as a new expression file with only the genes in that group by clicking the "Save Expression File" button. After you save a new expression file, you'll be asked if you want to explore the new file or keep the old one open. If you open the new one, you can use this for progressive query building – in the newly created expression file, form a new group by clicking the "Find Genes Matching Criteria…" button and you can query the new expression file.

A new window will appear that lets you view the list of genes in your newly formed group. You can modify this group if you want, or you can "save as" under the file menu. You can create many subgroups of genes and explore them individually using the "Select Existing Group" pull down menu. Once you have subsets of genes to explore, you can visualize them in a number of ways:

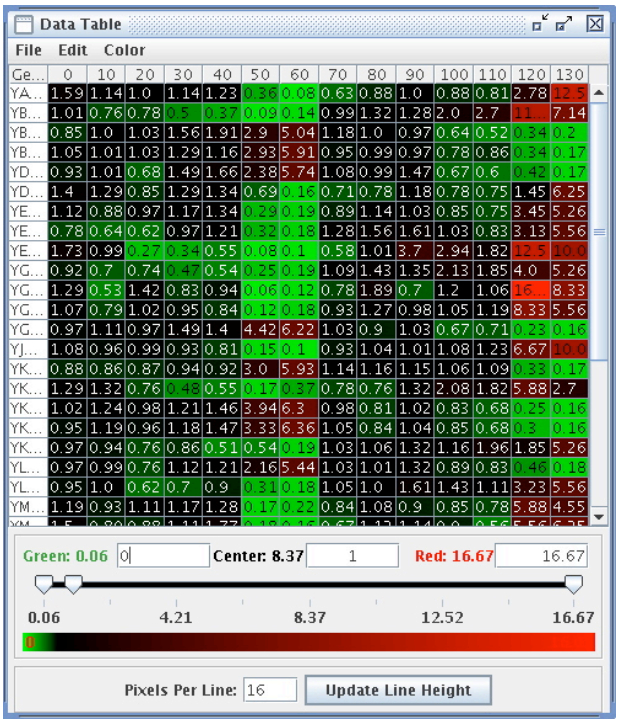




Plot Selected Group

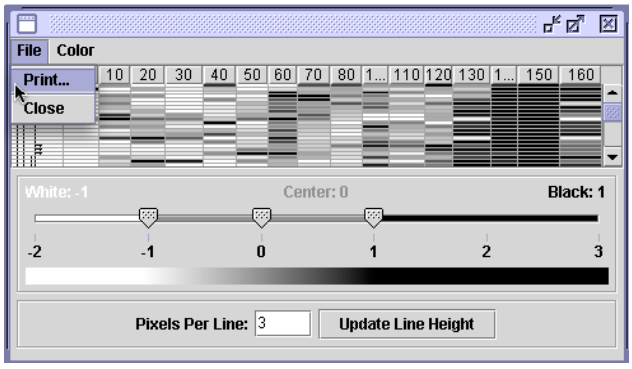
You can have the ratios plotted graphically. You can select one gene using the pull down menu in the bottom right corner. Or, as shown here, you can click on one node at a time and hold down the shift key to select multiple genes (in this case, those with the lowest ratios in the group). These selected genes are listed in the top window (which you can pull down to see) as well as any other information about these genes in your gene list. You can adjust the size of the plot, as well as zoom in on a section. For example, this group of genes was selected by having a ratio of 2 or more at 150 minutes. To untangle the crowded lines, you can zoom in on any region of interest. To do this, hold down the control button then click and drag a box around the crowded area to zoom in. You can unzoom using the Plot View menu at the top of the window.

In addition, you can label the axes, save this as a file, print this plot, normalize the data (if you have not already done so), change the size and shape of the points, and search for certain terms for the genes based on the gene list from which these genes are derived.



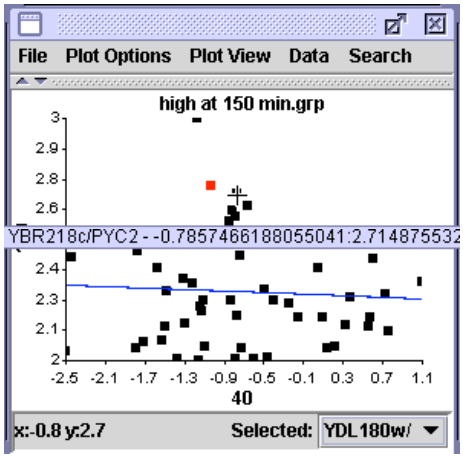
Create Table

This feature is unique to MAGIC Tool and creates a dynamic table. The default is a grayscale table, but you can change this to a red-green scale if you prefer. The most interesting feature of this interactive table is the scale bar and the three sliding tabs. Imagine a gene set that has one gene with a very high ratio (e.g. +16) and one gene with a very low ratio (-16) but with most genes having ratios between +3 and –3. Because of these two extreme genes, the color differences in the remaining genes would be lost. However, if you adjust the tabs, you can compress the color scale on the extreme ends and bring more color variation to the middle of the range of ratios, where most of your genes are located. You can use the mouse to drag the tabs, or enter numerical values in the boxes corresponding to each tab to change the colors. You can choose to view the gene info associated with the genes in the group by choosing the Show Gene Info option from the Edit menu; choose the option again to turn off gene info.



In this view, the gene lines have been reduced from

16 pixels high to 3 pixels high, the color scale changed to grayscale and the range reduced to –1 to +1. This reduction makes all high and low values either white or black, but allows the intermediate values to be on the grayscale.

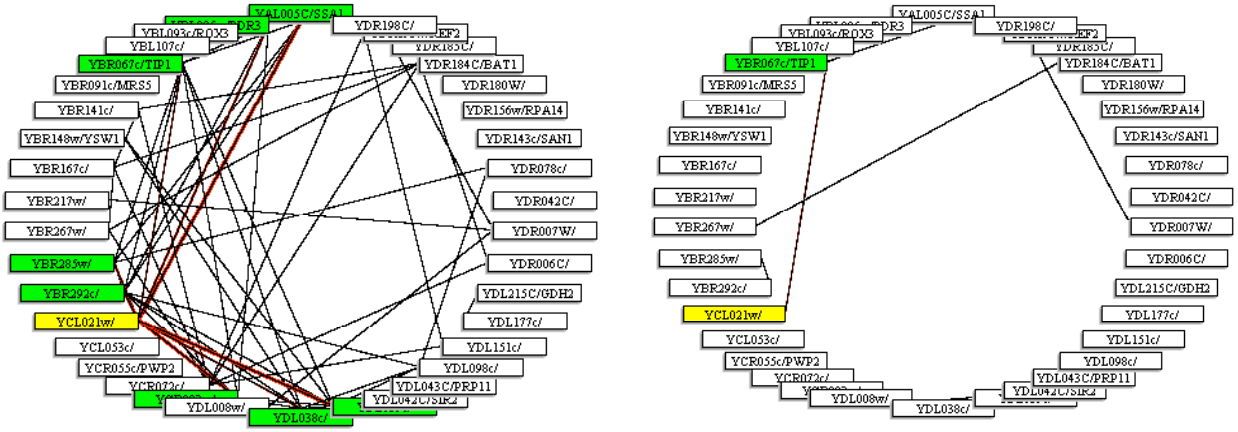


Two Column Plot

This plot allows you to select two columns of data and compare their ratios. As you can see, some comparisons are more similar than others.In this plot, you can select a single gene or many genes (hold down the shift key while clicking). If you mouse over a gene, the display will tell you the two ratios for the two time points. You can also see an approximation in the bottom left corner.

Circular Display

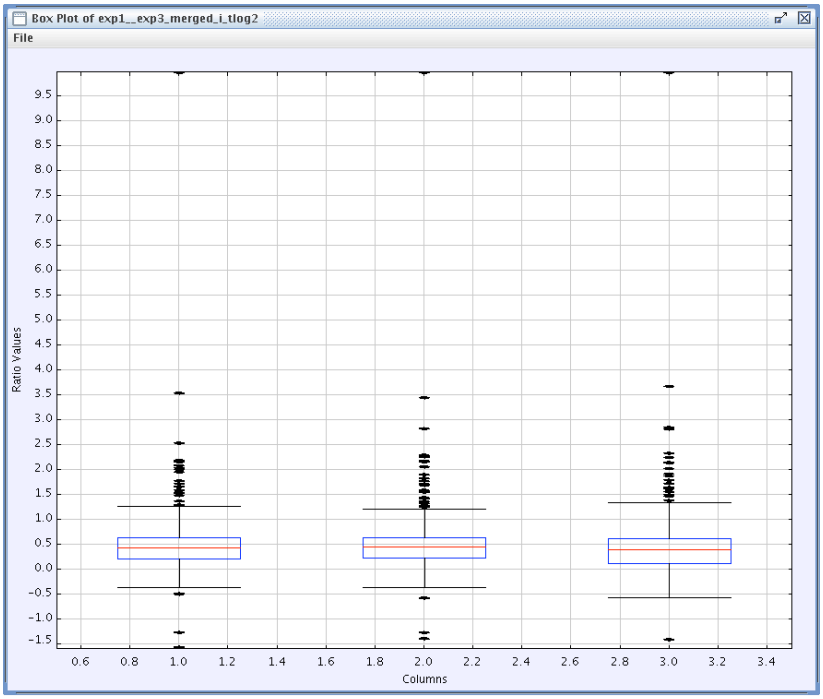
Another unique MAGIC Tool display is the circular one. Let's imagine you have created a group of genes and you want to know how correlation coefficient for these genes, and to which genes the correlation exists. The default setting is correlation coefficient of 0.8 which is shown on the left. Using the display menu, you can change the radius of the circle and the threshold for reporting correlations. Change the threshold to 0.1 (correlation of 0.9) and you see fewer lines connecting the genes (right). In this case, the same gene was clicked on (yellow) and the genes which met the threshold are colored green with the lines colored red.



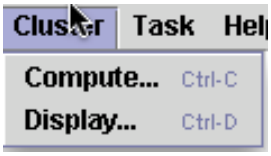
Box Plot

You can also create a standard "box plot," which displays the minimum, lower quartile, median, upper quartile, and maximum in a graphical format. When you choose the "Box Plot" button, a box plot of all the selected genes from all of the columns will appear, each column of data in a separate column of the box plot. The box shows the upper and lower quartile and the red line the median. The horizontal lines at the top and bottom represent the next point past 1.5 times the distance between the 25th and 75th percentiles from the median. The outlying dots are the positions of the outliers.

A box plot will allow you to visualize experiments across columns. This is especially useful if you created biological replicates or replicate chips of the same experiment.

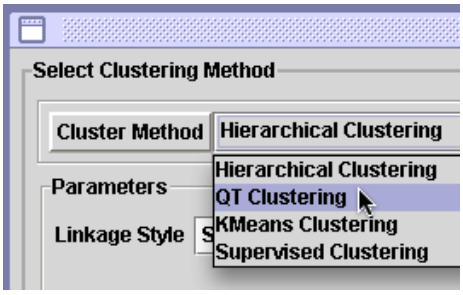


Cluster Menu



Compute… (Control C)

Once you have created dissimilarity file, you may cluster your data. To do this you must computer the cluster using one of four methods. Details for these four methods can be found in the Instructor's Guide.



Hierarchical Clustering

Hierarchical clustering produces a tree-like structure (a *dendrogram*) by connecting genes according to the similarity of their expression data. When a gene joins with another gene or group of genes in the tree, the entire collection of genes is represented as a single pseudo-gene. The similarity between a given gene and the gene (or pseudo-gene) to which it is connected, is indicated by the horizontal length of the branches joining them. At each stage in the algorithm, the two most similar genes or pseudo-genes are joined together. The process continues until all genes have joined the tree.

QT Clustering

QT Cluster takes every gene under consideration and one at a time, builds a temporary cluster for each gene with a user-defined cutoff value for similarity. Whichever gene garnered the most genes in its cluster is used to create permanent cluster and all the genes associated in this cluster are removed from the list of genes for the next round of creating permanent clusters. QT Cluster repeats the process of creating temporary clusters, one gene at a time, and then forms the second permanent cluster using the largest temporary cluster. This process is repeated until all the genes are in clusters, or the remaining genes form clusters smaller than a user-defined size. These remaining genes (called *singletons*) are not presented in the clustering displays unless the user defined 1 as the minimal size for a permanent cluster.

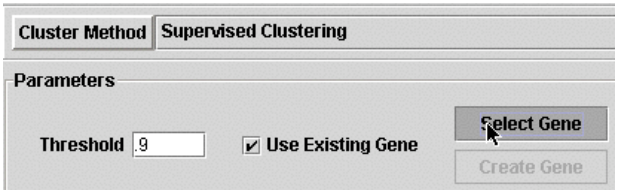
When you use QT Cluster, you should adjust the threshold value. The default of 0.9 means correlation coefficients of +0.1 through +1.0. If you change the threshold setting to 0.2, you will cluster genes only if their correlation coefficients are +0.8 through +1.0. The range of settings for threshold is from 0 (correlation of +1.0) through 1 (correlation of 0, i.e. not similar at all) to 2 (correlation of –1.0; track opposite each other). Therefore, by setting the threshold at 2, you would get every single gene placed in one cluster.

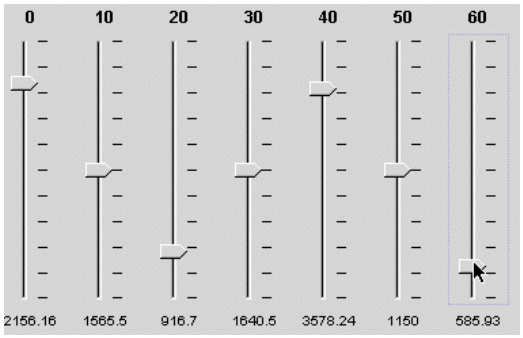
K-Means Clustering

In this method, you determine *a priori* how many clusters there will be (K = the number of clusters) and MAGIC tool will make sure all genes fit into this number of clusters. This is the first step in Self Organized Maps but both methods begin with the investigator determining how many clusters to generate.

Supervised Clustering

This method performs a QT cluster but you can define the threshold and choose one gene around which you want your cluster built. This allows you to focus your research on your favorite gene. On the left, you see that "Use Existing Gene" is selected. Click on the "Select Gene" button and then choose form the genes in your gene list of the currently active expression file.

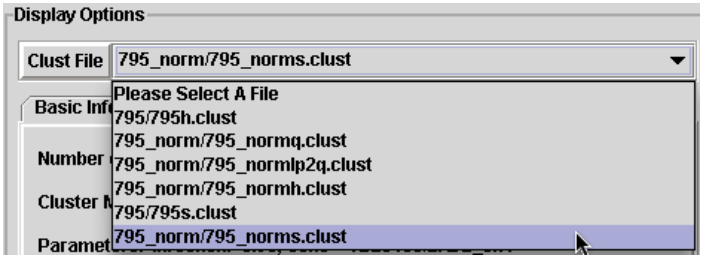




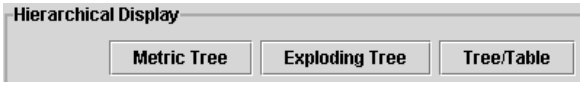
Alternatively, you can deselect the "Use Existing Gene" option and then click on "Create Gene". This produces a window that allows you to manipulate the sliders to create an expression profile for which you want to find genes with similar profiles (based on the threshold you choose). This is a quick way to find complex patters of interest to you.

Display…

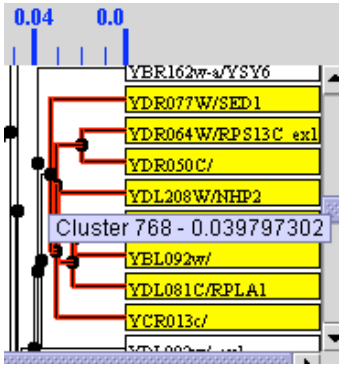
Once you have create a cluster or two, you can display them. First, choose the cluster file you want to display. Each type of cluster has its own display options.



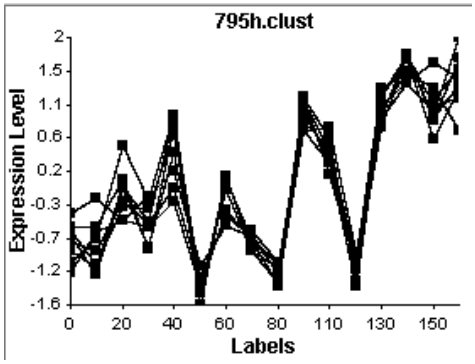
Hierarchical Cluster Display



You have three options for display, each of which has its own options. Metric Tree is unique to hierarchical clustering. It produces a dendrogram with nodes plotted at indicated thresholds. The smaller the threshold number, the higher the correlation coefficient.

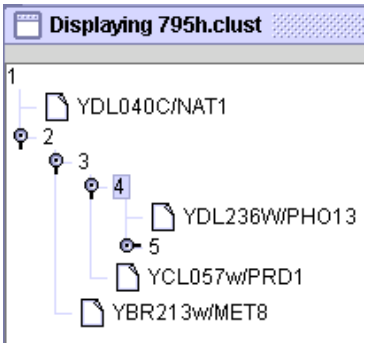


You can click on a branch point and highlight all the genes within this cluster as shown. If you mouse over the branch point, you can see the exact threshold which is 1 minus the correlation coefficient (~0.96). You can plot this cluster and as you would image with this high a correlation coefficient, the normalized data plot as a very tight group.



Exploding Tree is an efficient way to show clusters and gradually expand the contents of each node. In this example, there is one gene and then all other genes are within node number 2. As you click on the nodes, they expand

and if you click a second time, they collapse. You can explode the node completely by highlighting the number and clicking on the explode button, or explode it one at a time by clicking on the node directly. You can also plot any cluster within a node by clicking on the "Plot Node As Group" button.



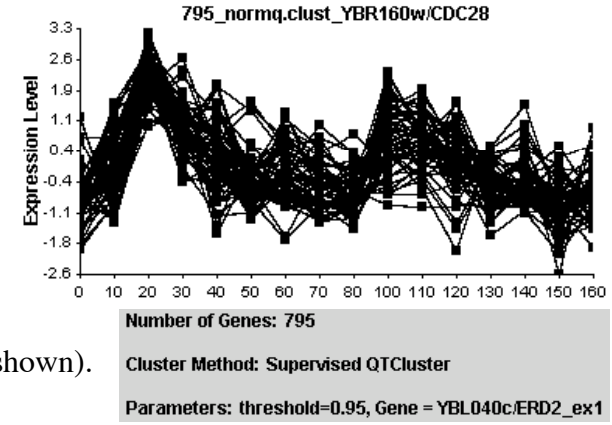
Tree/Table is a way to combine the Table view and the dendrogram. The dendrogram is on the far left and the colored table (the majority of the window) is displayed on the right (view not shown).

QT Cluster Display

QT cluster also allows Exploding tree and Tree/Table, but it has replaced the metric tree with List. List allows you to see the name of the root gene for each cluster. If you click on the root gene, then all the genes within this cluster are displayed. You can plot this cluster as shown here.

Supervised (QT) Cluster Display

Supervised Cluster hast the same display options as regular QT Cluster. However, when you are choosing your display, you should note the box that indicates what threshold was used and which gene was used as the root. In this case, ERD2, the KDEL receptor exon 1 was used as the root for this cluster with a correlation coefficient of 0.95 (plot not shown).



K-means Cluster Display

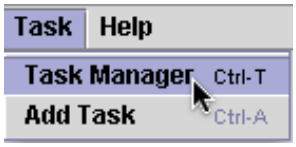
The three displays possible for K-means cluster display are described above.

Create Dendrogram with JTreeView

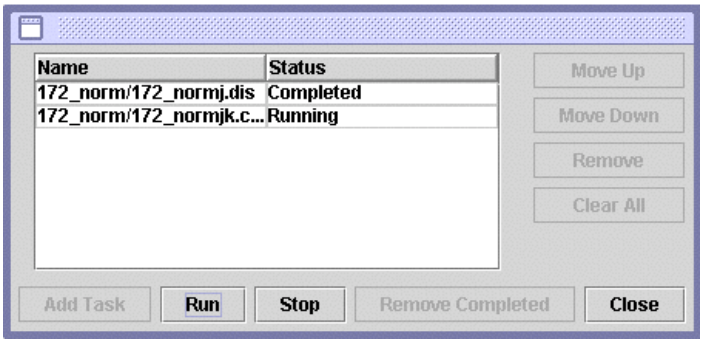
When gene lists get longer than about 5000 genes, displaying clusters becomes slow in MAGIC Tool. One way to handle this is to export a cluster computed by MAGIC Tool for viewing in other software. We export files that are readable by the open source software Java TreeView. Only files created using the hierarchical clustering method currently work with Java TreeView. When you click the Export button in the JTreeView Export Information dialog, the files required to visualize the cluster in JTreeView are created, JTreeView is automatically launched, the files are loaded, and a dendrogram displayed. You can also visualize the data in the files in a karyoscope which can help detect aneuploidy; to do so reopen the file in JTreeView (click File *

Open…, then choose the file you just exported and click Open), then choose “Karyoscope” from the Analysis menu.
For more information about Java TreeView, visit http://jtreeview.sourceforge.net.

Task Menu



As your datasets get bigger, the time it will take to make all the necessary calculations will increase rapidly. Therefore, MAGIC allows you to establish a list of tasks to be performed in sequence. You can tell MAGIC to begin a series of steps and then walk away from your computer. MAGIC will perform this sequence of tasks while you do other things. For example, you can establish a list of tasks to perform and go home for the night. When you return the next morning, MAGIC will have completed the series of tasks. At this time, the only tasks that can be performed are calculating dissimilarities and clusters.



Task Manager (Control Shift M)
The window above is the task manager. It allows you to add or remove a task, change the order of a task as well as various housekeeping chores.

Add Task (Control T)
This option allows you to add a task without going through the task manager.

Help (Control H)
This displays a modified version of this User’s Guide within MAGIC Tool.

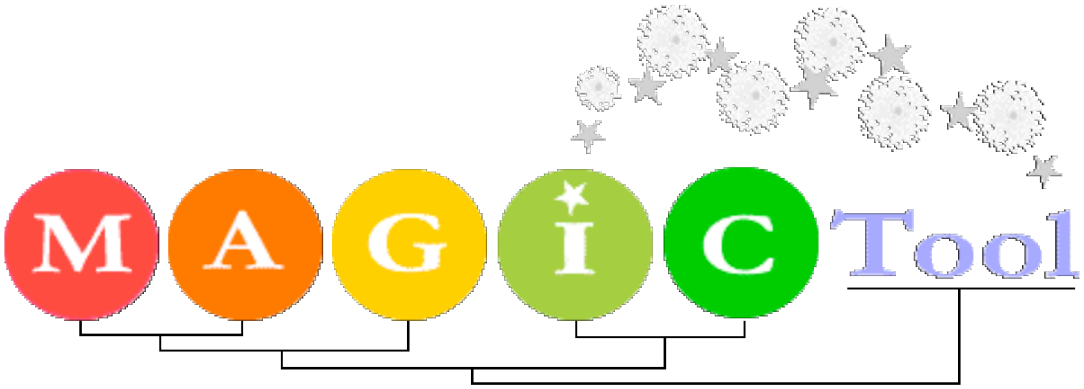
Credits

MAGIC Tool version 1.0 was written in JAVA by Adam Abele, Brian Akin, Danielle Choi, and Parul Karnik, David Moskowitz. Contributors to subsequent versions are Mackenzie Cowell, Gavin Taylor, Bill Hatfield, Nicholas Dovidio, and Michael Gordon. Laurie J. Heyer and A. Malcolm Campbell are advisors to the code-writing team. MAGIC Tool was developed at Davidson College and supported by the National Science Foundation, the Duke Endowment, and Davidson College.

Parts of the code were written by Alok Saldanha (JTreeview), The MathWorks and NIST (JAMA matrix library) and Jari Häkkinen and Nicklas Nordborg (BASE). These sections are licensed under GNU Public License Version 2 or compatible licenses.

We are grateful to the Open Source Physics project, particularly Wolfgang Christian and Mario Belloni, for sharing their knowledge and resources with us.

Exploring Diauxic Shift Microarray Data with



In this lab, we will use the free open-source software program MAGIC Tool to explore the yeast diauxic shift microarray data published by DeRisi et al. in Science (1997). Pat Brown, in whose lab the data were generated, has generously provided both raw and processed data for us to work with. The files you need to do this lab are linked to as you go along. You will also need a copy of the DeRisi et al. 1997 reprint.

You can start this lab in two different places, depending on your interests:

Creating the Gene List

or

Creating the Project

☐

References

Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown, Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science , Vol 278, Issue 5338, 680-686 , 24 October 1997
http://cmgm.stanford.edu/pbrown/explore/

Creating the Gene List

Understanding the Godlist

1. Using a spreadsheet program such as Excel, open the file DeRisiGodList.xls. The file is in tab delimited text format. It is the "godlist" associated with the DeRisi tiff files, describing where each gene is spotted on the microarray.
2. Open this jpeg snapshot of the file 1309_ch1_OD690_green.tif, the results of scanning one of the microarrays in the Cy3 channel.
3. Study the godlist opened in step #1 and the image file opened in step #2 to help you answer the following questions: How many spots are on each microarray? How many grids are on each microarray? How many rows and columns are in each grid? Answer
4. To use the godlist in MAGIC Tool, the spots must be listed systematically, first by grid, and by rows and columns within each grid. Using the Excel sorting function, modify the godlist so that the genes are listed in order of spot number. Note that this results in the grids, rows and columns being ordered sequentially.
5. Once the genes have been sorted in spot order, the MAGIC Tool orientation questions can only be answered in four ways; the other four ways are ruled out by the way the rows and columns are numbered for the consecutive spots. List the four ways that are feasible. Answer
6. To determine the grid order (which is grid 1, 2, 3 and 4), and whether the spots are numbered left to right or right to left horizontally, and whether the spots are numbered top to bottom or bottom to top vertically, we can use the godlist in conjunction with Figure 1 of the paper. For example, find YDL204 in Figure 1, and read its sector, row and column numbers. Use other ORF names (i.e. names that begin with Y) in Figure 1 to determine which is grid 2, and which is grid 4. Check your understanding of the godlist, and how it relates to the microarray image, by looking at the following Jpeg graphic of the array orientation.

Creating the Gene List

1. The gene list for MAGIC Tool will be created from the original godlist, which you should still have open in Excel, with the data sorted by spot number. This systematic ordering of genes is the first criterion for a MAGIC Tool gene list.
2. A second requirement for the MAGIC Tool gene list is that the ORF names must appear in the first column. Using Excel, modify the godlist to meet this second criterion.
3. The final requirement for the MAGIC Tool gene list is that there must not be any column headings. Modify the godlist to meet this third criterion.
4. Save the modified godlist, now a MAGIC Tool gene list, as a tab-delimited text file, calling it derisi_genelist.txt.
5. Continue to Creating the Project.

Generating Expression Ratios

1. In MAGIC Tool, load the Red and Green image files for OD 6.9. Use derisi_genelist.txt as the gene list. When you begin the addressing step, you can either practice creating a new grid, or open the saved grid 1309.grid.
2. During segmentation, create two different expression files:
    a. Using fixed circle with a radius of 3 pixels, and total signal (without background subtraction) create a new file named my3_10, labeling the column 10 (the number of hours that have passed between OD 0.14 and OD 6.9). Show me how.
    b. Using fixed circle with a radius of 5 pixels, and total signal (without background subtraction) create a new file named my5_10, once again labeling the column 10.
3. Repeat steps 1 and 2 for the OD 7.3 array, with the same settings as above. For the OD 7.3 array, the alternative to creating your own grid for addressing is to use 1313.grid. During segmentation, append the 7.3 data to the 6.9 data in files called my5_last2, and my3_last2. In each file, label the current column 12. Show me how.
4. Now we will see how your data compares to the published DeRisi data.
    a. Use the command Merge Expression Files to combine the two expression files that you have just created, my5_last2.exp and my3_last2.exp, calling the result my_last2.exp (override the default name by simply typing over it). Accept the default nicknames for the two files, which will be appended to the column names. The merge will take a few minutes; you will not be able to open any menus until it is done.
    b. Use the command Merge Expression Files to combine the existing expression file derisi_last2.exp with the merged expression file you just created, calling the result all_last2.exp. **Important**: you must select derisi_last2.exp as File #1, because all genes in File #1 need to be in File #2 for the merge to work properly.
    c. Log base 2 transform the expression file.
    d. From the Explore window, perform two-column plots comparing your 3 pixel segmentation to the published DeRisi data for the OD 7.3 array (12 hours into experiment), and your 5 pixel segmentation to the published DeRisi data for that same array. Each plot will take a minute or so to appear, so be patient.
        i. Click on an outlier point in one of the plots, turning the point red and causing the ORF name to appear in the bottom right corner.
        ii. Go to the other plot, and select the same gene from the drop-down menu in the bottom right corner. Is the ratio in the second plot closer to the published data, or even more different?
        iii. Go back to Segmentation in the Build Expression File Menu, which should still contain the OD 7.3 array. Jump to the gene you identified in step (i), and try to explain why the ratio at this particular spot was difficult to determine. Experiment with

different segmentation methods to see what you think the best answer is for the ratio at this spot. Answer

iv. As time permits, explore more outliers in the first set of plots, and/or repeat the analysis with the OD 7.3 array.

v. Explain why it was important to log transform the data before looking for outliers in the two-column plots. Answer

5. Continue to Exploring Expression Ratios

Exploring Expression Ratios

1. Use the command Merge Expression Files to combine the existing expression file derisi_first5.exp and the existing expression file derisi_last2.exp. Be sure to list the files in this order, and change the nicknames for both files to t. Call the merged file derisi.exp. After the merge is complete, examine derisi.exp using View / Edit Data, to be sure the column labels are in order.
2. Add the gene information in yeastgenes.info to derisi.exp, forming derisi_i.exp. Use this merged and annotated file, which is the complete time course published by DeRisi, to answer the remaining questions.
3. How many genes' expression change by at least a factor of 2 in the first two hours? (p. 680) Answer
4. How many genes' expression are greater than 2.0 or less than 0.5 in the time 0 microarray? How does this affect your interpretation of the answer to #3? Answer
5. How many genes' expression increases by a factor of at least 4 sometime during the time course? How many genes' expression diminishes by a factor of at least 4 sometime during the time course? (p. 680) Answer
6. Investigate the change in expression of ribosomal genes by forming a group of ribosomal genes, plotting the group, and highlighting the mitochrondrial genes in the plot. (p. 681) Answer
7. Find genes with the "late induction profile" described on p. 681, and graphed in Fig. 5B, in which levels increased by more than ninefold at the last timepoint, but less than threefold at the preceding timepoint. Compare your results to those in Fig. 5B, and use http://www.yeastgenome.org to help explain any discrepencies. Answer
8. Add the file derisi_lab_ i_tlog2.dis to the project to enable you to answer the remaining questions. This file was generated by transforming the ratios with log base 2 , then computing dissimilarities using 1-correlation. The process of computing dissimilarities takes a few hours, even on a fast computer, so we are skipping this step for this lab. If you do not have the file, you can download it by **right-clicking** here. WARNING, this file is HUGE (72 MB)!
9. Form a supervised cluster with SAM1 (YLR180W) as the seed, and compare your results to Fig 5E,
    a. using 0.2 as the threshold. Answer
    b. using 0.02 as the threshold. Answer
10. If you did not know what patterns to expect or search for, you might want to cluster the genes in to groups with similar patterns first. Use the (unsupervised) QT clust method with a threshold of 0.3 and maximum number of clusters 20. Answer

This concludes the online lab, "Exploring Diauxic Shift Microarray Data with MAGIC Tool."

Exploring Correlation

The accompanying Excel spreadsheet (*correl_explore_scenarios.xls*) illustrates the concept of the Pearson correlation coefficient as a measurement of similarity between gene expression patterns. Each of the four scenarios in the spreadsheet begins with log-transformed gene expression ratios of two genes, as measured in eight different samples. We will refer to the set of eight numbers for a particular gene as a "gene expression pattern," or simply "pattern." The correlation coefficient between the two gene expression patterns is calculated by Excel and displayed in the grey-shaded area to the right of the pattern data.

The first graph for each scenario (on the left hand side) plots the gene expression pattern for each of the two genes. One way to think about the correlation coefficient is as a measure of how well the two patterns "track" each other.

If the two patterns tend to go up and down together, from one sample to the next, then the patterns are highly positively correlated. The patterns in Scenario II have a fairly large positive correlation. The largest possible value for correlation is 1, and this occurs when the change from one sample to the next for one gene, divided by the change from one sample to the next for the other gene, is always the same number. In other words, the two gene expression patterns do not have to have to be the same order of magnitude to be highly correlated. For example, one gene may have values between –1 and 1, while the other gene has values between –100 and 100.

If the two patterns tend to be opposites of one another, i.e. one goes up while the other goes down, as you move from one sample to another, then the patterns are highly negatively correlated. The smallest possible value for correlation is –1.

The second graph for each scenario (on the right hand side) plots the log-transformed gene expression ratio for each sample as a point in the plane. The horizontal axis represents Gene 1, and the vertical axis represents Gene 2. The line of best fit (i.e. regression line) is shown on each graph of this type. If the line of best fit has a negative slope, the two patterns are negatively correlated; if the line has a positive slope, the two patterns are positively correlated. Note that the slope of the line does **not** measure the magnitude of the correlation. Rather, the magnitude of the correlation is determined by how close the points are to the line of best fit. If they are very close, the magnitude is large (near 1 or –1). If they are scattered far from the line, the magnitude is near 0. The patterns in Scenario I have a correlation near 0.

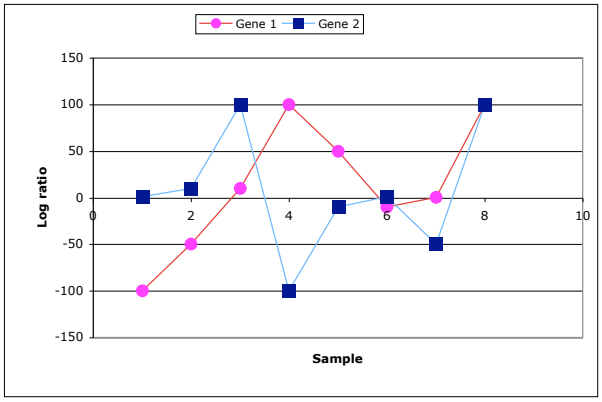
The following exercises guide you though a brief exploration of the correlation coefficient. Answers are on the last page of this document.

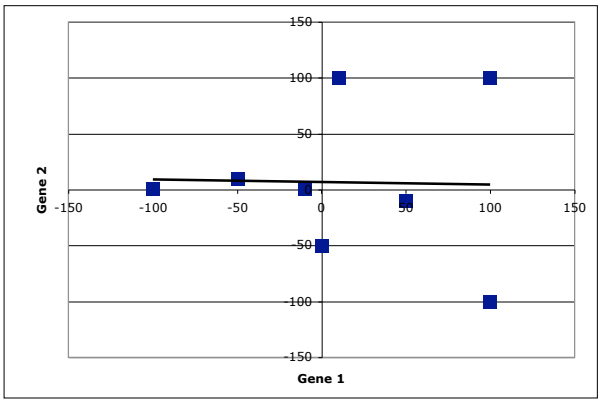
1. In Scenario I, a single number can be changed for Gene 1 that results in dramatic changes in the correlation. Use the two graphs for the scenario to guide your experimentation of the following changes.
    a. Change a single sample for Gene 1 that causes the correlation to jump up to approximately 0.68.
    b. Change Scenario I, Gene 1, Sample 8, from 100 to –150. Note that the correlation jumps down to approximately –0.63. Try to explain this jump by seeing what changes in each of the two graphs.
2. To help answer the following, first notice that in Scenario II, the pattern for Gene 2 is evenly spaced between 10 and 80, changing in increments of 10.
    a. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly 1. You will need to change all but one or two of the values.
    b. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly –1. You will need to change all but one or two of the values.
3. Scenario III illustrates how sensitive the correlation can be to small changes. Here we examine a gene whose log ratio changes substantially across samples and a gene with essentially constant log ratio across samples.
    a. Find a pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a much larger positive correlation.
    b. Return the two samples found in part (a) to their original values of 7, and find a new pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a fairly large negative correlation.
4. Scenario IV shows that correlation is undefined if one of the patterns is constant across samples. As in the previous scenario, changing just one of the values for Gene 2 has a significant effect on the correlation.
    a. Change the value for sample 1 from 4 to 3, and note the effect on correlation.
    b. Change the value for sample 8 from 4 to 3, and note the effect on correlation.
    c. Explain why one of these changes has a greater magnitude effect than the other.
    d. Which single change from 4 to 3 would give the correlation nearest to 0? Why?

SCENARIO I

Sample	Gene 1	Gene 2
1	-100	1
2	-50	10
3	10	100
4	100	-100
5	50	-10
6	-10	1
7	0	-50
8	100	100

Correlation
-0.0229278

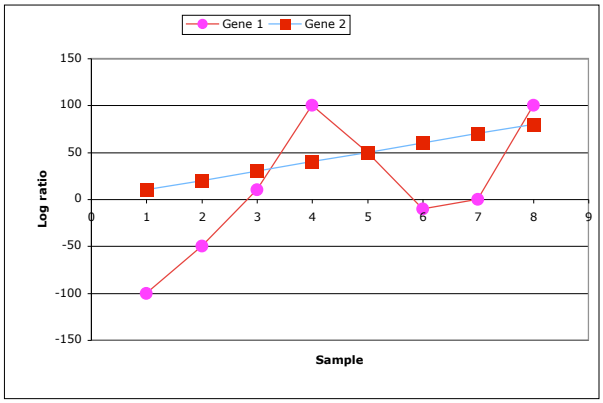


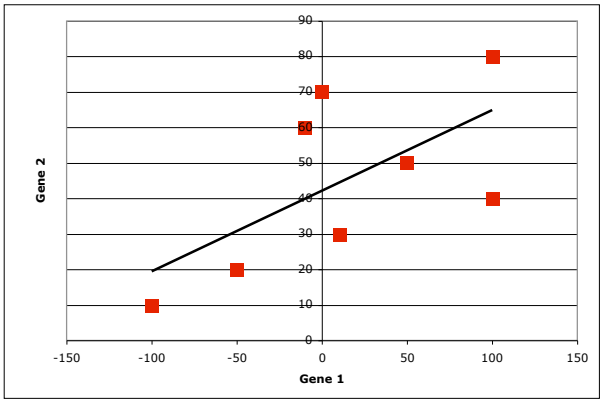


SCENARIO II

Sample	Gene 1	Gene 2
1	-100	10
2	-50	20
3	10	30
4	100	40
5	50	50
6	-10	60
7	0	70
8	100	80

Correlation
0.64483142

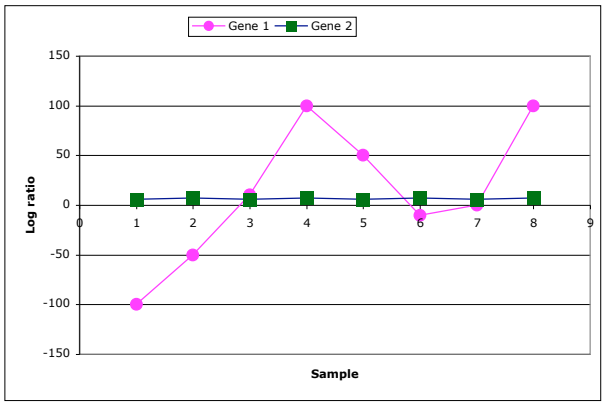


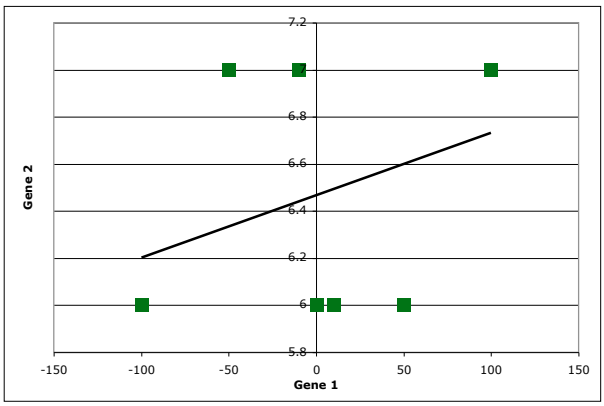


SCENARIO III

Sample	Gene 1	Gene 2
1	-100	6
2	-50	7
3	10	6
4	100	7
5	50	6
6	-10	7
7	0	6
8	100	7

Correlation 0.3453883

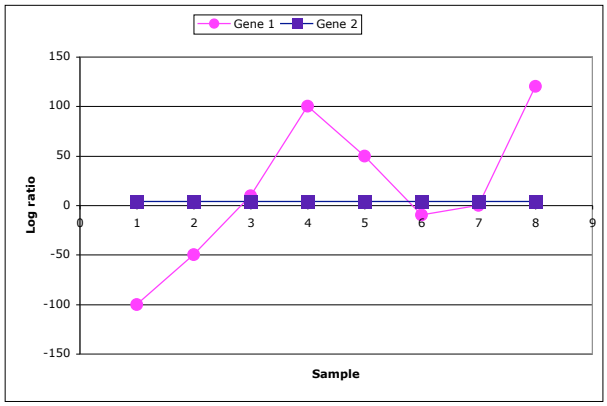


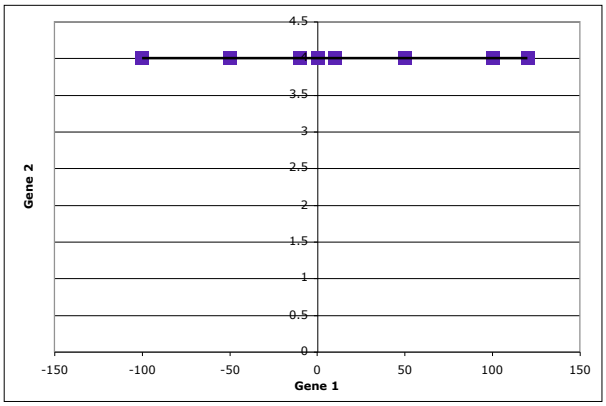


SCENARIO IV

Sample	Gene 1	Gene 2
1	-100	4
2	-50	4
3	10	4
4	100	4
5	50	4
6	-10	4
7	0	4
8	120	4

Correlation #DIV/0!





Exploring Correlation

The accompanying Excel spreadsheet (*correl_explore_scenarios.xls*) illustrates the concept of the Pearson correlation coefficient as a measurement of similarity between gene expression patterns. Each of the four scenarios in the spreadsheet begins with log-transformed gene expression ratios of two genes, as measured in eight different samples. We will refer to the set of eight numbers for a particular gene as a "gene expression pattern," or simply "pattern." The correlation coefficient between the two gene expression patterns is calculated by Excel and displayed in the grey-shaded area to the right of the pattern data.

The first graph for each scenario (on the left hand side) plots the gene expression pattern for each of the two genes. One way to think about the correlation coefficient is as a measure of how well the two patterns "track" each other.

If the two patterns tend to go up and down together, from one sample to the next, then the patterns are highly positively correlated. The patterns in Scenario II have a fairly large positive correlation. The largest possible value for correlation is 1, and this occurs when the change from one sample to the next for one gene, divided by the change from one sample to the next for the other gene, is always the same number. In other words, the two gene expression patterns do not have to have to be the same order of magnitude to be highly correlated. For example, one gene may have values between –1 and 1, while the other gene has values between –100 and 100.

If the two patterns tend to be opposites of one another, i.e. one goes up while the other goes down, as you move from one sample to another, then the patterns are highly negatively correlated. The smallest possible value for correlation is –1.

The second graph for each scenario (on the right hand side) plots the log-transformed gene expression ratio for each sample as a point in the plane. The horizontal axis represents Gene 1, and the vertical axis represents Gene 2. The line of best fit (i.e. regression line) is shown on each graph of this type. If the line of best fit has a negative slope, the two patterns are negatively correlated; if the line has a positive slope, the two patterns are positively correlated. Note that the slope of the line does **not** measure the magnitude of the correlation. Rather, the magnitude of the correlation is determined by how close the points are to the line of best fit. If they are very close, the magnitude is large (near 1 or –1). If they are scattered far from the line, the magnitude is near 0. The patterns in Scenario I have a correlation near 0.

The following exercises guide you though a brief exploration of the correlation coefficient. Answers are on the last page of this document.

1. In Scenario I, a single number can be changed for Gene 1 that results in dramatic changes in the correlation. Use the two graphs for the scenario to guide your experimentation of the following changes.
    a. Change a single sample for Gene 1 that causes the correlation to jump up to approximately 0.68.
    b. Change Scenario I, Gene 1, Sample 8, from 100 to –150. Note that the correlation jumps down to approximately –0.63. Try to explain this jump by seeing what changes in each of the two graphs.
2. To help answer the following, first notice that in Scenario II, the pattern for Gene 2 is evenly spaced between 10 and 80, changing in increments of 10.
    a. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly 1. You will need to change all but one or two of the values.
    b. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly –1. You will need to change all but one or two of the values.
3. Scenario III illustrates how sensitive the correlation can be to small changes. Here we examine a gene whose log ratio changes substantially across samples and a gene with essentially constant log ratio across samples.
    a. Find a pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a much larger positive correlation.
    b. Return the two samples found in part (a) to their original values of 7, and find a new pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a fairly large negative correlation.
4. Scenario IV shows that correlation is undefined if one of the patterns is constant across samples. As in the previous scenario, changing just one of the values for Gene 2 has a significant effect on the correlation.
    a. Change the value for sample 1 from 4 to 3, and note the effect on correlation.
    b. Change the value for sample 8 from 4 to 3, and note the effect on correlation.
    c. Explain why one of these changes has a greater magnitude effect than the other.
    d. Which single change from 4 to 3 would give the correlation nearest to 0? Why?

1. Possible answers:
    a. Change Scenario I, Gene 1, Sample 4, from 100 to –150. The correlation jumps up to approximately 0.68.
    b. Change Scenario I, Gene 1, Sample 8, from 100 to –150. The correlation jumps down to approximately –0.63.
2.
    a. Change the pattern for Gene 1 to be evenly spaced and increasing, for example, increasing from –100 to 110 in increments of 30. You can watch the correlation steadily approach 1 as you change the numbers for samples 1 through 8.
    b. Change the pattern for Gene 1 to be evenly spaced and decreasing, for example, decreasing from 110 to –100 in increments of 30.
3.
    a. By changing only samples 2 and 6 for Gene 2 from 7 to 6, the correlation jumps to nearly 0.78.
    b. By changing only samples 4 and 8 for Gene 2 from 7 to 6, the correlation falls to approximately –0.38.
4.
    a. Changing sample 1 causes correlation to jump to 0.632.
    b. Changing sample 8 causes correlation to jump to –0.577.
    c. The first change has greater magnitude impact on the correlation because Sample 4, with Gene 1 value of 100, keeps the line from dropping too far on the right when Sample 8 is changed. The line tries to be close to all sample points. The closest point to Sample 1 on the left end is Sample 2, but it is further from Sample 1 than Sample 4 is from Sample 8, so the "pull" on the line when Sample 8 is changed is not as great.
    d. Changing Sample 3, Gene 2, from 4 to 3 gives a correlation of 0.027. No other single change from 4 to 3 results in a correlation this close to 0. The reason this correlation is so near 0 is that this Gene 1 value (10) is closest to the average, so changing its Gene 2 value has little effect on the line of best fit.

Self-Quiz Sampler for Students

Do your students truly understand how clustering is peformed?
Can your students efficiently read color displays of microarray data?

Multiple Choice

1) If you change the correlation threshold for "cutting the tree" in hierarchical clustering from 0.8 to 0.5, you can be certain that the number of genes per cluster will:

a) decrease
b) increase
c) stay the same
d) not decrease
e) not increase
f) not stay the same

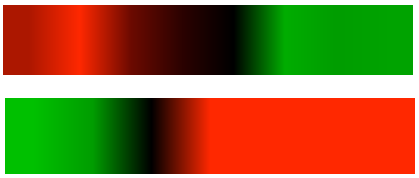
2) If you change the correlation threshold for "cutting the tree" in hierarchical clustering from 0.95 to 0.2, the number of clusters is likely to:

a) decrease
b) increase
c) stay the same
d) you cannot tell without seeing the data

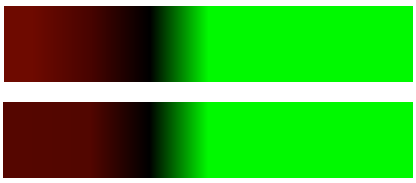
3) Which gene pair is likely to cluster together if the correlation threshold for "cutting the tree" in hierarchical clustering is set for 0.90?

A.

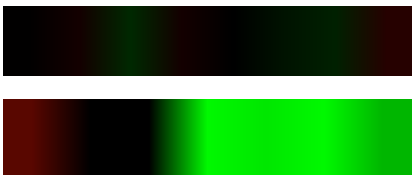
C.



B.



D.



True-False

1) Genes with similar function (e.g. cell cycle regulation) will likely have a correlation coefficient greater than 0.5.
2) Any two genes can be forced to be in the same cluster by decreasing the correlation threshold.
3) Two genes that cluster together under one set of experimental conditions will still cluster together under another set of experimental conditions at the same correlation threshold.

Open-ended Questions for Exploration

Using the Online Clustering Web Page, see if you can choose appropriate genes, conditions and correlation threshold to discover the following:

1) Six genes such that 3 are in one cluster and 3 are in a second cluster.
2) Six genes that fall into 6 different clusters, with a correlation threshold no greater than 0.7.
3) Under condition Heat Shock 1 (include all), genes YNL174W, YOL077C and YOR095C in
    a) one cluster
    b) two clusters
    c) three clusters
4) Five genes that cluster with YNL174W using a correlation threshold of 0.75.

Answer Key

Multiple Choice

1) d
2) a
3) b

True-False
1) False
2) True
3) False

Open-ended Questions for Exploration

1) There are many correct answers. One example: YNL007C, YOL151W, YNL134C are in one cluster and YOR361C, YPR190C, YOR095C are in another when using Heat Shock 1 and correlation threshold < 0.8.
2) There are many correct answers. One example: YAL015C, YHR104W, YKR024C, YEL055C, YKR066C, and YHL028W using Hydrogen Peroxide.
3) 0.9, 0.95, 0.99
4) YOR361C - YOR095C - YLR175W - YPR190C - YOL077C

Multiple-laboratory comparison of microarray platforms

Rafael A Irizarry[1], Daniel Warren[2], Forrest Spencer[3], Irene F Kim[4], Shyam Biswal[5], Bryan C Frank[6], Edward Gabrielson[7], Joe G N Garcia[8], Joel Geoghegan[9], Gregory Germino[4], Constance Griffin[10], Sara C Hilmer[11], Eric Hoffman[11], Anne E Jedlicka[12], Ernest Kawasaki[9], Francisco Martínez-Murillo[13], Laura Morsberger[10], Hannah Lee[5], David Petersen[9], John Quackenbush[6,14], Alan Scott[12], Michael Wilson[15,17], Yanqin Yang[2], Shui Qing Ye[8] & Wayne Yu[16]

Microarray technology is a powerful tool for measuring RNA expression for thousands of genes at once. Various studies have been published comparing competing platforms with mixed results: some find agreement, others do not. As the number of researchers starting to use microarrays and the number of cross-platform meta-analysis studies rapidly increases, appropriate platform assessments become more important. Here we present results from a comparison study that offers important improvements over those previously described in the literature. In particular, we noticed that none of the previously published papers consider differences between labs. For this study, a consortium of ten laboratories from the Washington, DC–Baltimore, USA, area was formed to compare data obtained from three widely used platforms using identical RNA samples. We used appropriate statistical analysis to demonstrate that there are relatively large differences in data obtained in labs using the same platform, but that the results from the best-performing labs agree rather well.

Microarray technology has become an important tool in medical science and basic biology research. A first time user will find many platform options and little guidance on which is the most appropriate for their application. Various comparison studies have been published presenting contradictory results. Some have observed agreement in results obtained with different platforms[1–6], others have not[7–10]. Here we demonstrate that the disagreement observed in some studies may be due to disputable statistical analyses. In particular, none of the prior studies have considered lab-to-lab variability (lab effect). The lab effect has been observed in all scientific fields[11]. Therefore, it is essential to assess this effect before drawing conclusions about platform performances.

A consortium of ten labs from the Washington, DC–Baltimore, USA, area was formed to compare the performance of three leading platforms. Researchers in each lab were given identical RNA samples that were processed according to what was considered best practice in each lab. Affymetrix GeneChips were used in five of the labs (Affymetrix labs 1–5), two-color spotted cDNA arrays were used in three labs (two-color cDNA labs 1–3), and two-color long oligonucleotide arrays were used in two labs (two-color oligo labs 1 and 2). Here we describe the features of our experiment that are necessary for such studies to be informative and a set of simple assessment measures useful for summarizing and interpreting the observed data.

To decide among various strategies for measuring the same quantity, one looks to optimize accuracy and precision. Because in many situations precision can be improved at the cost of accuracy, and vice versa, one tries to find the strategy providing the 'best' balance. Because the definition of best depends on the application, it is important to consider precision and accuracy in the context of a realistic problem. We mimicked the most common application of microarray technology: screening for a few candidate genes that appear to be differentially expressed among thousands of genes that are not. In this context, an appropriate comparison experiment requires at least the following three features. (i) To appropriately assess precision we should make a comparison with an *a priori* expectation of no-fold change for most or all genes. (ii) To appropriately assess accuracy, an *a priori* expectation of nonzero

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. [2]Department of Surgery, Johns Hopkins University, Baltimore, Maryland 21205, USA. [3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [4]JHU NIDDK Gene Profiling Center, Department of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA. [5]Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. [6]The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, Maryland 20878, USA. [7]Department of Pathology, Johns Hopkins University, Baltimore, Maryland 21231, USA. [8]Division of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, Mason F. Lord Bldg., Center Tower #665, Baltimore, Maryland 21224, USA. [9]NCI's Microarray Core Facility, Advanced Technology Center, Gaithersburg, Maryland 20877, USA. [10]Department of Pathology, Johns Hopkins University, School of Medicine, Baltimore, Maryland 21287, USA. [11]Research Center for Genetic Medicine, Children's National Medical Center, George Washington University, Washington, DC 20052, USA. [12]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. [13]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, Maryland 21205, USA. [14]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115-6084, USA. [15]Microarray Research Facility, Research Technologies Branch, DIR, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, USA. [16]Oncology Microarray Facility, Johns Hopkins University, Baltimore, Maryland 21231, USA. [17]Present address: Ambion, Inc., Austin, Texas 78744, USA. Correspondence should be addressed to R.A.I. (rafa@jhu.edu).

PUBLISHED ONLINE 21 APRIL 2005; DOI:10.1038/NMETH756

Table 1 | Assessment measures for all ten labs

		Precision			Proportion of agreement		
Platform	Lab number	Correlation	s.d.	Accuracy signal (s.e.m.)	25	50	100
Affymetrix oligo	1	0.48	0.32	0.62 (0.05)	0.72	0.56	0.54
Affymetrix oligo	2	0.76	0.17	0.64 (0.05)	0.80	0.70	0.70
Affymetrix oligo	3	0.67	0.24	0.66 (0.05)	0.68	0.66	0.60
Affymetrix oligo	4	0.79	0.15	0.59 (0.04)	0.80	0.70	0.65
Affymetrix oligo	5	0.59	0.25	0.58 (0.05)	0.64	0.68	0.55
Two-color cDNA	1	0.65	0.23	0.41 (0.12)	0.68	0.64	0.65
Two-color cDNA	2	0.68	0.21	0.13 (0.04)	0.28	0.30	0.38
Two-color cDNA	3	0.46	0.23	0.54 (0.09)	0.72	0.68	0.50
Two-color oligo	1	0.68	0.51	0.21 (0.09)	0.40	0.36	0.33
Two-color oligo	2	0.90	0.10	0.76 (0.13)	0.44	0.72	0.81

To summarize precision we used the correlation across replicate $\log_2$-fold change measurements and standard deviation (s.d.) of the difference between replicate $\log_2$-fold change measurements. To quantify accuracy we regressed the observed $\log_2$-fold changes of 16 genes against nominal $\log_2$-fold changes obtained using RT-PCR. The slope of the regression line defines what we refer to as accuracy signal. The proportion of agreement in interesting genes lists—ranked by fold change—of sizes 25, 50 and 100, created with replicate $\log_2$-fold change measurements, are also used to assess precision.

log–fold change of a few genes is needed. (iii) To be able to distinguish between platform effect and lab effect, at least two labs should provide data from each platform. We have designed the first platform comparison experiment that includes all of these features.

In general, the Affymetrix labs achieved better accuracy and precision. But overall, the best-performing lab was two-color oligo lab 2. Furthermore, two-color cDNA labs 1 and 3 outperformed most Affymetrix oligo labs in some categories. The worst performance was observed from a two-color oligo lab; thus the best and worst overall performance was achieved using the same platform. This underscores the importance of considering the lab effect. In general, we found that the lab had a larger effect on, for example, precision than did the platform, and that the results from the best-performing labs agreed rather well.

RESULTS

Assessment measures and plots

We created two samples in which we expect a few genes to be differentially expressed. To do this we developed a strategy based on mixtures from four knockout human cell lines that resulted in four specific genes with *a priori* expectation of fold change different from 1 (**Supplementary Methods** online). We refer to these genes as the altered genes. For each of these two samples we created an exact copy, or technical replicate, for a total of four samples. Exact copies of these four samples where hybridized by the ten labs using their platform of choice, and the resulting data were processed as described. We quantified relative expression between the two duplicate pairs of samples with $\log_2$-fold change. This resulted in two replicate $\log_2$-fold change measurements for each gene, from each lab.

To summarize precision we used two simple measures: correlation across replicate $\log_2$-fold change measurements and standard deviation (s.d.) of the difference between replicate $\log_2$-fold change measurements. These assessment measures can also be used to quantify the similarity between measurements made using different platforms. We refer to these two assessment measures as correlation and s.d. (**Table 1**, columns 3 and 4). A box plot of the differences used to compute the s.d. for each lab provides a graphical summary (**Fig. 1a**).

To assess accuracy we validated 16 genes using RT-PCR (**Supplementary Methods**). The 16 genes included the four altered genes, four randomly selected genes from those that were

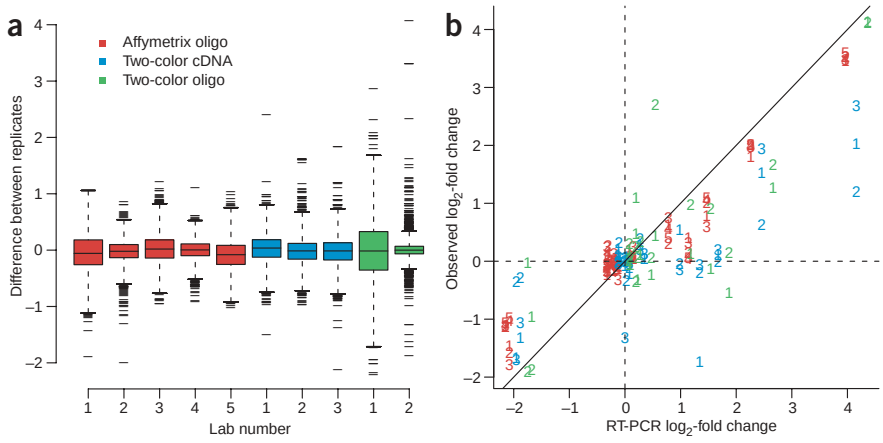


Figure 1 | Precision and accuracy assessment figures. (**a**) Box plot of the difference in $\log_2$-fold change between replicate measurements of gene expression from each of the ten labs. The platform used is represented by different colors defined in the figure. (**b**) Observed $\log_2$-fold change versus nominal (calculated from RT-PCR experiments) $\log_2$-fold change for the four altered genes and 12 other genes. The results for each of the 10 labs are represented by the lab number and color for the different platforms as in **a**. The solid diagonal line is the identity function and represents perfect accuracy.

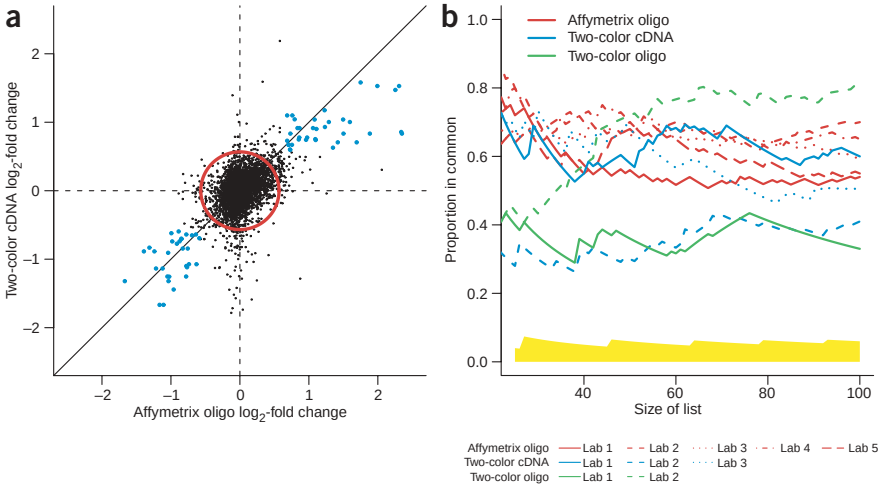


Figure 2 | Plots demonstrating agreement for differentially expressed genes. (**a**) Scatter plot of observed $\log_2$-fold change from two-color cDNA lab 1 and Affymetrix oligo lab 4. Points inside red circle represent genes that do not appear to be differentially expressed. Blue points are genes that appear to be differentially expressed. The solid diagonal line is the identity function and represents perfect accuracy. (**b**) CAT plot showing agreement between differential expression calls, defined by ranking genes by fold change, using replicate measurements from each lab. We considered list sizes less than 100 because we do not expect more than 100 genes to be differentially expressed, thus correspondence of larger lists is not of interest. The three colors represent the different platforms as in **Figure 1a**. The different line types represent the different labs within each platform so that a color and line-type pair uniquely represents each lab. The yellow strip represents critical values for rejecting the null hypothesis of no agreement at the 0.001 level.

consistently found not to be differentially expressed across all platforms, four genes that were consistently found to be differentially expressed across all platforms, and four genes found to be differentially expressed using one platform and not the others. To quantify accuracy we regressed the observed $\log_2$-fold changes of these 16 genes against nominal $\log_2$-fold changes obtained by RT-PCR analysis. The slope of the regression line defines our assessment measure, which we refer to as the signal (**Table 1**, column 5). A graphical summary is the scatter plot of the observed versus nominal values obtained by all labs (**Fig. 1b** and **Supplementary Fig. 1** online).

A scatter plot of the $\log_2$-fold changes obtained by the best-performing Affymetrix oligo and two-color cDNA labs showed no correlation for about 95% of genes (**Fig. 2a**). These genes had $\log_2$-fold changes close to zero and were probably not differentially expressed. Because for these genes it is likely that we measured zero $\log_2$-fold change plus random measurement error, we did not expect across-platform measurements to correlate. But for the few genes that appeared to be differentially expressed there was good agreement. In practice, we typically screen a small subset of genes that appear to be differentially expressed. Therefore, it is more important to assess agreement for genes that are likely to pass this screen. To account for this, we introduced a new descriptive plot: the correspondence at the top (CAT) plot. This plot is useful for comparing two procedures for detecting differentially expressed genes. To create a CAT plot we made a list of n candidate genes for each of the two procedures and plotted the proportion of genes in common against the list size n (**Fig. 2b**). As assessment measures, we reported the value of these curves for list sizes 25, 50 and 100. We refer to these assessment measures as the proportion of agreement (**Table 1**, columns 6, 7 and 8).

Preprocessing

We found that within- and across-platform performance can be greatly improved using alternative preprocessing algorithms to the defaults offered by the array manufacturers. For our analysis, probe-level data from the Affymetrix oligo arrays were preprocessed with the robust multiarray analysis (RMA)[12]. Print-tip normalization with no background correction was used to preprocess probe-level data from the two-color platforms[13]. Spot-quality information was ignored because we found it did not have substantial impact on downstream results. Because algorithms implementing these methodologies are available from the Bioconductor project[14], we will refer to them as the Bioconductor procedures. We compared the results obtained with this approach to those obtained with what we consider to be the default approaches: Affymetrix's MAS 5.0 algorithms for Affymetrix oligo arrays and median adjustment normalization with background correction for the two-color technologies. Although in general the default procedures had slightly better accuracy (not statistically significant), the gains in precision given by the Bioconductor procedures were dramatic. Because of the great improvement provided by the Bioconductor procedures (**Supplementary Fig. 2** online and **Supplementary Table 1** online), we use them for all of the experiments presented in this paper.

Annotation

To match features across platforms, we used mappings that match features to genomic entities that are available from various public databases. Resourcerer[15] provides mappings that link features to UniGene, LocusLink and RefSeq for all the platforms used in our experiment. Resourcerer also provides its own annotation in a

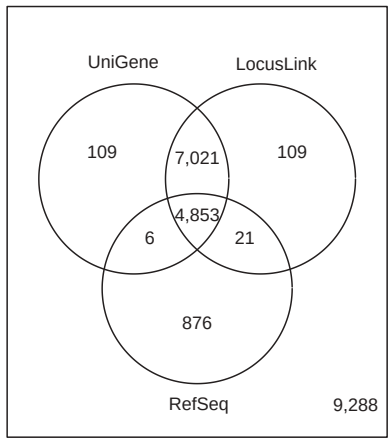


Figure 3 | Venn diagram illustrating agreement between annotation databases. For each mapping (UniGene, LocusLink and RefSeq) we obtained a different set of genes that had identifiers for each platform. This Venn diagram shows the agreement between these three different lists.

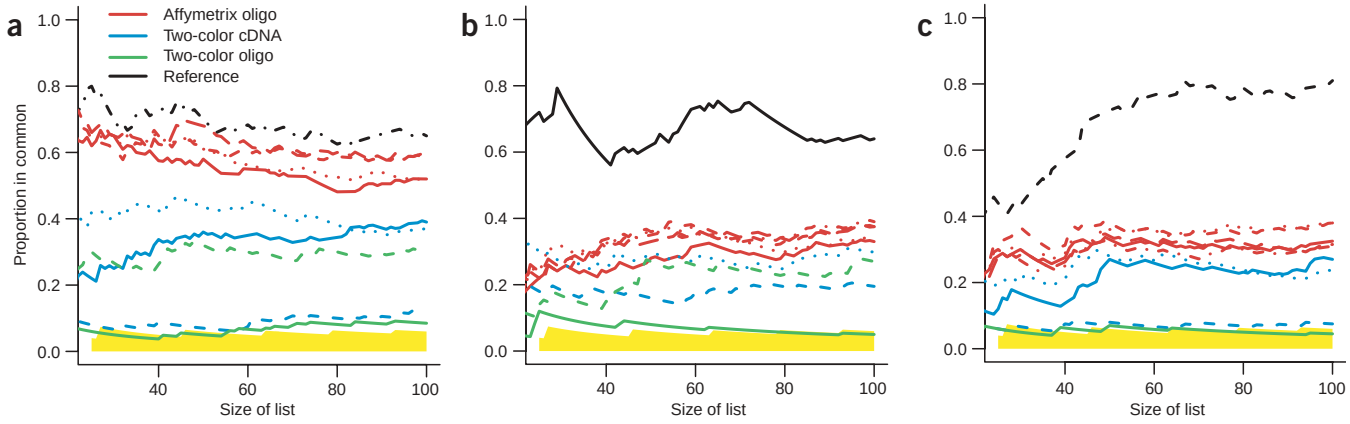


Figure 4 | CAT plots showing agreement in differential expression calls, based on fold change, between each lab and a reference lab. (**a**–**c**) The different line types represent the individual labs, and the three colors represent the different platforms as in **Figure 2b**. The black curve is the CAT curve comparing replicates from the reference lab. (**a**) CAT plot using data from the best-performing Affymetrix oligo lab as the reference. (**b**) CAT plot using data from the best-performing two-color cDNA lab as the reference. (**c**) CAT plot using data from the best-performing two-color oligo lab as a reference.

eukaryotic gene orthologs (EGO) database. Unfortunately, none of these mappings are one to one: not all the features in the arrays are annotated and/or some are annotated with more than one genomic identifier. Therefore, for a particular annotation only a subset of the array features will have an entry for each platform. Furthermore, these subsets differ depending on which annotation was used (**Fig. 3**). The annotation used had an effect on the across-platform agreement. For example, the correlation between measurements from Affymetrix oligo lab 4 and two-color cDNA lab 1 was 0.39–0.44 when using UniGene and EGO, respectively. We found that using the genes having entries in all databases for all platforms provided the best agreement. For all the analyses presented here we used the subset of genes obtained from this intersection (**Supplementary Table 2** online).

Platform comparison

Our results demonstrated that precision is comparable across platforms (**Table 1** and **Fig. 1a**). With the exception of two-color oligo lab 1, all the labs performed similarly, and it is clear that the lab effect is stronger than the platform effect. All the labs provided attenuated $\log_2$-fold change estimates, and this is consistent with previous observations[12] (**Fig. 1b**). In general, the labs using the Affymetrix platform seem to attain better accuracy than the labs using two-color platforms, although the best signal measure was attained by two-color oligo lab 2. Two-color cDNA lab 2 and two-color oligo lab 1 were clearly underperforming. The differences in data obtained by the other eight labs were not statistically significant.

We used CAT plots to assess across-platform agreement. It is important to note that these were used to compare results from single array experiments, and thus we did not expect perfect agreement. Note, for example, that the agreement of lists of the top 100 genes created from replicate fold-change measurements ranged from 33–81 percent (**Fig. 1b**). CAT plots comparing across-lab agreement demonstrate that the Affymetrix oligo labs consistently provided results similar to those from the best-performing labs (**Fig. 4**). This suggests that the Affymetrix platform provides by far the most consistent data across labs. Apart from two labs, there appears to be good agreement regardless of the platform used (**Fig. 4** and **Supplementary Table 3** online).

Table 2 | Correlation and s.d. measurements computed for absolute and relative measurements of expression

	Correlation		s.d.	
	Absolute	Relative	Absolute	Relative
Affymetrix oligo versus Affymetrix oligo	0.98	0.79	0.16	0.15
Two-color cDNA versus two-color cDNA	0.91	0.65	0.29	0.23
Affymetrix oligo versus two-color cDNA	0.40	0.44	0.91	0.25

Affymetrix oligo lab 4 and two-color cDNA lab 1 were used for this comparison.

DISCUSSION

We defined a series of assessment measures and plots used to compare three leading microarray platforms. These were justified by questions of scientific interest and have practical interpretations. The signal measure represents the expected $\log_2$-fold change in expression of a gene that should be differentially expressed with a nominal fold change of two, and the s.d. measure gives us the expected $\log_2$-fold change of a null gene. These two measures gave us a clear idea of the signal-to-noise ratio. Although, overall, the Affymetrix platform performed best, it is important to keep in mind that this platform is typically more expensive than the alternatives.

We also demonstrated that there was relatively good agreement between the Affymetrix labs and the best-performing two-color labs. These results contradict some previously published results that find disagreement across platforms[7–10]. The conclusions reached by these studies are likely due to three misconceptions. The first misconception is that absolute measurements of gene expression can be used to assess data across platforms. Note that both studies using absolute measurements had found disagreement[7,10]. Results established based on absolute measurements are misleading because they are adversely affected by platform-dependent probe effects that can be removed by considering relative measurements

of expression. The statistical model used to motivate our assessment measures, described in the Methods section, can be used to demonstrate this point. Note that in all studies interested in differential expression of genes, relative expression is the quantity of interest; thus this type of measurement is always available. The second misconception is that preprocessing has no significant effect on final results. With one exception[4], all previous studies had used algorithms that have been shown to be inferior to alternatives developed by the academic community[12,13]. Finally, the third misconception is that platform performance is not affected by lab. The existence of the sizable lab effect was ignored in all previously published comparison studies. This permits the possibility that studies done by, for example, experienced technicians may find agreement and studies done by less-experienced technicians may find disagreement (**Supplementary Fig. 3** online).

Although we found relatively good across-platform agreement, it is quite far from being perfect. In all across-platform comparisons, there was a small group of genes that had relatively large fold changes from data obtained using one platform but not using the others (**Fig. 2b**). We conjecture that some genes were not measured correctly, not because the technologies are not performing adequately, but because transcript information and annotation can still be improved.

Our results provide a useful assessment of three leading technologies and demonstrate the need for continued cross-platform comparisons. In fact, Affymetrix has released a new platform for measuring gene expression in humans, which yields slight improvements in accuracy and precision (**Supplementary Figs. 1** and **4** online and **Supplementary Table 4** online). We expect our study to serve as a starting point for larger, more comprehensive comparisons. Furthermore, our findings show that improved quality assessment standards are needed. Assessments of precision based on comparisons of technical replicates appear to be standard operating procedure among, at least, academic labs. We have demonstrated that precision and accuracy assessments are not informative unless performed simultaneously. We hope that our study serves as motivation to create such standards. This will be essential for the success of microarray technology as a general measurement tool.

METHODS

Data analysis. A commonly used statistical model for microarray data is $Y_{ijk} = \theta_i + \phi_{ij} + \varepsilon_{ijk}$, in which Y_{ijk} represents measurement k of $\log_2$-scale expression of gene i measured by platform j. Here θ_i represents absolute gene expression in the $\log_2$ scale. ϕ_{ij} denotes the platform-specific probe or spot effect. Measurement error is represented by ε_{ijk}. For illustrative purposes we considered each of the effects in this model to be random and statistically independent from each other. We represented their variances with ν_θ, ν_ϕ and ν_ε.

Many researchers have observed a sizeable probe effect in microarray data, which implies that ν_ϕ is large[12]. This will result in artificially large correlations when comparing absolute measurements obtained using the same platform. To see this, note that within-platform correlation is $\mathrm{corr}(Y_{ij1}, Y_{ij2}) = (\nu_\theta + \nu_\phi)/(\nu_\theta + \nu_\phi + \nu_\varepsilon)$. This correlation is typically close to one, but only because ν_ϕ can be much larger than ν_θ and ν_ε. If we compare across platforms, the correlation will not be as large, but only because the probe effect is not common to the two platforms and therefore does not affect the correlation $\mathrm{corr}(Y_{i1k}, Y_{i2k}) = \nu_\theta/(\nu_\theta + \nu_\phi + \nu_\varepsilon)$. These theoretical predictions were confirmed empirically (**Table 2**).

A simple solution to the probe effect problem is to consider relative expression instead of absolute expression. Most experiments compare between different samples, thus in general this type of measure is readily available. By considering the difference of the Y_{ijk} from the two samples, the $\phi_{i,j}$ are cancelled out. Because these are $\log_2$-scale measurements this difference is simply the $\log_2$ ratio of the absolute expression levels.

For the relative expression measurements, the within-platform correlations were substantially smaller and the across lab correlation was a bit larger (**Table 2**). We propose that only assessments based on relative expression are useful. All the results presented in this paper deal with relative expression.

Additional methods. Sample preparation, RT-PCR and microarray hybridization and experimental design are described in **Supplementary Methods** online. The code and data necessary to reproduce this work are available online (http://www.biostat.jhsph.edu/~ririzarr/techcomp).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank A. Nones and K. Broman for useful suggestions. The work of R.A.I. is partially funded by the National Institutes of Health Specialized Centers of Clinically Oriented Research (SCCOR) translational research funds (212-2494 and 212-2496). The work of G. Germino and I. Kim was partially funded by NIDDK U24DK58757.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 30 November 2004; accepted 22 March 2005
Published online at http://www.nature.com/naturemethods/

1. Kane, M. *et al.* Assessment of the sensitivity and specificity of oligonucleotide (50-mer) microarrays. *Nucleic Acids Res.* **28**, 4552–4557 (2000).
2. Hughes, T. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347 (2001).
3. Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. & Sealfon, S.C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48 (2002).
4. Barczak, A. *et al.* Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* **13**, 1775–1785 (2003).
5. Carter, M. *et al. In situ*-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. *Genome Res.* **13**, 1011–1021 (2003).
6. Wang, H. *et al.* Assessing unmodified 70-mer oligonucleotide performance on glass-slide microarrays. *Genome Biol.* **4**, R5 (2003).
7. Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L. & Kohane, I. Analysis of mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
8. Kothapalli, R., Yoder, S., Mane, S. & Loughran, T.P. Jr. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
9. Li, J., Pankratz, M. & Johnson, J. Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.* **69**, 383–390 (2003).
10. Tan, P. *et al.* Evaluation of gene expression measurements from commercial platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
11. Youden, W. Enduring values. *Technometrics* **14**, 1–11 (1972).
12. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
13. Dudoit, S. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
14. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
15. Tsai, J. *et al.* Resourcerer: a database for annotating and linking microarray resources within and across species. *Genome Biol.* **2** software0002.1–0002.4 (2001).

Addendum: Standardizing global gene expression analysis between laboratories and across platforms

Members of the Toxicogenomics Research Consortium
Nat. Methods **2**, 351–356 (2005).

The authors were listed as Members of The Toxicogenomics Research Consortium, with a complete list of authors available in a **Supplementary Note** online. A complete list of authors in alphabetical order and their affiliations follows.

Theodore Bammler[1], Richard P Beyer[1], Sanchita Bhattacharya[2], Gary A Boorman[3], Abee Boyles[4], Blair U Bradford[5], Roger E Bumgarner[6], Pierre R Bushel[7], Kabir Chaturvedi[8], Dongseok Choi[9], Michael L Cunningham[3], Shibing Deng[5], Holly K Dressman[4], Rickie D Fannin[7], Fredrico M Farin[1], Jonathan H Freedman[4], Rebecca C Fry[2], Angel Harper[8], Michael C Humble[10], Patrick Hurban[8], Terrance J Kavanagh[1], William K Kaufmann[5], Kathleen F Kerr[11], Li Jing[12], Jodi A Lapidus[9], Michael R Lasarev[13], Jianying Li[7], Yi-Ju Li[4], Edward K Lobenhofer[8], Xinfang Lu[14], Renae L Malek[15], Sean Milton[2], Srinivasa R Nagalla[14], Jean P O'Malley[14], Valerie S Palmer[13], Patrick Pattee[14], Richard S Paules[7], Charles M Perou[5], Ken Phillips[8], Li-Xuan Qin[11], Yang Qiu[8], Sean D Quigley[1], Matthew Rodland[14], Ivan Rusyn[5], Leona D Samson[2], David A Schwartz[4], Yan Shi[5], Jung-Lim Shin[12], Stella O Sieber[7], Susan Slifer[4], Marcy C Speer[4], Peter S Spencer[13], Dean I Sproles[13], James A Swenberg[5], William A Suk[10], Robert C Sullivan[12], Ru Tian[5], Raymond W Tennant[7], Signe A Todd[14], Charles J Tucker[7], Bennett Van Houten[10], Brenda K Weis[10], Shirley Xuan[2] & Helmut Zarbl[12]

[1]Department of Environmental and Occupational Health Sciences and the Center for Ecogenetics and Environmental Health, University of Washington, Box 357234, Seattle, Washington 98195-7234, USA. [2]Biological Engineering Division, Center for Environmental Health Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 56-235 Cambridge, Massachusetts 02139, USA. [3]National Toxicology Program, National Institute of Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [4]Duke University Medical Center, Department of Medicine, DUMC 2629, Room 275 MSRB, Research Drive, Durham, North Carolina 27710, USA. [5]Lineberger Comprehensive Cancer Center, University of North Carolina, CB#7295, Chapel Hill, North Carolina 27599, USA. [6]Department of Microbiology, Box 358070, University of Washington, Seattle, Washington 98195, USA. [7]National Center for Toxicogenomics, National Institute of Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [8]Icoria, Inc., 108 TW Alexander Drive, Building 1A, Research Triangle Park, North Carolina 27709, USA. [9]Division of Biostatistics, School of Public Health and Preventive Medicine, School of Medicine, Oregon Health & Science University, Portland, Oregon 97239, USA. [10]Division of Extramural Research and Training, National Institute for Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [11]Department of Biostatistics, University of Washington, Box 357234, Seattle, Washington 98195-7234 USA. [12]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave North, Mailstop C1-1015, Seattle, Washington 98109, USA. [13]Center for Research on Occupational and Environmental Toxicology, Oregon Health & Science University, Portland, Oregon 97239, USA. [14]Center for Biomarker Discovery, Department of Pediatrics, Oregon Health & Science University, Portland, Oregon 97239, USA. [15]The Institute for Genomic Research, Rockville, Maryland, USA.

Corrigendum: Multiple-laboratory comparison of microarray platforms

Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martínez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye & Wayne Yu
Nat. Methods **2**, 345–349 (2005).

The GEO accession number for the array data is GSE2521.

Standardizing global gene expression analysis between laboratories and across platforms

Members of the Toxicogenomics Research Consortium[1]

To facilitate collaborative research efforts between multi-investigator teams using DNA microarrays, we identified sources of error and data variability between laboratories and across microarray platforms, and methods to accommodate this variability. RNA expression data were generated in seven laboratories, which compared two standard RNA samples using 12 microarray platforms. At least two standard microarray types (one spotted, one commercial) were used by all laboratories. Reproducibility for most platforms within any laboratory was typically good, but reproducibility between platforms and across laboratories was generally poor. Reproducibility between laboratories increased markedly when standardized protocols were implemented for RNA labeling, hybridization, microarray processing, data acquisition and data normalization. Reproducibility was highest when analysis was based on biological themes defined by enriched Gene Ontology (GO) categories. These findings indicate that microarray results can be comparable across multiple laboratories, especially when a common platform and set of procedures are used.

Transcriptional profiling using DNA microarrays is one of many genomic tools that is now being used to characterize biological systems. Despite the increasing reliance on this technology by the scientific community, the reproducibility of microarray data between laboratories and across platforms has not been adequately addressed. Now there is a range of DNA microarray platforms including one- and two-channel formats, cDNA and oligonucleotide microarrays, in-house spotted microarrays, and commercially developed microarrays. There is also great diversity in the protocols used by different laboratories for RNA preparation and labeling, as well as in the instrumentation and software used for these procedures. Moreover, there are many computational and statistical tools for analyzing microarray images, quantitating spot intensities, normalizing and background-correcting these data, and for determining which transcripts are differentially expressed[1–3]. The impact of these multifaceted approaches toward assessing global gene expression remains inadequately characterized[4–7]. The issue of data reproducibility and reliability is crucial to the generation of, and ultimately to the utility of, large databases of microarray results[8,9]. Although the Microarray Gene Expression Data (MGED) Society has coordinated an impressive effort to develop guidelines for publishing microarray data through the minimal information about microarray experiments (MIAME) standards[10,11], these efforts have focused on documentation of experimental details and results, and therefore do not directly address issues of reproducibility between laboratories or across platforms. It is thus critical to determine the effect of methodological variables on the reproducibility, validity and generalizability of the results.

The Toxicogenomics Research Consortium was established in November 2001, with advancing the application of gene expression technologies in toxicology as one of it goals. The first Consortium study systemically assessed microarray data and reproducibility of the results within and between laboratories, as well as within and between microarray platforms. In doing so, potential sources of inter- and intralaboratory error and variability in the microarray experimental results were identified. Two standard microarrays were used by the Consortium laboratories: a spotted long oligonucleotide microarray, produced by one of the Consortium laboratories (designated the standard spotted array), and a commercially produced long oligonucleotide microarray (designated the standard commercial array). Consortium members also used a variety of other microarray platforms that were 'resident' at each laboratory (**Fig. 1**). The resident arrays included both commercial microarrays and in-house spotted microarrays, in long oligonucleotide, short oligonucleotide and cDNA formats.

Each laboratory was provided with aliquots of two different RNA samples that were prepared in one of the Consortium laboratories—a sample prepared from mouse livers (liver RNA; L), and a sample prepared from equal amounts of RNA isolated from five mouse tissues, liver, kidney, lung, brain and spleen (pooled RNA; P). The microarray hybridizations were designed to determine the reproducibility of gene expression measurements, the reproducibility of measuring differential transcript representation when comparing liver RNA to pooled RNA, and the feasibility of deriving comparable results across disparate laboratories and platforms[12]. These common RNA samples allowed us to focus on variation in the technical and analytical approaches to microarray experimentation without biological variation[1,13,14]. The results demonstrate that the highest level of reproducibility between laboratories was observed when a commercial microarray was used together with standardized protocols for RNA labeling, hybridization, microarray processing, data acquisition and data normalization. Whereas this

[1]A list of authors and their affiliations appears in the **Supplementary Note** online. Correspondence should be addressed to B.K. Weis (weis@niehs.nih.gov).

PUBLISHED ONLINE 21 APRIL 2005; DOI:10.1038/NMETH754

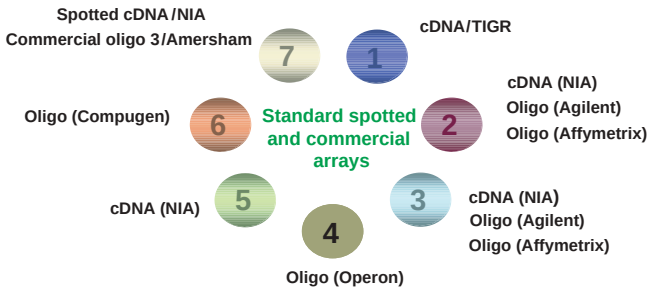


Figure 1 | DNA microarray platforms used across laboratories. Seven laboratories used a total of 12 microarray platforms: seven spotted resident cDNA and oligonucleotide platforms, three commercial resident platforms and two standard microarray platforms. Each colored circle represents a laboratory. Resident platforms are represented around the outer periphery of the ring of circles; the standard array platforms are represented in the middle of ring of circles (see Methods for description of platforms). An eighth laboratory (not shown), the provider of the standard commercial array, contributed to dataset D.

may be expected, the extent of the improvement in reproducibility that was obtained by such standardization was surprising. Notably, even with low levels of data correlation, the biological themes that emerge from these results are remarkably consistent.

RESULTS

Reproducibility of expression intensity with standard arrays

The standard spotted arrays and common RNA samples were used to generate dataset A (**Fig. 1** and **Supplementary Table 1** online). Researchers in laboratories 1–7 each carried out eight hybridizations: four that cohybridized liver RNA labeled with both Cy3 and Cy5 (LvsL), and four that cohybridized liver RNA and pooled RNA (LvsP). Each set comprised two dye-swapped samples[15,16]. In each of these seven laboratories researchers used their own protocols for mRNA labeling, microarray hybridization, image acquisition and data analysis (**Supplementary Methods** online). These data were combined based on Unigene IDs. The reproducibility of raw intensity values was fairly high within each laboratory for LvsL, with median correlation coefficients ranging from 0.73 to 0.90 (**Fig. 2a**). When the data from each laboratory were compared to the collective data from the other laboratories, however, the correlations were significantly lower, between 0.21 and 0.41 across laboratories (**Fig. 2b**). The intensity values for the pooled RNA samples were extracted from LvsP data and used to make PvsP comparisons *in silico*. The same trends were revealed (**Figs. 2c**,**d**, green symbols): correlation coefficients for PvsP ranged from 0.68 to 0.91 within laboratories and from 0.23 to 0.44 across laboratories. Thus, inter-laboratory differences negatively impact data reproducibility as measured by raw intensity values.

The first step toward evaluating reproducibility across laboratories was to standardize methods for the entry, storage and retrieval of the microarray data generated from different platforms and analysis software packages. Although this step did not significantly affect data correlation within each laboratory, the correlations across laboratories were improved dramatically (Dataset B; **Fig. 2** and **Supplementary Table 1** online). Specifically, the median LvsL correlation across laboratories improved from 0.33 to 0.56, and the PvsP correlation improved from 0.32 to 0.59. This indicates that an important source of variability for Dataset A was in the data handling methods; using standardized file formats and gene nomenclature improved the ability to detect correlations that are inherent in the data.

To evaluate the impact of image analysis methods on data reproducibility, we reanalyzed each microarray image from all laboratories (for Dataset B) using the same software package (GenePix Pro v4.1.1.28, Axon) and a common set of feature extraction parameters (dataset C; **Supplementary Table 1**). Standardizing the method for image analysis did not significantly affect the within-laboratory correlations, but did result in a modest increase in the correlation of intensity data across the seven laboratories (**Fig. 2**). The median correlation coefficient across laboratories for LvsL increased from 0.56 to 0.59, and for LvsP from 0.59 to 0.64.

Notwithstanding these standardization efforts, the best intensity correlation coefficient values across laboratories (0.59–0.64) were relatively poor (**Supplementary Table 1**). This was likely due to the diverse RNA labeling and hybridization methods used by the laboratories. To address this issue, the LvsL and LvsP experiments described above were repeated in each of eight laboratories using common protocols for RNA labeling and hybridization, and the standard commercial array (Dataset D). We applied a standard file format, nomenclature and image analysis protocol to these data. Intensity correlation coefficients were markedly improved for the within-laboratory comparisons and marked improvements in across-laboratory correlations were realized (**Fig. 2**). This was observed for both LvsL (**Figs. 2a**,**b**) and PvsP values (**Figs. 2c**,**d**), with median correlation coefficients improving to 0.87–0.92

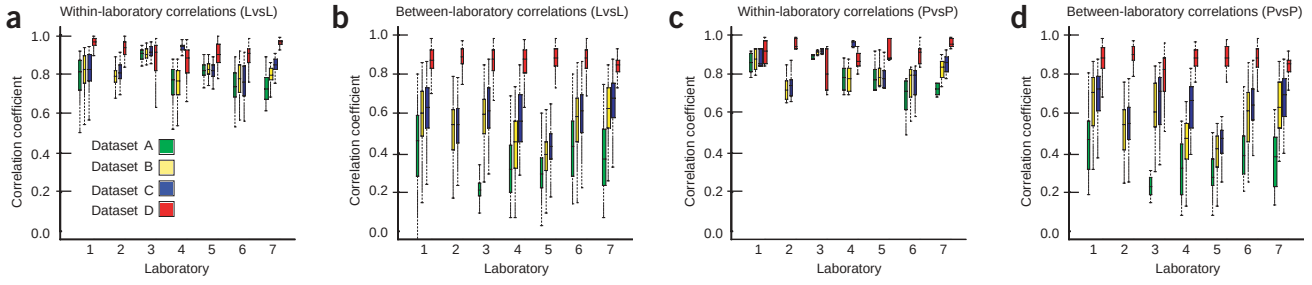
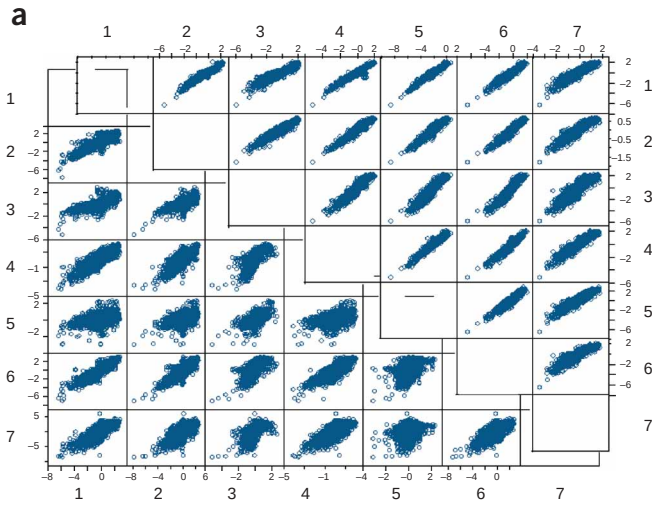


Figure 2 | Within- and between-center Pearson correlation coefficients for gene expression intensity using standard arrays. (**a**–**d**) Liver and pooled RNA samples were hybridized to two common platforms. Pearson correlations of raw intensity measurements were calculated for all possible pairwise combinations either within a laboratory (**a**,**c**) or between laboratories (**b**,**d**) on either the standard spotted arrays (datasets A–C) or the standard commercial array (dataset D). The box plots represent median values with upper and lower quantiles; the dotted lines represent maximum and minimum values.



	1	2	3	4	5	6	7
1		0.91	0.84	0.79	0.93	0.93	0.87
2	0.86		0.93	0.90	0.93	0.93	0.90
3	0.67	0.68		0.87	0.90	0.91	0.86
4	0.80	0.78	0.69		0.89	0.87	0.88
5	0.23	0.27	0.39	0.25		0.92	0.91
6	0.88	0.84	0.61	0.76	0.22		0.87
7	0.74	0.65	0.49	0.67	0.13	0.67	

Figure 3 | Within and between laboratory Pearson correlation coefficients for $\log_2$ gene expression ratios using standard arrays. (**a**,**b**) Liver and pooled RNA samples were hybridized to two common array platforms in seven laboratories (1–7). Average $\log_2$ gene expression ratios were calculated across laboratories and were used to calculate correlation coefficients. Graphic display of pair-wise comparisons of average $\log_2$ gene expression ratios for liver versus pooled RNA samples plotted for all genes on the standard spotted array (**a**, lower panel) and the standard commercial array (**a**, upper panel). Pearson correlation coefficients for pair-wise comparisons across laboratories for the standard spotted array (**b**, lower panel) and the standard commercial array (**b**, upper panel). Correlation coefficients greater than 0.80 are highlighted (red).

(**Supplementary Table 1**). Thus, standardization of RNA labeling and hybridization protocols is an important contributor to signal intensity correlations across laboratories.

Reproducibility of expression ratios with standard arrays

It is arguably important to evaluate reproducibility in gene expression ratio measurements between laboratories and across platforms. Indeed, in most transcriptional profiling studies, it is ultimately the relative changes in gene expression ratios that are used to infer biological mechanisms and state changes. The first task was to establish how the transcript level ratios between liver and pooled RNA (LvsP) varied depending on the method used for data normalization and background subtraction. We applied four different approaches to Dataset C, wherein file format, nomenclature and image analysis were standardized, but different labeling and hybridization protocols were used in each laboratory. We found the highest median correlation (0.69) of the LvsP $\log_2$ ratios between laboratories using Lowess normalization without background subtraction (**Supplementary Table 2** online). Thus, we applied Lowess normalization without background adjustment to the gene expression measurements generated from all LvsP sample comparisons.

We calculated Pearson correlation coefficients comparing the average expression ratios across laboratories for each transcript feature on each platform, approximately 18,000 for datasets B and C and 20,000 for dataset D (**Fig. 3**). Similar to the raw intensity correlations (**Fig. 2**), the highest reproducibility was observed for dataset D, in which essentially all procedures were standardized. Once again, there was a marked increase in reproducibility in dataset D relative to dataset C.

Reproducibility of expression ratios with resident arrays

To compare expression ratios across noncommercial resident arrays (**Figs. 1** and **4**), we identified a set of common transcripts present on all 12 platforms (**Supplementary Table 3** online and **Supplementary Methods**). Using stringent criteria, only 502 transcripts were matched across the 12 microarray platforms. We limited our analysis to these 502 genes to minimize the possibility that poor correlations between platforms could be due to gene misidentification. We applied Lowess normalization without

Figure 4 | Resident array Pearson correlation. Liver and pooled RNA samples were hybridized to seven different resident microarray platforms at eight laboratories (1–8). The average gene expression ratios were calculated across replicate microarrays within each laboratory and Pearson correlation coefficients were calculated for the set of 502 common genes (white boxes). In addition, average pair-wise correlations between different array replicates for each laboratory/platform combination were calculated (grey boxes). Labels comprise the laboratory number, type of array used and the source of probes for the array.

	1 Spotted cDNA/TIGR	2 Spotted cDNA/NIA	2 Commercial oligo/Agilent MD	2 Commercial oligo/Affymetrix	3 Spotted cDNA/NIA	3 Commercial oligo/Agilent MD	3 Commercial oligo/Affymetrix	4 Spotted oligo/Operon	5 Spotted cDNA/NIA	6 Spotted oligo/Compugen	7 Commercial oligo/Amersham	7 Spotted cDNA/NIA	8 Commerical oligo/Agilent MD
1 Spotted cDNA/TIGR	0.87												
2 Spotted cDNA/NIA	0.76	0.95											
2 Commercial oligo/Agilent MD	0.58	0.62	0.45										
2 Commercial oligo/Affymetrix	0.69	0.75	0.60	0.97									
3 Spotted cDNA/NIA	0.46	0.49	0.35	0.37	0.42								
3 Commercial oligo/Agilent MD	0.68	0.75	0.75	0.67	0.39	0.97							
3 Commercial oligo/Affymetrix	0.68	0.75	0.58	0.91	0.36	0.65	0.90						
4 Spotted oligo/Operon	0.48	0.46	0.45	0.50	0.26	0.47	0.48	0.79					
5 Spotted cDNA/NIA	0.26	0.25	0.25	0.15	0.33	0.23	0.13	0.15	0.21				
6 Spotted oligo/Compugen	0.64	0.72	0.53	0.65	0.29	0.64	0.65	0.40	0.17	0.73			
7 Commercial oligo/Amersham	0.58	0.65	0.46	0.64	0.20	0.57	0.64	0.41	0.11	0.57	0.96		
7 Spotted cDNA/NIA	0.57	0.70	0.42	0.56	0.29	0.48	0.60	0.39	0.17	0.52	0.49	0.37	
8 Commerical oligo/Agilent MD	0.54	0.66	0.67	0.67	0.35	0.73	0.69	0.38	0.17	0.60	0.50	0.48	0.95

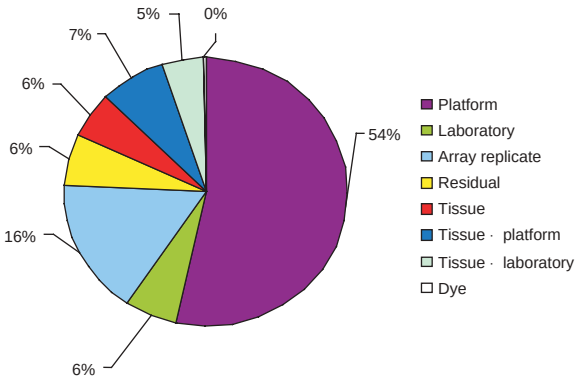


Figure 5 | Sources of variation in gene expression measurements across microarray platforms and laboratories for resident arrays. Contributions of different sources of variability were estimated with an ANOVA mixed model. Microarray platform was the largest source of variability, followed by laboratory and array-to-array replication (array replicate).

background subtraction as described above. For the single-color resident arrays (Affymetrix), we applied quantile normalization[17].

We ran the two-color resident arrays in quadruplicate and the one-color arrays in duplicate. We calculated average expression ratios, as described above, across the replicate microarrays. We calculated median Pearson correlation coefficients for the set of 502 common transcripts (**Fig. 4**). As before, the reproducibility for each platform within its resident laboratory was generally very good, in particular for the commercial platforms. Overall, the cross-platform correlations were extremely poor both within and between laboratories, although we noted a few acceptable correlations (>0.75). We performed hierarchical clustering of the $\log_2$-ratio values for the 502 common genes (**Supplementary Fig. 1** online). Overall, we obtained similar ratios for a considerable percentage of the common genes for a majority of laboratory and platform combinations. For example, 69% of all laboratory and platform combinations had correlations greater than 0.70; the highest correlation was observed within dataset D (0.93). The remaining laboratory and platform combinations had lower overall $\log_2$ ratios and did not correlate as well.

Microarray platform contributes most to reproducibility

To assess the relative contribution of the different sources of technical variability in our gene expression measurements, we fitted an ANOVA random effects model to the LvsL and LvsP normalized data from the resident array platforms. For each of the 502 common transcripts, the model was used to partition the observed variability in the data into variability owing to platform, laboratory, microarray replicate, residual, tissue, tissue × platform, tissue × laboratory and dye[18] (**Fig. 5** and **Supplementary Methods**). These results indicate that more than half of the variability observed in these data is attributable to the microarray platform; differences between replicate microarrays and between different laboratories contributed substantially less.

Emergence of biological themes using Gene Ontology

We found considerable variability in gene expression using gene-by-gene comparisons. Subsequently, we determined whether consistent biological themes could nonetheless be identified among different microarray platforms and laboratories. We identified the differentially expressed genes in the LvsP data for each laboratory and platform combination, and used the lists to identify enriched GO categories using EASE[19] (**Supplementary Methods**). During the generation of the lists of differentially expressed genes, we noted a marked improvement in concordance (percent overlap in significantly induced or repressed genes based on pair-wise comparisons of gene lists across laboratories) of gene lists for datasets with increased standardization of technical methods (**Supplementary Table 4** online). For example, we observed good overall concordance for dataset D, up to 80%. In addition, we found 277 transcripts that were significantly regulated, as defined by fold change plus an error term (described in **Supplementary Methods**) in all eight laboratories that contributed to Dataset D. For Dataset C, concordance was considerably lower, only as high as 52.4 percent, and only 13 genes were found to be significantly regulated in the six laboratories contributing to Dataset C.

We identified a list of 106 significant GO nodes that clustered into three main branches across 24 laboratory and platform combinations (**Fig. 6** and **Supplementary Table 5** online). Concordance was highest for branch 2, which primarily represented dataset D. More than 50% of the functionally enriched GO nodes had 70% concordance within branch 2, whereas less than 50%

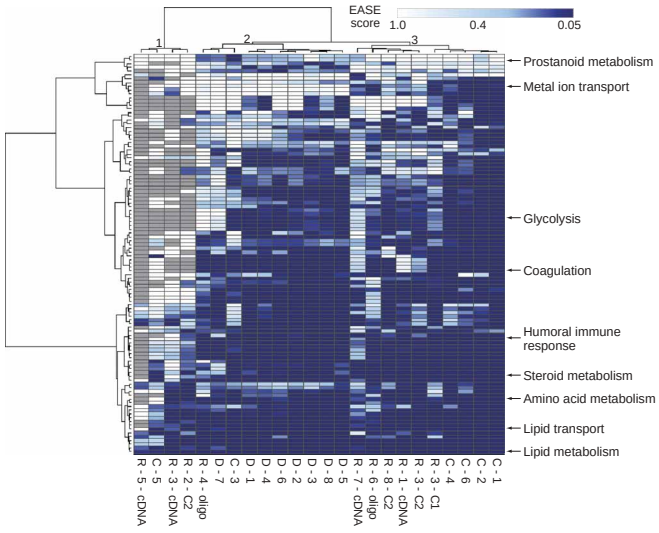


Figure 6 | Clustering of 24 laboratory and platform combinations based on common GO nodes. Common GO nodes were selected by two criteria: an EASE score was calculated for at least 20 of the 24 laboratory and platform combinations, and the EASE score was significant ($P < 0.05$) for at least one of the laboratory and platform combinations. This resulted in a list of 106 common GO nodes. Hierarchical clustering of both the laboratory and platforms, and the common GO nodes was performed using the calculated EASE scores. The relationship between the color intensity and EASE score is illustrated by the color bar. Gray indicates that an EASE score was not calculated for that GO node. The laboratory and platform is denoted by the letter and number combination at the bottom of every column. C = dataset C (standard spotted array with data extracted from a common image analysis software package), D = dataset D (standard commercial array), R = resident array. The number details which of the eight laboratories performed the hybridizations (see **Figure 1**). Resident arrays are also described by the type of array they are cDNA = spotted cDNA, oligo = spotted oligonucleotide, C#1 = commercial oligonucleotide arrays from Affymetrix and C#2 = commercial oligonucleotide arrays from Agilent. Numbers on the upper *x*-axis refer to branches of the dendogram.

concordance was observed across all three branches. The decline in concordance across datasets is likely due to the impact of branch 1, which had relatively little GO node enrichment.

We found many similar biological themes across most laboratory and platform combinations (**Fig. 6**). For example, three GO nodes demonstrated enrichment across multiple laboratory and platform combinations: steroid metabolism, humoral immune response and coagulation. Enrichment of these nodes is readily explained by the samples used in this study. The liver is a principal site of steroid metabolism; it was thus expected to be an enriched node in liver RNA when compared to pooled RNA. Likewise, the spleen is an initiating organ in the humoral immune response; the presence of spleen RNA in the pooled sample resulted in an enrichment of this node relative to the liver sample. Finally, the liver has a role in coagulation through the synthesis of coagulation factors (for example, coagulation factor IX) and hepatocyte nuclear factors, thus explaining the enrichment of this GO node.

Notably, the EASE score for the coagulation nodes on three resident platforms (R7-cDNA, R1-cDNA and R3-C#2) was not as significant (EASE score >0.05) as for other laboratory and platform combinations. This observation can be explained by the different transcript representation across the arrays. For both the standard spotted and standard commercial arrays, approximately 60 genes map to the coagulation GO node. In contrast, this node is represented by far fewer genes on the three resident arrays: 19 genes (R7-cDNA), 22 genes (R1-cDNA) and 25 genes (R3-C#2). The EASE score is a function of both the number of genes for a given GO node present on the array and the number of genes present in the list of differentially expressed genes for that node. Evaluating this further, several genes within the coagulation GO node (for example, fibrinogen, coagulation factor X and serine (or cysteine) proteinase inhibitor; *Serpind1*) were identified as differentially expressed in the majority of the laboratory and platform combinations; however, none of these genes were represented on the arrays that were used for R7-cDNA, R1-cDNA, and R3-C#2 (all of which represent distinct platforms). Therefore, the less significant EASE scores (>0.05) for these resident arrays were likely due to a decreased representation of this GO node on these arrays.

DISCUSSION

Our results indicate that technical variables such as the microarray platform, the labeling and hybridization protocols, and the approaches to data analysis can profoundly affect the comparability of gene expression experiments between laboratories. Comparability is highest when these technical variables are standardized. We found that comparable biological themes emerge from data across disparate platforms and laboratories when GO nodes are used to analyze collections of genes representative of biological themes in lieu of direct gene-by-gene comparisons. This method of analysis may therefore prove useful for mitigating potentially confounding factors inherent in multisite and multiplatform data. The relationships between GO categories, however, take the form of directed acyclic graphs, meaning that 'child' categories can have multiple 'parents'. Thus, differences (and similarities) between datasets can be exaggerated because related nodes (for example, regulation of body fluids, hemostasis and coagulation) can contain some of the same genes. Therefore, the identity of nodes and their interrelatedness should be considered when attempting to assess reproducibility and concordance of disparate datasets.

Our findings have important lessons for the field of genomics. First, as one begins to use genomics to identify biological responses or states, one must carefully assess the platform and experimental (analytical) protocols used by the investigators. Our results demonstrate that the microarray platform can be a source of substantial gene expression variability and that commercial microarrays, for a variety of reasons (such as uniform labeling and hybridization techniques, consistent quality of the microarray itself), yield results that are more comparable between laboratories. Second, using genomics to identify environmental-response genes and biological pathways will require external validation, preferably through focused, independent hypothesis-testing experiments. Thus, gene expression results from microarray studies originating from one laboratory should be considered to be a foundation for developing testable hypotheses that can be addressed in subsequent experiments. Third, our findings demonstrate that the generalizability of gene expression studies can be limited between independent laboratories and across platforms. Although independent laboratories can clearly achieve similar results, this can be greatly facilitated by a substantial commitment to using harmonized experimental protocols, similar approaches to image and data analysis, and similar or identical microarray platforms. Fourth, similar biological themes can emerge from results obtained in the absence of harmonization, although the findings should be treated with caution. Although we found common GO categories across laboratories and platforms, there were also several distinct differences, indicating that it is easy to overinterpret microarray results. Finally, our findings indicate that creation of gene expression databases that incorporate results from multiple laboratories will be most useful if experimental standards are developed, and data filters are applied before consolidation of the individual experimental results. Without these steps, the results from comparisons across laboratories can be misleading and should be considered with appropriate caution.

METHODS

Microarrays. Twelve total microarray platforms, including seven spotted resident microarray platforms, three commercial resident platforms and two standard platforms were used (**Figs. 1** and **4**). Seven laboratories used the following four types of spotted resident platforms: (i) a spotted cDNA microarray with target cDNAs obtained from TIGR (laboratory 1); (ii) four different spotted cDNA microarrays with target cDNAs obtained from the National Institute of Aging (NIA) mouse clone sets (laboratories 2, 3, 5 and 7); (iii) a spotted oligonucleotide microarray with target oligonucleotides obtained from Operon (laboratory 4); and (iv) a spotted oligonucleotide microarray with target oligonucleotides purchased from Compugen (laboratory 6). Three laboratories used the following commercial resident platforms: (i) a long oligonucleotide Agilent Mouse Development (MD) microarray, (laboratories 2 and 3); (ii) a short oligonucleotide Affymetrix microarray (laboratories 2 and 3); and (iii) a long oligonucleotide Amersham microarray (laboratory 7). Two laboratories used the following standard platforms: (i) the standard spotted array, made in laboratory 1, by depositing 70-mer oligonucleotides (Operon) representing 18,000 unique mouse transcripts onto poly(L-lysine)-coated slides using a GeneMachines OmniGrid Arrayer. Control spots corresponding to the 10-gene *Arabidopsis thaliana* set (http://pga.tigr.org) were randomly spotted throughout the microarray; and (ii) the standard commercial array

that comprised *in situ* synthesized 60-mer oligonucleotides representing ~20,000 mouse transcripts, designed through collaboration between the Toxicogenomics Research Consortium and Agilent Technologies.

RNA labeling and hybridization. RNA labeling and hybridization procedures used with the standard spotted array and non-commercial resident arrays (datasets A, B and C) are described in **Supplementary Methods**. Standard protocols used with the standard commercial array are described in **Supplementary Methods**. For the commercial resident arrays, individual laboratories performed labeling and hybridization according to the manufacturer's recommendations.

Microarray scanning and image analysis. Scanning and image analysis methods used by the individual laboratories for the standard spotted array (datasets A and B) and resident arrays are described in **Supplementary Methods**. For dataset C, raw image files for the standard spotted array were reanalyzed using Axon GenePix Pro v4.1.1.28 using uniform image extraction parameters (**Supplementary Methods**). Standardized protocols for microarray scanning and image analysis used by all laboratories for the standard commercial arrays (dataset D) are described in **Supplementary Methods**.

Data preprocessing. Non-normalized intensity measurements obtained from the standardized image processing protocol were used to generate four normalized datasets by applying: (i) global intensity normalization, (ii) global intensity normalization with background adjustment, (iii) Lowess normalization with background adjustment applied to a $\log_2$-ratio versus $\log_2$–geometric-mean intensity (R-I or M-A plot), and (iv) Lowess normalization without background subtraction applied to an R-I plot[20,21] (**Supplementary Methods**).

Statistical methods. To examine data reproducibility, Pearson correlation coefficients were calculated between background-corrected $\log_2$ intensity values for all nucleotide sequences represented on all microarrays. Transcripts represented across all microarray platforms were identified by mapping to the NIA mouse gene index (**Supplementary Methods**). When there was more than one correlation coefficient in a comparison, a median of the relevant correlations was presented. To assess the contributions of different potential sources of variability, an ANOVA mixed model[16] was fitted for each of the 502 genes represented on all the platforms (**Supplementary Methods**). For each laboratory, a list of statistically significant up- or downregulated genes was generated based on a prespecified false discovery rate of 0.05. This rate was calculated in a step-up fashion for the mixed model[22].

File format. For dataset A, microarray images were analyzed at individual laboratories using different (in-house) image analysis software packages (**Supplementary Methods**). Raw image extracted data files were stored in a shared database by extracting data from columns of interest from the image analysis output files. For datasets B, C and D, the output files from extracted images were directly stored on an ftp server in a common file location and flat file format without parsing. The datasets were combined based on Unigene IDs.

Scoring and evaluation of Gene Ontology categories. Methods for scoring and evaluating GO categories are described in **Supplementary Methods**.

Additional methods and information. Tissue extraction and RNA isolation procedures are described in **Supplementary Methods**. Additional material and primary (raw) data are available online (http://dir.niehs.nih.gov/microarray/trc/) and via GEO database (accession number GSE2458).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank J. Quackenbush from The Institute for Genomic Research, L. Hartwell from Fred Hutchinson Cancer Research Center and R. Wolfinger from the SAS Institute for their scientific contributions. We thank K.J. Yost (Science Applications International) and P. Cozart (NIEHS ITSS) for their information technology support. Research support was provided by National Institutes of Environmental Health Sciences grants ES11375, ES11384, ES11387, ES11391 and ES11399, and Contract # N01-ES-25497.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 December 2004; accepted 25 March 2005
Published online at http://www.nature.com/naturemethods/

1. Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427 (2001).
2. Salter, A.H. & Nilsson, K.C. Informatics and multivariate analysis of toxicogenomics data. *Curr. Opin. Drug Discov. Devel.* **6**, 117–122 (2003).
3. Nadon, R. & Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271 (2002).
4. Spruill, S.E., Lu, J., Hardy, S. & Weir, B. Assessing sources of variability in microarray gene expression data. *Biotechniques* **33**, 916–923 (2002).
5. Tan, P.K. *et al.* Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
6. Yang, Y.H. & Speed, T. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3**, 579–588 (2002).
7. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630–631 (2004).
8. Becker, K.G. The sharing of cDNA microarray data. *Nat. Rev. Neurosci.* **2**, 438–440 (2001).
9. Miles, M.F. Microarrays: lost in a storm of data. *Nat. Rev. Neurosci.* **2**, 440–443 (2001).
10. Ball, C.A. *et al.* Standards for microarray data. *Science* **298**, 539 (2002).
11. Campbell, P. Microarray standards at last. *Nature* **418**, 323 (2002).
12. Kim, H. *et al.* Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analyses. *Biotechniques* **33**, 924–930 (2002).
13. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
14. Cronin, M. *et al.* Universal RNA reference material for gene expression. *Clin. Chem.* **50**, 1464–1471 (2004).
15. Kerr, M.K. & Churchill, G.A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201 (2001).
16. Kerr, M.K. Experimental design to make the most of microarray results. *Methods Mol. Biol.* **224**, 137–147 (2003).
17. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
18. Wolfinger, R. *et al.* Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625–637 (2001).
19. Hosack, D.A., Dennis, G., Sherman, B.T., Lane, H.C. & Lempicki, R.A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4**, R60 (2003).
20. Hyduke, D.R., Rohlin, L., Kao, K.C. & Liao, J.C. A software package for cDNA microarray normalization and assessing confidence intervals. *OMICS* **7**, 227–234 (2003).
21. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray Analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a powerful approach to multiple testing. *J. R. Stat. Soc. (Ser A)* **57**, 289 (1995).

Addendum: Standardizing global gene expression analysis between laboratories and across platforms

Members of the Toxicogenomics Research Consortium
Nat. Methods **2**, 351–356 (2005).

The authors were listed as Members of The Toxicogenomics Research Consortium, with a complete list of authors available in a **Supplementary Note** online. A complete list of authors in alphabetical order and their affiliations follows.

Theodore Bammler[1], Richard P Beyer[1], Sanchita Bhattacharya[2], Gary A Boorman[3], Abee Boyles[4], Blair U Bradford[5], Roger E Bumgarner[6], Pierre R Bushel[7], Kabir Chaturvedi[8], Dongseok Choi[9], Michael L Cunningham[3], Shibing Deng[5], Holly K Dressman[4], Rickie D Fannin[7], Fredrico M Farin[1], Jonathan H Freedman[4], Rebecca C Fry[2], Angel Harper[8], Michael C Humble[10], Patrick Hurban[8], Terrance J Kavanagh[1], William K Kaufmann[5], Kathleen F Kerr[11], Li Jing[12], Jodi A Lapidus[9], Michael R Lasarev[13], Jianying Li[7], Yi-Ju Li[4], Edward K Lobenhofer[8], Xinfang Lu[14], Renae L Malek[15], Sean Milton[2], Srinivasa R Nagalla[14], Jean P O'Malley[14], Valerie S Palmer[13], Patrick Pattee[14], Richard S Paules[7], Charles M Perou[5], Ken Phillips[8], Li-Xuan Qin[11], Yang Qiu[8], Sean D Quigley[1], Matthew Rodland[14], Ivan Rusyn[5], Leona D Samson[2], David A Schwartz[4], Yan Shi[5], Jung-Lim Shin[12], Stella O Sieber[7], Susan Slifer[4], Marcy C Speer[4], Peter S Spencer[13], Dean I Sproles[13], James A Swenberg[5], William A Suk[10], Robert C Sullivan[12], Ru Tian[5], Raymond W Tennant[7], Signe A Todd[14], Charles J Tucker[7], Bennett Van Houten[10], Brenda K Weis[10], Shirley Xuan[2] & Helmut Zarbl[12]

[1]Department of Environmental and Occupational Health Sciences and the Center for Ecogenetics and Environmental Health, University of Washington, Box 357234, Seattle, Washington 98195-7234, USA. [2]Biological Engineering Division, Center for Environmental Health Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 56-235 Cambridge, Massachusetts 02139, USA. [3]National Toxicology Program, National Institute of Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [4]Duke University Medical Center, Department of Medicine, DUMC 2629, Room 275 MSRB, Research Drive, Durham, North Carolina 27710, USA. [5]Lineberger Comprehensive Cancer Center, University of North Carolina, CB#7295, Chapel Hill, North Carolina 27599, USA. [6]Department of Microbiology, Box 358070, University of Washington, Seattle, Washington 98195, USA. [7]National Center for Toxicogenomics, National Institute of Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [8]Icoria, Inc., 108 TW Alexander Drive, Building 1A, Research Triangle Park, North Carolina 27709, USA. [9]Division of Biostatistics, School of Public Health and Preventive Medicine, School of Medicine, Oregon Health & Science University, Portland, Oregon 97239, USA. [10]Division of Extramural Research and Training, National Institute for Environmental Health Sciences, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [11]Department of Biostatistics, University of Washington, Box 357234, Seattle, Washington 98195-7234 USA. [12]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave North, Mailstop C1-1015, Seattle, Washington 98109, USA. [13]Center for Research on Occupational and Environmental Toxicology, Oregon Health & Science University, Portland, Oregon 97239, USA. [14]Center for Biomarker Discovery, Department of Pediatrics, Oregon Health & Science University, Portland, Oregon 97239, USA. [15]The Institute for Genomic Research, Rockville, Maryland, USA.

Corrigendum: Multiple-laboratory comparison of microarray platforms

Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martínez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye & Wayne Yu
Nat. Methods **2**, 345–349 (2005).

The GEO accession number for the array data is GSE2521.

Independence and reproducibility across microarray platforms

Jennie E Larkin[1], Bryan C Frank[1], Haralambos Gavras[2], Razvan Sultana[1,3] & John Quackenbush[1,3–5]

Microarrays have been widely used for the analysis of gene expression, but the issue of reproducibility across platforms has yet to be fully resolved. To address this apparent problem, we compared gene expression between two microarray platforms: the short oligonucleotide Affymetrix Mouse Genome 430 2.0 GeneChip and a spotted cDNA array using a mouse model of angiontensin II–induced hypertension. RNA extracted from treated mice was analyzed using Affymetrix and cDNA platforms and then by quantitative RT-PCR (qRT-PCR) for validation of specific genes. For the 11,710 genes present on both arrays, we assessed the relative impact of experimental treatment and platform on measured expression and found that biological treatment had a far greater impact on measured expression than did platform for more than 90% of genes, a result validated by qRT-PCR. In the small number of cases in which platforms yielded discrepant results, qRT-PCR generally did not confirm either set of data, suggesting that sequence-specific effects may make expression predictions difficult to make using any technique.

DNA microarrays have afforded biological research scientists the opportunity to assay patterns of gene expression on a global scale. Although there have been many successful applications of this technology, often with high rates of validation using an alternate technology such as northern blot analysis or qRT-PCR, several published studies have called into question the validity of microarray assays, in part because of observed disparities between results obtained by different groups analyzing similar samples[1–8]. As confident practitioners of spotted cDNA microarray technology, we have often been puzzled by the apparent dichotomy of these competing views. In many instances, it seems that the lack of concordance between microarray platforms designed to assay biologically relevant patterns of expression is a failure not of the platform or the biological system, but rather a reflection of the metrics used to evaluate concordance. Other meta-analyses have focused on examining overlapping lists of significant genes, neglecting the fact that in many instances these were derived from not only different platforms but also using vastly different approaches to data analysis[5,9,10]—an effect we have seen even when analyzing a single dataset generated on a single platform.

Based on our experience with hybridization-based approaches, we decided to test platform dependence in assessing a simple biological system, asking whether platform or treatment was the major factor influencing the patterns of observed gene expression. We chose a model system we have previously studied using cDNA microarrays, the effects of short and long-term angiotensin II exposure on cardiac gene expression in a mouse model of hypertension[11] (although in this analysis we used an independent collection biological replicate RNA samples). We chose to compare treated animals to matched controls using cDNA microarrays and Affymetrix GeneChips; the former because it is a two-color platform with which we have a great deal of experience and the latter as it is a widely used commercial oligonucleotide-based platform.

One challenge in designing this experiment was due to the difference in the way that data are collected on these two platforms. Whereas individual Affymetrix arrays are used to assess single biological samples, for cDNA arrays typically two RNA samples are cohybridized: a treated sample and a reference or control sample. Consequently, for the cDNA array assays, we chose to use a common reference design in which each experimental RNA sample is cohybridized with a reference RNA pool as this design most closely mimics the Affymetrix approach.

At every step in the process, from the initial amplification of the RNA to the final steps in the analysis, care was taken to treat the biological samples and the data in an identical fashion as to not introduce artifacts; only in the platform-specific stages of the experiment (RNA labeling, hybridization, data extraction and normalization) was a distinction made between the platforms.

The resulting data were transformed to produce a comparison between treated and matched control animals to address a simple biological question: what is the difference in response to elevated levels of angiotensin II as the length of the exposure increases? In the context of comparing platforms, the question then becomes: given a biological question evaluated on two different microarray platforms, are there platform-specific differences that mask the underlying biological response? The answer to this second question is no, but this answer is qualified by some minor effects we

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [2]Boston University Medical Center, 715 Albany Street, Boston, Massachusetts 02118, USA. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [4]Department of Biochemistry, The George Washington University, Washington, DC 20037, USA. [5]Department of Statistics, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, Maryland 21205, USA. Correspondence should be addressed to J.Q. (johnq@jimmy.harvard.edu).

PUBLISHED ONLINE 21 APRIL 2005; DOI:10.1038/NMETH757

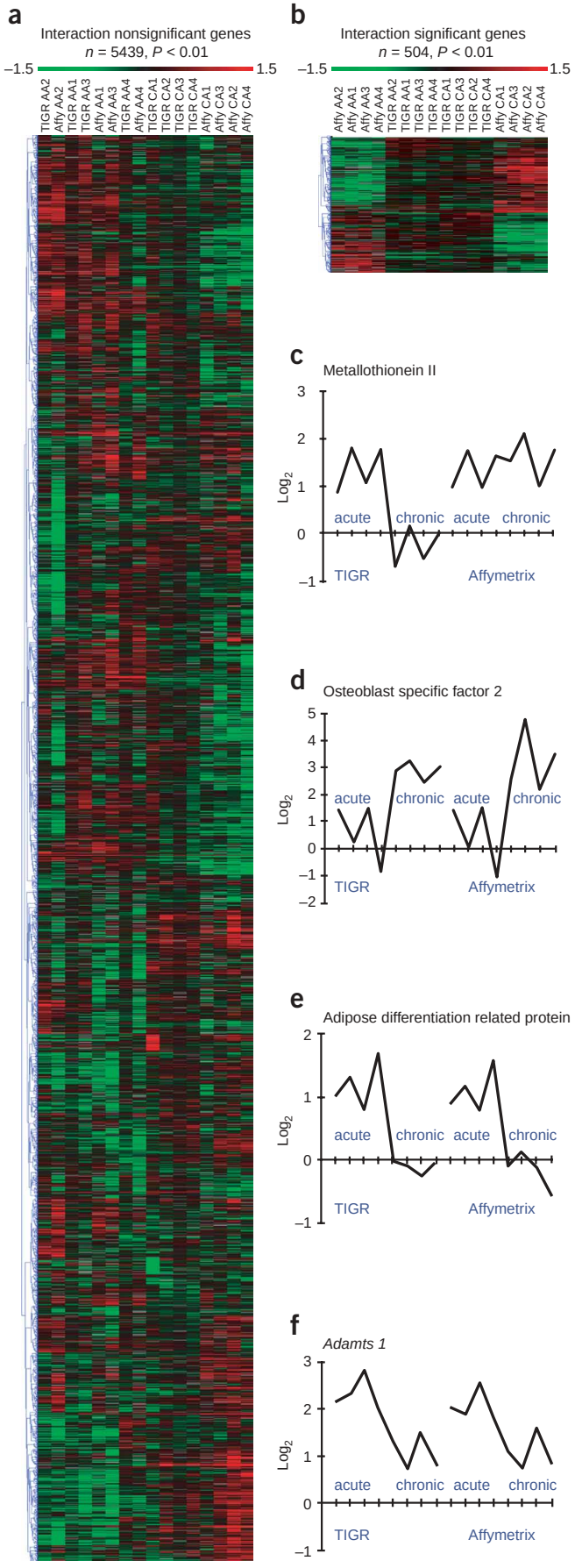


Figure 1 | HCL and gene profiles of two-factor ANOVA results comparing Affymetrix PM-MM to TIGR cDNA microarray data. (**a**,**b**) Heat map and dendrogram representations of genes that are nonsignificant (**a**) and significant (**b**) for platform-specific effects; here each row represents a gene and each column represents a particular sample. (**c**–**f**) Also shown are examples of gene expression profiles showing the expression level for particular genes in each sample measured on spotted cDNA and Affymetrix microarrays for interaction significant (**c**) and interaction nonsignificant (**d**–**f**) genes. The profiles on the left of each plot represent expression levels for individual samples measures on cDNA arrays, and those on the right represent measured levels in the corresponding samples assayed on Affymetrix GeneChips.

observed that point to challenges inherent in using any hybridization based assay.

RESULTS

Study design

Previous experience with DNA microarray assays led us to believe that in a test of a biological system, biological differences between samples should be more significant than platform-specific effects in assaying gene expression. Our previous work, from a large number of studies in diverse species, indicated that approximately 90% of genes identified as 'significant' in microarray assays are ultimately confirmed by an alternate technique such as qRT-PCR. The biological system we chose to analyze here was one we had previously studied, the effects of short- and long-term exposure to angiotensin II (ref. 11). Ten-week-old C57BL/6J male mice were treated with either angiotensin II or saline for either 24 h (acute exposure) or 14 d (chronic exposure), at which time heart tissue was collected for RNA extraction. As the quantities of RNA available for this study were limited, total RNA was amplified using a modification of the Eberwine protocol[12,13], producing antisense cRNA. Amplification of 2.0 μg of total RNA resulted in 47.0 ± 4.3 μg cRNA. The cRNA was then processed separately for hybridization on TIGR 25K cDNA arrays and Affymetrix Mouse Genome 430 2.0 GeneChip.

TIGR cDNA microarray procedure

For cDNA arrays, fluorescently-labeled cDNA targets were prepared from cRNA; the labeled cDNAs were purified, combined as appropriate, and hybridized to the cDNA arrays constructed[14] using 27,010 cDNA clones from the NIA 15k and BMAP mouse cDNA clone sets; these arrays contain approximately 22,000 unique transcripts. All samples were hybridized in duplicate with dye-reversal replicates, against amplified cRNA prepared from the Universal Mouse Reference RNA (Stratagene). Hybridization data were saved as 16-bit .tiff image files and expression data were extracted using TIGR Spotfinder[15]. Data were consistent across biological replicates (RNA derived from independent animals hybridized to independent arrays) and technical replicates (individual RNA samples hybridized to multiple independent arrays), with 87.7 ± 0.7% 'good' spots identified by SpotFinder quality control parameters. Prior to data analyses, signals were normalized using a locally weighted scatterplot-smoothing regression (LOWESS) algorithm[16] implemented in MIDAS[15] followed by standard deviation regularization between array subgrids, and dye-flip consistency checking.

To facilitate comparisons between platforms, the expression data, measured relative to the Stratagene Reference RNA, were used to

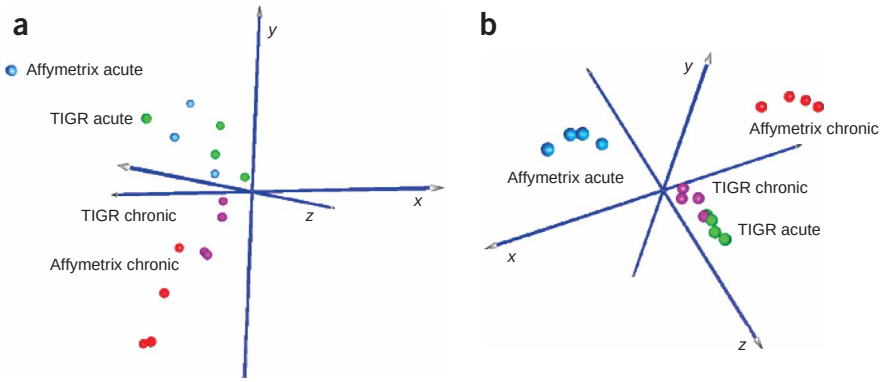


Figure 2 | PCA of microarray data from TIGR cDNA array and Affymetrix mouse GeneChip arrays. (**a**,**b**) Three-dimensional plots show the relative relationship between samples based on 5,144 genes whose expression was found to be platform independent (**a**) and 514 genes found to exhibit platform-specific effects (**b**). In each plot, samples are represented as spheres and each is labeled by platform and treatment (acute or chronic angiotensin II treatment relative respective controls). Data from acute angiotensin II–treated samples on the Affymetrix platform are blue and on the TIGR platform are green. Data from chronic angiotensin II–treated samples on the Affymetrix platform are red and on the TIGR platform are pink. PCA of genes that do not differ statistically between platforms (**a**) show that biological treatment (acute or chronic) groups data more strongly than does platform, with acute samples above the *x*-*z* plane and chronic samples below. In contrast, PCA of discordant genes shows strongly divergent grouping of acute and chronic samples profiled on Affymetrix arrays from each other and from samples profiled on the TIGR cDNA arrays.

calculate the $\log_2$ of gene expression in treated animals relative to the average expression level for the appropriate saline control. This approach had the added benefit of providing robust values for each array element, incorporating both biological and technical replicates. Consistency of expression across biological replicates within each of the two saline control groups was assessed by using the same procedure to compare each individual measurement to the mean of the appropriate group. All subsequent analyses were performed on these normalized datasets of biologically relevant measures, using only those 24,759 array elements for which detectable hybridization signals were available for more than 50% of the hybridization assays.

Affymetrix GeneChips procedure

For Affymetrix GeneChips, antisense cRNA from each biological sample was used as the starting material for the second cycle of the Affymetrix small-sample protocol, to produce biotinylated antisense cRNA; these samples were then chemically fragmented and labeled, then hybridized to Affymetrix Mouse Genome 430 2.0 GeneChips. The two acute saline biological replicate samples were each subjected to three independent amplification and hybridization cycles; as these technical replicates had very high reproducibility, the other 12 biological samples (4 chronic saline, 4 acute angiotensin II, 4 chronic angiotensin II) had a single technical replicate each. Affymetrix GCOS quality control parameters indicated high quality, with consistent hybridizations for all samples; background measurements were nominal (28.1 ± 0.6) and noise was low (1.02 ± 0.06) across all 18 chips. We exported CEL files from the Affymetrix GCOS software and normalized in dChip[17] to the median intensity using two models, the PM-MM model (which subtracts the background mismatch probe intensity from the perfect match probe intensity) and the PM-only model (which ignores the mismatch probes and uses only the perfect match probes for each gene). As was the case with the cDNA arrays, the measured expression levels were $\log_2$ transformed and used to calculate $\log_2$ (angiotensin

Table 1 | Results of qRT-PCR 'validation' for genes that agreed and disagreed across platforms. Correlation coefficients are given between $\log_2$ expression ratios between the Affymetrix GeneChip microarray platform, for both PM-only and PM-MM models, the TIGR mouse cDNA array and qRT-PCR. Correlation table for the gene plasminogen activator inhibitor 1, which showed high correlations of Affymetrix gene expression measurements to qRT-PCR, but TIGR expression measurements were divergent. Correlation tables of microarray measurements to qRT-PCR values for ten genes that had no agreement in expression between Affymetrix and TIGR microarray platforms and for ten genes whose expression was consistent across platforms.

Plasminogen activator inhibitor-1 (*Serpine1*)

	Affy PM-only	Affy PM-MM	TIGR	qRT-PCR
Affy PM-only	1			
Affy PM-MM	0.998	1		
TIGR	0.089	0.087	1	
qRT-PCR	0.893	0.914	0.082	1

Genes that disagree across microarray platforms, without *Serpine1*

	Affy PM-only	Affy PM-MM	TIGR	qRT-PCR
Affy PM-only	1			
Affy PM-MM	0.756	1		
TIGR	0.081	0.044	1	
qRT-PCR	0.386	0.119	0.024	1

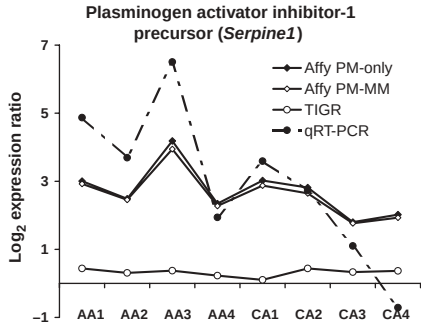


Figure 3 | qRT-PCR validation of microarray results for plasminogen activator inhibitor 1 (*PAI-1*). Expression levels of treated animals relative to the average of the appropriate controls are shown for each of the four platforms, TIGR cDNA, Affymetrix GeneChips with the PM-MM model, for the PM-only model, and qRT-PCR. *PAI-1* was the only example where a discordant result between the array platforms had one platform confirmed by qRT-PCR as can be seen by the correlation between expression profiles.

II–treated/mean saline control) for both acute and chronic exposure to angiotensin II.

Comparative analysis of the two platforms

To ensure standard treatment of both datasets in the analysis, comparisons between platforms were made based on probe-level associations provided by the Resourcerer[18] database. The Affymetrix Mouse Genome 430 2.0 GeneChip contains more than 39,000 probe sets, and the TIGR platform contains greater than 27,000 array elements. These two platforms overlap by 11,710 TIGR Tentative Consensus sequences, but not all of these elements provided useful hybridization data; although the PM-only model provided data for all probe sets, the Affymetrix PM-MM model had a 49% absent call rate and the TIGR platform had a 12% absent call rate. Of the 11,710 tentative consensus sequences annotated using both platforms 10,177 were present in 50% (8/16) of the experiments, but these generally were missing data on one of the two platforms. But 5,853 genes had data in 80% of the 16 combined TIGR and Affymetrix PM-MM experiments, and these data for these genes were used for subsequent analysis.

The expression patterns for the 5,853 'good' genes were subjected to comparative analysis of the platforms. A two-factor ANOVA was used to quantify the impact of platform (Affymetrix mouse GeneChip 430 2.0 or TIGR mouse cDNA array) and experimental treatment (acute or chronic angiotensin II treatment) on measured gene expression values. For most of the genes shared between the two arrays, the gene expression data were remarkably consistent and independent of platform (**Fig. 1**), as biological treatment had a greater impact on gene expression values than did microarray platform. We found that 88% of the genes had no significant effect ($P \leq 0.01$) of microarray platform on the expression values ($n = 5{,}144$). Analysis of these genes indicated that in most instances, the pattern of expression across samples was similar, independent of platform, but that the relative amplitude of the change was greater on one platform than the other. The interaction term in the ANOVA model identifies genes with divergent gene expression responses between the two platforms (**Fig. 1**). These terms were of particular interest as they defined a small subset of genes for which the two platforms gave strongly divergent measurements, both in amplitude and direction of gene expression. Only 9% of genes ($n = 504$) of the 5,853 genes had significant interaction terms in the two-factor ANOVA ($P \leq 0.01$); the majority of these showed a strong transcriptional response on the Affymetrix GeneChip but not the TIGR cDNA array (**Fig. 1**).

Principal components analysis (PCA) is used to reduce the dimensionality of multidimensional datasets. PCA was performed on the two Affymetrix data models (PM-MM and PM-only) and on the TIGR cDNA microarray data, to determine whether experiments clustered primarily by platform or by experimental treatment. The primary principle component accounted for 32% of the variation in the data and differentiated between acute and chronic angiotensin II treatments (**Fig. 2**). The second and third principle components accounted for 28% and 11% of the variation

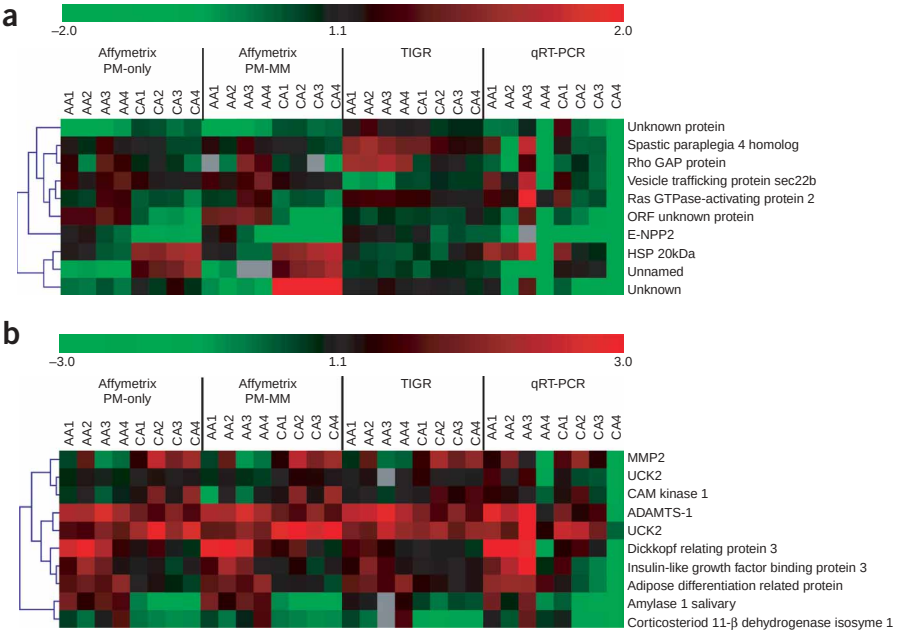


Figure 4 | Hierarchical clustering of expression profiles for genes selected for validation. Shown are heat map representations for measured expression on Affymetrix arrays for the PM-only and PM-MM models, cDNA arrays (TIGR mouse array), and qRT-PCR. (**a**) For the ten genes that had no correlation between array platforms, qRT-PCR had poor correlation with all hybridization-based assays. (**b**) In contrast, for gene that had good correlation between arrays, qRT-PCR also correlated well with all measured expression levels. All data are presented as $\log_2$ (angiotensin II–treated/mean saline control).

in the data and differentiated between biological replicates within each treatment and platform differences. Biological replicates were more tightly clustered in the acute angiotensin II samples than in the chronic angiotensin II samples.

Validation of measurements for shared expression profiles

We used qRT-PCR to validate gene expression for ten genes that shared similar expression profiles across both platforms, representing the group of genes that had nonsignificant interaction terms in the 2-factor ANOVA. We also performed qRT-PCR on 11 genes for which there was a significant interaction term; these genes had disparate expression profiles across the two microarray platforms. Our goal was to use qRT-PCR to identify which platform gave the more accurate result when the data from the two platforms were discrepant. As noted previously, all gene expression values, whether derived from microarray or from qRT-PCR, were represented as the $\log_2$-transformed ratios of the experimental (angiotensin II–treated) gene expression relative to the mean of the time-matched saline-control values. The expression vectors for each gene on each of the four platforms was recorded and pair-wise Pearson correlation coefficients were calculated to assess the degree of concordance between platforms (**Table 1**). The list of individual genes assayed and the primers used for qRT-PCR are available (**Supplementary Methods** online and **Supplementary Table 1** online).

For the ten genes that shared similar expression profiles across Affymetrix and TIGR microarrays, there was strong concordance between Affymetrix and TIGR values (0.81 for PM-MM and 0.85 for PM-only), as was expected. Correlations were very tight (0.98) between the two Affymetrix data models, PM-only and PM-MM, for these ten genes. When the microarray platforms gave consistent results, qRT-PCR also had a robust correlation between both platforms, with correlations between 0.61 and 0.67 (**Figs. 3** and **4**).

For the eleven genes with disparate profiles between platforms, only one gene, plasminogen activator inhibitor 1 (*Serpine1*, also known as *PAI-1*), gave robust confirmation of one platform over the other; qRT-PCR values for *PAI-1* expression mirrored Affymetrix PM-only and PM-MM values, with correlation coefficients exceeding 0.89. For the remaining ten genes with disparate gene expression profiles, qRT-PCR validated neither platform. This was not the result of poor-quality qRT-PCR runs, as each reaction was run in quadruplicate, with common and disparate genes assayed in the same run. Thus, for the majority of the genes whose profiles disagreed across microarray platforms, qRT-PCR validated neither platform but provided yet a third expression profile.

DISCUSSION

Despite the common perception that gene expression values are not reproducible across platforms[1–8], our analysis of cardiac gene expression yielded consistent results for greater than 90% of genes in common between the Affymetrix GeneChip and TIGR cDNA arrays. qRT-PCR analysis was used to independently verify expression for the genes that had similar expression values in both platforms. There are a variety of factors that may contribute to the reproducibility of the results, and the independence of these results from platform.

The first, and most obvious, reason is that the science and expertise of using microarrays as a reliable research tool and repeatability within any one platform has progressed rapidly over the last five years. Whereas earlier microarray experiments sometimes could not reproduce results between laboratories using the same RNA and same microarray technology[8], each platform (GeneChip and cDNA arrays) has progressed substantially in recent years in reliability and reproducibility. If only one of the two platforms being compared gives consistent, reliable data, then comparisons between the two platforms are meaningless as they cannot give consistent results.

As previously stated[9], it is essential to have a reliable, consistent method of identifying genes on both platforms. Gene expression values be compared effectively only if the genes are accurately identified on both platforms. This can be challenging, as oligonucleotide arrays may be generated from very different information than are EST-based cDNA arrays. In this study, we used TIGR tentative consensus sequences for both platforms. The cross-platform comparisons can only be as good as the gene identification method. The results of this study indicate that tentative consensus sequences reliably identify the vast majority of genes correctly, be they based on short oligonucleotide sequences or long cDNAs generated from ESTs.

The methods used for data handling may also influence the repeatability of gene expression values across platforms[9]. In this study, measurements on both platforms were presented as the $\log_2$ ratio of gene expression in response to angiotensin II treatment relative to the mean value of the matching saline control. One reason expressing gene expression as a relative ratio may give more consistent results across platforms, is that this represents a more biologically meaningful value than intensity measures. Hybridization-based assays rely on a wide range of parameters reflecting the properties of the probe and target molecules, which can make it difficult to compare across platforms. But these factors should be less important when comparing a single RNA species between different conditions within the same platform—the question which most often is the one we want to address in biological assays. Knowing how a transcript level changes in response to a particular stimulus often is the more relevant parameter to use in comparing platforms.

Having optimally pure and consistent starting material may also improve the reliability between platforms. It is possible that in previous studies, some of the difference in results between Affymetrix GeneChip results and cDNA arrays may have been due to the discrepancy in RNA handling: a round of amplification is built into the Affymetrix procedure but not into the cDNA protocol. As initial RNA quantities were limited in this study, all RNA was subjected to a round of RNA amplification, producing antisense mRNA. Thus both Affymetrix and TIGR cDNA arrays used amplified RNA as a starting material, minimizing any difference that amplification versus no amplification may have caused, otherwise[19].

A fraction of the genes examined (8–10%) had different results between the two platforms. The common perception upon encountering this type of inconsistency between platforms is to identify one platform as providing superior and more consistent results than does the other platform. However, when gene expression was verified using qRT-PCR, results of only one of the eleven genes tested supported one microarray platform over the other. One possibility is that these nonverifiable genes may represent splice variants. The qRT-PCR primer is based on the TIGR tentative

consensus sequence, which is derived from the assembly of multiple expressed sequence tags (ESTs) and gene sequences. Because the genome is still imperfectly annotated, the Affymetrix probe set may target one or more variants, the cDNA probe on the spotted array one or more others, and the qRT-PCR probe yet others. If these potential splice variants are differently expressed from each other, it would not be surprising that each platform would measure slightly different patterns of expression. Alternately, gene family members may cross-hybridize with both the cDNAs and Affymetrix probe sets and assaying different combinations of these family members, and qRT-PCR, possibly, only a single member.

To address these questions, we mapped the Affymetrix probe sets and the EST sequences corresponding to the cDNA clones arrayed on the TIGR arrays to both the genome sequence and each other. Not surprisingly, corresponding probes on each platform mapped consistently to the same genomic locus. But when searching the Affymetrix probe sequences against the corresponding TIGR EST sequences, a distinct difference was found between those genes that agreed between platforms versus those that did not. For the genes that had similar expression patterns across platforms, nine out of ten had 100% perfect sequence matches for Affymetrix Probe sequence to TIGR EST sequence. The one that did not, was also borderline nonsignificant ($P = 0.013$), suggesting some level of platform specificity. In contrast, only five of the eleven tentative consensus sequences that disagreed across all three platforms had Affymetrix probes that mapped to the corresponding TIGR EST, supporting the hypothesis that the two platforms were interrogating different sequences for the genes that disagreed across platforms. The alignment of multiple EST sequences to the genome in the region containing the array probes generally suggested multiple splice variants in regions where there was generally a single annotated gene structure in the EnsEMBL database. These data suggest that unannotated splice variants may be the major contributing factor to our observation, but without experimental validation for the gene structure it is difficult to claim that a particular platform maps to one splice variant or another.

This study demonstrates that microarray measurement of gene expression and RNA abundance can be a robust method, providing comparable results from different platforms and validates the findings of a recent, related report[20] that demonstrated consistency between laboratories and platforms, provided a consistent analytic approach. But this requires not only careful attention to the experimental details surrounding data collection and analysis, but consistent gene annotation and reliable means of assessing the quality of each experimental assay. If careful attention is paid to these elements, our data indicate that for the majority of genes expression is independent of platform in the sense that biological effects are greater than platform effects.

In doing such analyses, researchers should carefully consider the methods used to compare microarray results as these can have a profound effect on the conclusions that are ultimately derived. In this study we used biological end-points, that is, changes in expression levels in treated animals relative to saline controls, rather than arbitrary measurements that were based on technology. Multiple comparative techniques (two-factor ANOVA, principle components analysis, hierarchical clustering) all gave similar results, affirming our conclusion that microarrays can produce reliable, consistent data that are largely independent of platform. As public databases of microarray experiments (GEO and ArrayExpress) continue to rapidly accumulate expression data, the results presented here should provide some level of confidence that high-quality microarray results can provide a valuable resource for meta-analysis directed at uncovering biological phenomena.

METHODS

TIGR cDNA microarray data analysis. Gene expression levels, were a measure of the $\log_2$ of expression in experimental samples relative to those in the Stratagene Mouse Universal Reference RNA were normalized using local Lowess and filtered to eliminate inconsistent data amongst replicates. Log_2 (angiotensin II–treated/Saline control), mean $\log_2$ values for each array element were determined for both the acute saline ($n = 2$) and chronic saline ($n = 4$) control treatments. The appropriate mean saline control $\log_2$ value was then subtracted from the associated $\log_2$-transformed acute angiotensin II–treated samples ($n = 4$) or chronic angiotensin II samples ($n = 4$), as $\log_2$ (angiotensin II–treated/mean saline control) = $\log_2$ (angiotensin II–treated) – $\log_2$ (mean saline control).

Affymetrix Genechip data analysis. We exported .cel files from Affymetrix GCOS software and normalized in dChip[17] to the median intensity using two models, the PM-MM model and the PM-only model. Gene expression values were then $\log_2$ transformed. For comparison with the biological measures on the TIGR cDNA arrays, mean values for each probe set were calculated for both the acute saline ($n = 6$) and chronic saline ($n = 4$) control groups. The normalized, $\log_2$-transformed experimental values from the acute and chronic angiotensin II treatments were calculated to express $\log_2$ (angiotensin II–treated/mean saline control). In the PM-MM model, only genes with present calls ($51.3 \pm 0.6\%$) were included in subsequent data analysis, resulting in 22,212 probe sets from the PM-MM model. The PM-only model provided data for all probe sets on the array.

Data comparison. All microarray data were represented as $\log_2$ (angiotensin II–treated/mean saline control) for comparison across platforms. To ensure standard treatment of both datasets in analysis, all Affymetrix probe sets and TIGR cDNA clones were mapped to TIGR Mouse Gene Index tentative consensus sequences using RESOURCERER[18] (http://www.tigr.org/tdb/tgi). In instances when a tentative consensus sequence was represented by two or more probes on the array, which occurred on both TIGR and Affymetrix platforms, the mean of the $\log_2$ ratios of gene expression for that gene was calculated in each experiment. All functional analyses were based on tentative consensus sequence assignments; GO terms were mapped directly to TIGR tentative consensus sequences, whereas KEGG and GenMAPP pathways were mapped to tentative consensus sequences via LocusLink identifiers. Expression data from all assays on all platforms is available via ArrayExpress (accession numbers E-TIGR-121, E-TIGR-122, A-TIGR-5 and A-AFFY-45).

Real-time RT-PCR. Validation of gene expression levels performed using SYBR Green assays performed on the ABI Prism 7900HT Sequence Detection System as described in **Supplementary Methods**. Expression for each gene was determined as the ratio of the $\log_2$ (angiotensin II–treated/mean saline control) = $\log_2$ (angiotensin II–treated) – $\log_2$ (mean saline control) for comparison

with array data. Correlation coefficients between platforms were calculated using Excel (Microsoft).

Additional Methods. Details of all experimental protocols, including animal and tissue handling procedures, RNA extraction and labeling, hybridization, data extraction and data analysis, are available in the **Supplementary Methods** online.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

The authors wish to thank F. Pollock of Affymetrix, Inc. for providing the mouse GeneChips used in this study. Thanks also to N. Bhagabati and J. Braisted for valuable discussions. This work was supported by grants U01 HL66580-01 (J.Q.), R33 HL73712 (J.Q.), and U01 HL66617-01 (H.G.) from the National Institutes of Health.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 2 December 2004; accepted 25 March 2005
Published online at http://www.nature.com/naturemethods/

1. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
2. Shippy, R. *et al.* Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61 (2004).
3. Yauk, C.L., Berndt, M.L., Williams, A. & Douglas, G.R. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32**, e124 (2004).
4. Park, P.J. *et al.* Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J. Biotechnol.* **112**, 225–245 (2004).
5. Mah, N. *et al.* A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics* **16**, 361–370 (2004).
6. Rogojina, A.T., Orr, W.E., Song, B.K. & Geisert, E.E., Jr. Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol. Vis.* **9**, 482–496 (2003).
7. Maitra, A. *et al.* Multicomponent analysis of the pancreatic adenocarcinoma progression model using a pancreatic intraepithelial neoplasia tissue microarray. *Mod. Pathol.* **16**, 902–912 (2003).
8. Ulrich, R.G., Rockett, J.C., Gibson, G.G. & Pettit, S.D. Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. *Environ. Health Perspect.* **112**, 423–427 (2004).
9. Jarvinen, A.-K. *et al.* Are data from different gene expression microarray platforms comparable? *Genomics* **83**, 1164–1168 (2004).
10. Tan, P.K. *et al.* Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
11. Larkin, J.E. *et al.* Cardiac transcriptional response to acute and chronic angiotensin II treatments. *Physiol. Genomics* **18**, 152–166 (2004).
12. Phillips, J. & Eberwine, J.H. Antisense RNA Amplification: A Linear Amplification Method for Analyzing the mRNA Population from Single Living Cells. *Methods* **10**, 283–288 (1996).
13. Marko, N.F., Frank, B., Quackenbush, J. & Lee, N.H. A robust method for the amplification of RNA in the sense orientation. *BMC Genomics* **6**, 27 (2005).
14. Hegde, P. *et al.* A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548–550, 552–544, 556 (2000).
15. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
16. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
17. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
18. Tsai, J. *et al.* Resourcerer: a database for annotating and linking microarray resources within and across species. *Genome Biology* **2**, software0002.0001–0002.0004 (2001).
19. Park, P.J. *et al.* Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J. Biotechnol.* **112**, 225–245 (2004).
20. Irizarry, R.A. *et al.* Multiple lab comparison of microarray platforms. *Nat. Methods* **2**, 345–349 (2004).

THE MAGIC OF MICROARRAYS

Research tools known as DNA microarrays are already clarifying the molecular roots of health and disease and speeding drug discovery. They could also hasten the day when custom-tailored treatment plans replace a one-size-fits-all approach to health care

BY STEPHEN H. FRIEND AND ROLAND B. STOUGHTON

DOT PATTERNS EMERGE when DNA microarrays analyze tissue samples. Individual differences in those patterns could one day help doctors match treatments to the unique needs of each patient.

MOST PEOPLE STRICKEN with a cancer called diffuse large B cell lymphoma initially respond well to standard therapy. Yet in more than half of cases, the cancer soon roars back lethally. Physicians have long assumed that the reason some individuals succumb quickly while others do well is that the disease actually comes in different forms caused by distinct molecular abnormalities. But until two years ago, investigators had no way to spot the patients who had the most virulent version and thus needed to consider the riskiest, most intensive treatment.

Then a remarkable tool known as a DNA microarray, or DNA chip, broke the impasse. It enabled a team of researchers from the National Institutes of Health, Stanford University and elsewhere to distinguish between known long- and short-term survivors based on differences in the overall pattern of activity exhibited by

hundreds of genes in their malignant cells at the time of diagnosis. That achievement should lead to a diagnostic test able to identify the patients in greatest danger.

DNA microarrays, first introduced commercially in 1996, are now mainstays of drug discovery research, and more than 20 companies sell them or the instruments or software needed to interpret the information they provide. The devices are also beginning to revolutionize how scientists explore the operation of normal cells in the body and the molecular aberrations that underlie medical disorders. The tools promise as well to pave the way for faster, more accurate diagnoses of many conditions and to help doctors personalize medical care—that is, tailor therapies to the exact form of disease in each person and select the drugs likely to work best, with the mildest side effects, in those individuals.

Tiny Troupers

THE ARRAYS COME IN several varieties, but all assess the composition of genetic material in a tissue sample, and all consist of a lawn of single-stranded DNA molecules (probes) that are tethered to a wafer often no bigger than a thumbprint. These chips also capitalize on a very handy property of DNA: complementary base pairing.

DNA is the material that forms the more than 30,000 genes in human cells—the sequences of code that constitute the blueprints for proteins. It is built from four building blocks, usually referred to by the first letter of their distinguishing chemical bases: A, C, G and T. The base A in one strand of DNA will pair only with T (A's complement) on another strand, and C will pair only with G.

Hence, if a DNA molecule from a tissue sample binds to a probe having the sequence ATCGGC, an observer will be able to infer that the molecule from the sample has the complementary sequence: TAGCCG. RNA, which is DNA's chemical cousin, also follows a strict base-pairing rule when binding to DNA, so the sequence of any RNA strand that pairs up with DNA on a microarray can be inferred as well.

Complementary base-pairing reactions have been integral to many biological tests for years. But amazingly, DNA microarrays can track tens of thousands of those reactions in parallel on a single chip. Such tracking is possible because each kind of probe—be it a gene or a shorter sequence of code—sits at an assigned spot within a checkerboardlike grid on the chip and because the DNA or RNA molecules that get poured over the array carry a fluorescent tag or other label that can be detected by a scanner. Once a chip has been scanned, a computer converts the raw data into a color-coded readout.

Scientists rely on DNA microarrays for two very different purposes. So-called genotype applications compare the DNA on a chip with DNA in a tissue sample to determine which genes are in the sample or to decipher the order of code letters in as yet unsequenced strings of DNA. Frequently, however, investigators these days use the devices to assess not merely the presence or sequence of genes in a sample but the expression, or activity level, of those genes. A gene is said to be expressed when it is transcribed into messenger RNA (mRNA) and translated into protein. Messenger RNA molecules are the mobile transcripts of genes and serve as the templates for protein synthesis.

Gene Hunters

RESEARCHERS have employed the genotype approach to compare the genes in different organisms (to find clues to the evolutionary history of the organisms, for example) and to compare the genes in tumors with those in normal tissues (to uncover subtle differences in gene composition or number). One day gene comparisons performed on DNA chips could prove valuable in medical practice as well.

Carefully designed arrays could, for instance, announce the precise cause of infection in a patient whose flulike symptoms (such as aches, high fever and breathing difficulty) do not point to one clear culprit. A surface could be arrayed with DNA representing genes that occur only in selected disease-causing agents, and a medical laboratory could extract and label DNA from a sample of infected tissue (perhaps drawn from the person's nasal passages). Binding of the patient's DNA to some gene sequence on the chip would indicate which of the agents was at fault. Similarly, chips now being developed could signal that bioterrorists have released specific types of anthrax or other exotic germs into a community.

For better or worse, gene-detecting microarrays could also identify an individual's genetic propensity to a host of disorders. Most genetic differences in people probably take the form of single nucleotide polymorphisms, or SNPs (pronounced "snips"), in which a single DNA letter substitutes for another. A chip bearing illness-linked gene variants could be constructed to reveal an individual's SNPs and thus predict the person's likelihood of acquiring Alzheimer's disease, diabetes, specific cancers and so on. Those people at greatest risk could then receive close

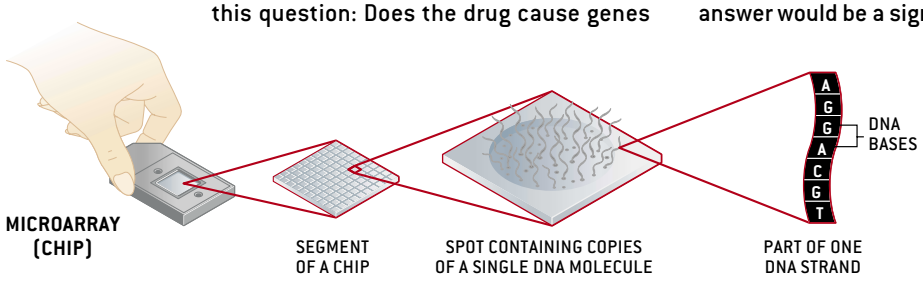
Overview/*Microarrays*

- DNA microarrays, also known as DNA or gene chips, can track tens of thousands of molecular reactions in parallel on a wafer smaller than a microscope slide. The chips can be designed to detect specific genes or to measure gene activity in tissue samples.
- These properties are proving immensely valuable to cell biologists, to scientists who study the roots of cancer and other complex diseases, and to drug researchers. Microarrays are also under study as quick diagnostic and prognostic tools.
- Protein arrays, which have great promise as diagnostic devices and as aids to biological research, are being developed as well.
- The research and diagnostic information provided by DNA chips and protein arrays should eventually help physicians provide highly individualized therapies.

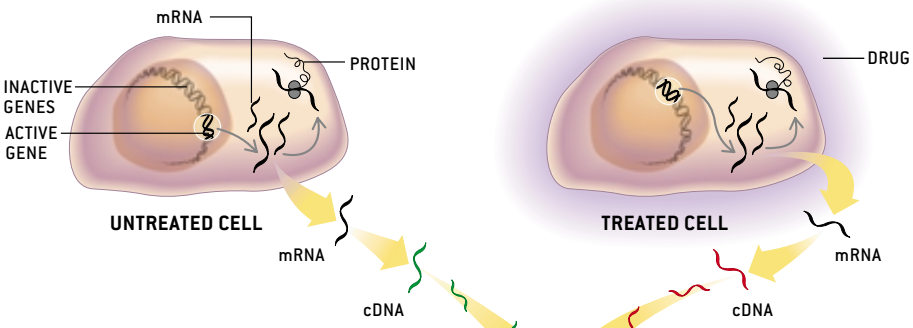
DENNIS GALANTE PHOTOGRAPHY (*preceding pages*)

HOW ARRAYS WORK

TO DETERMINE QUICKLY whether a potential new drug is likely to harm the liver, a researcher could follow the steps below, asking this question: Does the drug cause genes (the blueprints for proteins) in liver cells to alter their activity in ways that are known to cause or reflect liver damage? A "yes" answer would be a sign of trouble.

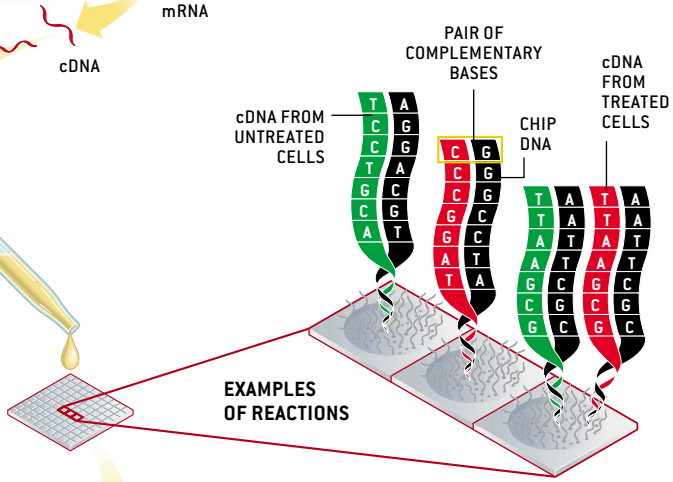


1 Construct or buy a microarray, or chip, containing single-stranded DNA representing thousands of different genes, each assigned to a specified spot on the one-by-three-inch or smaller device. Have every spot include thousands to millions of copies of a DNA strand.

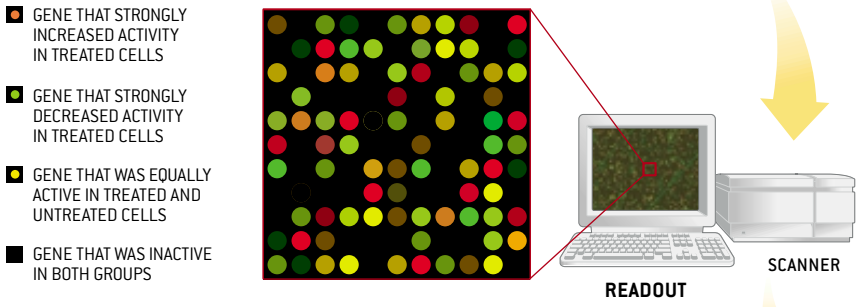


2 Obtain two samples of liver cells; apply the drug to one sample. Then, from each sample, collect molecules of messenger RNA (mRNA)—the mobile copies of genes and the templates for protein synthesis in cells.

3 Transcribe the mRNA into more stable complementary DNA (cDNA) and add fluorescent labels—green to cDNAs derived from untreated cells, red to those from treated cells.

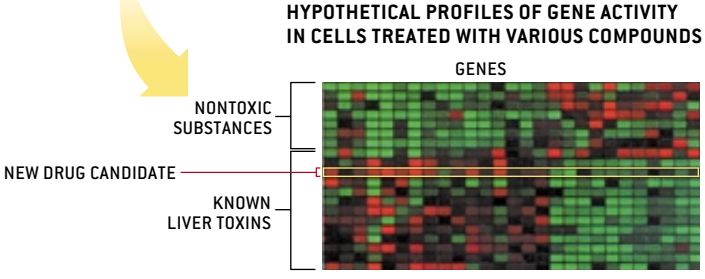


4 Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip (*detail at right*). Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.



5 Put the chip in a scanner. Have a computer calculate the ratio of red to green at each spot (to quantify any changes in gene activity induced by the drug) and generate a color-coded readout.

6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (*right*). Close similarity would indicate that the new candidate was probably toxic as well. In the diagram, each box represents a single gene's response to a compound.



JARED SCHNEIDMAN DESIGN

monitoring, intensive preventive care and early intervention. Whether these kinds of tests would appeal to the public is an open question, though; the downside of such knowledge can be increased anxiety and the potential for discrimination by employers and insurers.

Other valuable information provided by SNP chips would pose no threat to people's mental state, employability or insurability. The gene variants we possess influence how our bodies process the medicines we take, which in turn influences the effectiveness of the drugs and the intensity of their side effects. Chips that highlighted our unique genetic sensitivities would help physicians choose the drugs that work best and pose the fewest dangers in each of us. SNP chips displaying genetic mutations that increase the aggressiveness of tumors might also help pathologists determine whether benign-looking tumors are actually fiercer than they seem based on microscopic analyses. Both types of arrays are already being investigated for use in medical care.

Choice Expressions

AS EXCITING AS such applications are, it is the other major use of arrays—expression profiling—that has increasingly captivated researchers over the past few years. Laboratory workers produce these profiles by measuring the amounts of different mRNAs in a tissue sample. Generally, the more copies of mRNA a cell makes, the more copies of protein it will make, so the quantities of the various mRNAs in a sample can indirectly indicate the types and amounts of proteins present. Proteins are often of interest because they control and carry out most activities in our bodies' cells and tissues. Chips that directly measure protein levels are being developed [*see box on page 52*], but constructing them remains challenging.

By using the genome as a sensor pad to detect activity changes in a cell's various genes, scientists can gain exquisitely detailed "snapshots" of how a cell's functions have been altered by drugs or disease states. At times, knowing the overall on-off pattern of gene activity in a sample can actually be more useful than knowing which particular genes turn on and off in response to some influence. In those cases, as will be seen, the pattern serves as a shorthand "signature" reflecting the molecular state of a sample under some specific condition.

Expression profiling has proved invaluable on many fronts. Cell biologists like it because knowledge of the proteins that predominate after a tissue is exposed to different conditions can provide insight into how the tissue normally compensates for disruptions and what goes wrong when diseases develop.

These scientists are also using expression arrays to learn the functions of genes that have been discovered as a result of the recent sequencing of nearly all the DNA in the nucleus of the human cell. Several techniques that do not involve microarrays can reveal the jobs performed by newly discovered genes (or, more properly, by the proteins those genes encode), but those approaches do not always work well or quickly. In what has come to be called the guilt-by-association application, expression arrays can help fill in the blanks, even in the absence of any prior clues to a gene's role in the body.

This method derives from the awareness that no gene is an island. If genes in a tissue switch on and off together in response to some influence—say, a drug, an infection or an induced gene mutation—workers can surmise that those like-acting genes operate in the same regulatory pathway; that is, the genes work together or in series to induce a cellular response. Investigators can reasonably guess, then, that the jobs of any originally mysterious genes in the group resemble those of genes whose responsibilities are already understood.

Drug Discovery Tools

DRUG RESEARCHERS, too, take advantage of the guilt-by-association method—to discover proteins not previously known to operate in biological pathways involved in diseases. Once those proteins are found, they can be enlisted as targets for the development of new and better medicines.

In one example, Peter S. Linsley, our colleague at Rosetta Inpharmatics, wanted to identify fresh targets for drugs that might combat inflammatory illnesses, in which the immune system perversely damages parts of the body. He therefore asked which genes in white blood cells of the immune system increase and decrease their protein production in parallel with the gene for a protein called interleukin-2 (IL-2), which is strongly implicated in inflammatory disorders.

He got the answer by producing expression profiles for white blood cells exposed to various chemicals and then having a computer run a sophisticated pattern-matching program to pinpoint a set of genes that consistently switched on or off when the IL-2 gene was activated. This set included a gene whose function in the body had not been determined by other means. At about the same time, investigators at the Pasteur Institute in

AN ARRAY OF COMPANIES

The following are just some of the companies that sell or are developing array-related products and services:

DNA MICROARRAYS

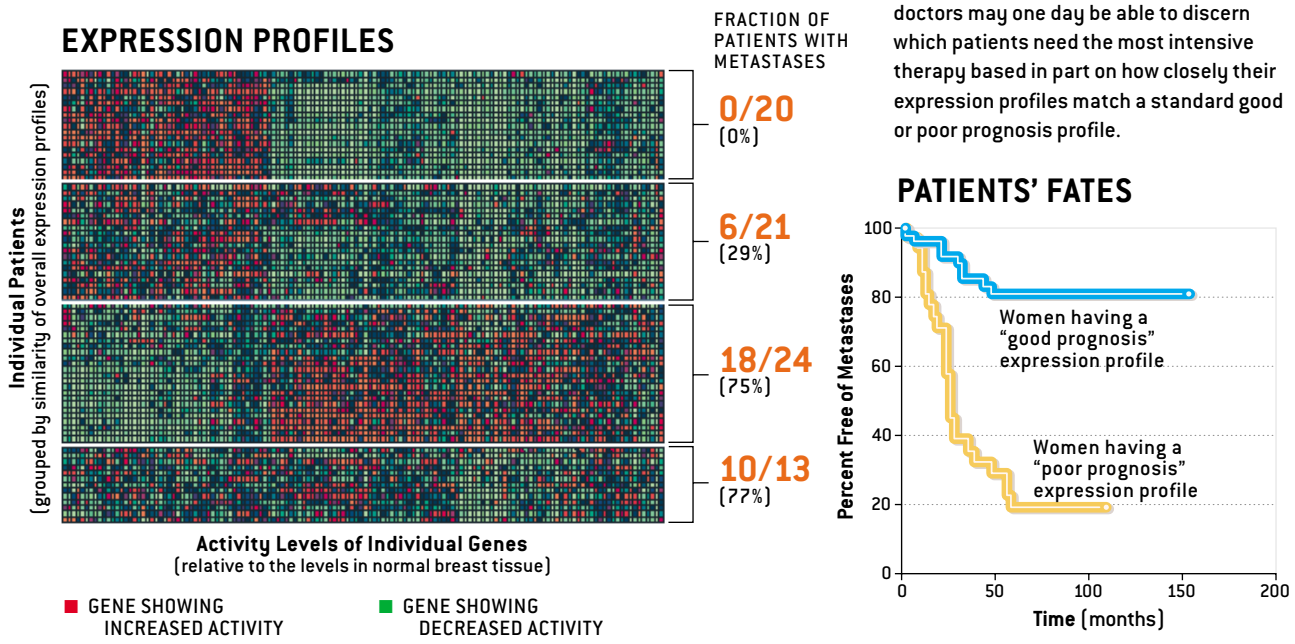
Affymetrix, *Santa Clara, Calif.*
Agilent Technologies, *Palo Alto, Calif.*
Amersham Biosciences, *Piscataway, N.J.*
Axon Instruments, *Union City, Calif.*
BioDiscovery, *Marina del Rey, Calif.*
Clontech, *Palo Alto, Calif.*
Genomic Solutions, *Ann Arbor, Mich.*
Mergen, *San Leandro, Calif.*
Motorola Life Sciences, *Northbrook, Ill.*
Nanogen, *San Diego, Calif.*
Partek, *St. Peters, Mo.*
PerkinElmer, *Boston, Mass.*
Rosetta Inpharmatics, *Kirkland, Wash.*
Spotfire, *Cambridge, Mass.*
Virtek Vision International, *Ontario, Canada*

PROTEIN ARRAYS

Biacore International, *Uppsala, Sweden*
Biosite Diagnostics, *San Diego, Calif.*
Ciphergen, *Fremont, Calif.*
Large Scale Biology, *Germantown, Md.*

PREDICTING CANCER'S COURSE

WORK AT ROSETTA INPHARMATICS and the Netherlands Cancer Institute suggests that microarrays can help distinguish cancer patients with different prognoses. After determining the activity (expression) levels of genes in small, localized breast tumors from young women who were followed for at least five years after surgery, the researchers found that the expression profiles—the overall patterns of activity across a selection of genes in the tumors—differed among the patients (*left*). A mathematical analysis (*right*) then revealed that patients whose expression profiles resembled a "poor prognosis" signature (the average pattern in tumors that metastasized) were much more likely to suffer a quick recurrence than were patients whose profiles resembled a "good prognosis" signature (the typical pattern in tumors that did not spread). If such results are confirmed by others, doctors may one day be able to discern which patients need the most intensive therapy based in part on how closely their expression profiles match a standard good or poor prognosis profile.



Paris independently confirmed, with other methods, that this gene operates in the IL-2 pathway. Together the findings suggest that the protein encoded by the gene could be a good target for anti-inflammatory drugs.

Pharmaceutical scientists use expression profiling in a different way: to pick out—and eliminate—drug candidates that are likely to produce unacceptable side effects. Workers who want to determine whether a given compound could damage the heart, for example, can compile a compendium of expression profiles for heart cells exposed to existing drugs and other chemicals. If they also treat heart cells with the drug candidate under study, they can ask a computer to compare the resulting signature with those in the compendium. A signature matching those produced by substances already known to disrupt cardiac cells would raise a red flag.

A compendium of expression profiles can also help explain why a drug produces particular side effects. A pressing question today, for instance, is why protease inhibitors, which are lifesavers to people infected with HIV (the virus that causes AIDS), can lead to high cholesterol and triglyceride levels in the blood, strange redistributions of body fat, and insulin resistance. Aware that the liver influences the production and breakdown of lipids (the group that includes cholesterol and triglycerides) and of lipid-containing proteins, we and others at Rosetta, in collaboration with Roger G. Ulrich and his team at Abbott Laboratories, decided to see whether one protease inhibitor—ritonavir—produced some of its side effects by acting on the liver.

With an array representing about 25,000 rat genes, we produced expression profiles of rat liver tissue exposed to an assortment of compounds that can be toxic to the liver. After that, we grouped the compounds according to similarities of expression signatures across some 2,400 genes that responded strongly to those substances. Next we delivered ritonavir to rat livers and compared the resulting expression profiles with those generated earlier.

THE AUTHORS

STEPHEN H. FRIEND and *ROLAND B. STOUGHTON* are colleagues at Rosetta Inpharmatics in Kirkland, Wash., which was founded in 1996 to develop molecular profiling methods involving computers and DNA microarray technology. Merck & Co. acquired the company last year. Friend is vice president of basic research at Merck and president of Rosetta. He was a pediatric oncologist and molecular biologist at Harvard University before becoming director of molecular pharmacology at the Fred Hutchinson Cancer Research Center and co-founding Rosetta. Stoughton, who has a Ph.D. in physics, is senior vice president for informatics at Rosetta. Before turning his attention to biotechnology, he worked on developing signal-processing and pattern-recognition tools for geophysics and astrophysics.

DATA DISPLAY BY HONGYUE DAI *Rosetta Inpharmatics*; GRAPH BY SARA CHEN

Protein Arrays—A New Option

by N. Leigh Anderson and Gunars Valkirs

LIKE DNA MICROARRAYS, protein-based chips—which array proteins instead of DNA molecules on a small surface—can measure the levels of proteins in tissues. In fact, they do the job more directly and, some evidence says, more accurately. Protein arrays also stand alone in being able to reveal which of thousands of proteins in a tissue interact with one another.

All these properties make protein arrays quite appealing to biological researchers. But the average person would most likely be intrigued for a different reason. Hope is high that such chips will dramatically expand the number of conditions that doctors can diagnose quickly in their offices.

These devices should be very useful as diagnostic tools in part because, unlike DNA microarrays, they can glean information from blood plasma, which is easy to obtain. Most medical disorders—from infectious diseases to heart or kidney damage—leave identifiable traces in the blood, in the form of secreted or leaked proteins. Moreover, in a single test, the arrays might measure many or all of the proteins known to flag the presence of medical problems. In contrast, standard diagnostic tests detect only one or a few disease-specific proteins at a time.

The design of protein arrays resembles that of DNA chips. Hundreds to thousands of distinct proteins sit (in millions of copies) at specified spots in a grid on a wafer-thin plate. Binding of proteins from a blood sample to proteins on a chip reveals the nature and quantities of the sample proteins.

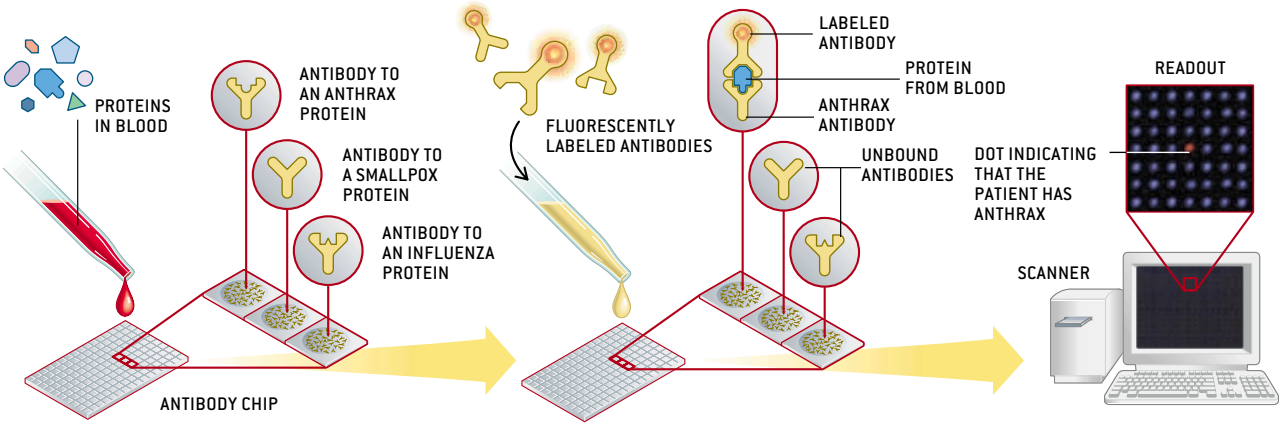
The kinds of proteins displayed on the chips can vary depending on the questions being asked. But the chips closest to commercialization (initially for use by researchers) rely on the remarkable immune system molecules called antibodies—each of which recognizes and binds to one specific protein or, more precisely, to a specific segment of a protein. Some of these antibody chips work by what is called the sandwich method: proteins recognized by a chip get sandwiched between two different antibodies, one that grabs the protein and a second that attaches a fluorescent label to the snagged molecule (*diagram below*).

For antibody-based arrays to deliver fully on their potential for advancing research and diagnostics, scientists will have to topple at least two major impediments. One is the need for techniques that mass-produce many different antibodies at once, and not just any antibodies—those that bind tightly to one target, so as to reveal even small quantities in a sample. This problem is already being surmounted. The second obstacle is more fundamental. Medical science has so far uncovered only dozens of the perhaps thousands of proteins able to signal the presence or progress of a disease. Until chipmakers know which proteins to look for, they will be able to seek only a limited number of disease markers in a tissue sample. Fortunately, droves of investigators are now hunting for new disease-specific proteins. As advances in antibody manufacture and protein discovery converge, they will yield a second generation of protein arrays that could well transform both medical research and clinical practice.

N. Leigh Anderson and Gunars Valkirs collaborate on protein array research. Anderson is chief scientific officer at Large Scale Biology Corporation in Germantown, Md. Valkirs is chief technology officer at Biosite Diagnostics in San Diego, Calif.

A PROTEIN ARRAY IN ACTION

DOCTORS MIGHT ONE DAY use a "sandwich assay" to identify the infectious agent responsible for a patient's illness. Is it a common flu bug or a new, deadly variety? Might the tuberculosis bacterium be at fault—or even anthrax, smallpox or Q fever microorganisms unleashed by bioterrorists? Following the steps below would reveal the answer.



1 Apply blood from a patient to a chip, or array, consisting of antibodies assigned to specific squares on a grid. Each square includes multiple copies of an antibody able to bind to a specific protein from one organism and so represents a distinct disease-causing agent.

2 Apply fluorescently labeled antibodies able to attach to a second site on the proteins recognizable by the antibodies on the chip. If a protein from the blood has bound to the chip, one of these fluorescent antibodies will bind to that protein, enclosing it in an antibody "sandwich."

3 Feed the chip into a scanner to determine which organism is present in the patient's body. In this case, the culprit is shown to be a strain of anthrax.

JARED SCHNEIDMAN DESIGN

Ritonavir, we learned, leads to activation of genes that are usually quieted in response to a well-known lipid-lowering agent; ritonavir also decreases the production of proteins that normally assemble into proteosomes, structures that break down no-longer-useful proteins, including lipid-containing types. These findings suggest that ritonavir raises lipid levels in the liver—and hence in the blood—in part by elevating the liver's synthesis of lipids and inhibiting its breakdown of lipid-containing proteins. Further study of exactly how ritonavir interacts with the lipid- and proteosome-producing pathways will provide ideas for reducing its side effects.

Treatment Tailors

HAVING AN ENLARGED arsenal of drugs, and more drugs with fewer side effects, would be a great outcome of the molecular profiling made possible by DNA array studies. But many physicians are hoping for an even better result: rapid diagnostic tools that would divide patients with similar symptoms into separate groups that would benefit from different treatment plans. As the lymphoma study mentioned at the start of this article demonstrated, cancer specialists in particular desperately need ways to identify patients who require maximally aggressive treatment from the beginning.

Research into breast cancer by our group at Rosetta, working with collaborators from the Netherlands Cancer Institute in Amsterdam, demonstrates how expression arrays can help [*see box on page 49*]. In this case, we wanted to invent a test able to determine which young patients with early-stage breast cancer (with no evidence of cancer in the lymph nodes) need systemic drug therapy to prevent tumor spread (metastasis) after surgery and which do not. Although current guidelines recommend systemic treatment for about 90 percent of these women, a good many of them would probably avoid distant metastases even if they did not have such treatment. Unfortunately, standard tools cannot single out the women at greatest risk.

We began by generating expression profiles for tumors from close to 100 women under age 55 whose clinical course had been followed for more than five years after surgery. We initially worked with a microarray representing 25,000 human genes. In the end, we found that one particular signature produced by about 70 genes strongly indicated that metastases would soon appear. In addition, the opposite pattern was strongly indicative of a good prognosis. Clearly, some tumors are programmed to metastasize before they grow to a size smaller than half a dime, whereas other, larger masses are programmed not to spread.

Our results have to be confirmed by others before expression profiling can become a routine part of breast cancer workups. Within two years, many medical centers will probably begin to test expression profiling as a guide to therapy, not just for breast cancer but for other types as well. Other diseases need improved diagnostic tools, too. Expression profiling might help distinguish subgroups of patients with such disorders as asthma, diabetes or obesity who have special treatment needs. Those applications are now under study.

Before microarrays can live up to their full potential as research and diagnostic tools, several roadblocks have to be toppled. The chips, scanners and other accoutrements remain expensive (engendering "array envy" in many underfunded academics). Presumably, however, costs will drop with time.

Yet even if prices fall, the technologies may prove infeasible, at least initially, for doctors' offices or standard medical laboratories. Few physicians or technicians have the equipment and the skill to prepare tissue samples properly for use with arrays. What is more, to diagnose, say, liver disease based on changes in gene expression in liver cells, a doctor would ideally need to obtain tissue from the liver. But that organ is not readily accessible.

These problems loom large right now but are probably surmountable with ingenuity. At times, for instance, accessible tissues might function as acceptable stand-ins for inaccessible ones. Moreover, in some instances, microarrays themselves may not have to be used; they might provide the research information needed for devising new diagnostic tests, which can then take other forms.

As the operations of cells and the entire body become better understood, physicians will be able to make more precise diagnoses, to offer patients more sophisticated therapies (possibly including gene therapies), and to tailor these interventions to an individual's genetic background and current state of physiological functioning. By the year 2020, health maintenance organizations and their ilk could conceivably keep *in silico* models of the personal molecular states of their subscribers—virtual simulations that could be updated constantly with microarray and other data from doctor visits and with new scientific information about cell biology. Perhaps some subscribers won't like that idea and will forgo a rate discount—and quite possibly the best care—in return for a feeling of privacy. Those who go along with the program, though, will probably delay the effects of aging more successfully and lead healthier lives. SA

MORE TO EXPLORE

The Chipping Forecast. Supplement to *Nature Genetics,* Vol. 21, pages 1–60; January 1999.

Genomics, Gene Expression and DNA Arrays. David Lockhart and Elizabeth Winzeler in *Nature,* Vol. 405, pages 827–836; June 15, 2000.

Experimental Annotation of the Human Genome using Microarray Technology. D. D. Shoemaker et al. in *Nature,* Vol. 409, pages 922–927; February 15, 2001.

Web sites listing links and publications on microarrays can be found at:

http://bioinformatics.phrma.org/microarrays.html

http://industry.ebi.ac.uk/~alan/MicroArray/

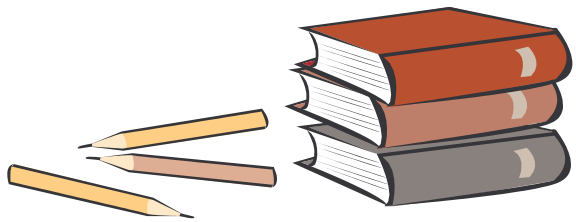
www.rii.com/publications/default.htm

http://ihome.cuhk.edu.hk/~b400559/2001j_mray.html

www.biologie.ens.fr/en/genetiqu/puces/links.html#news

COLLABORATIVE PROGRAMS

Genome Consortium for Active Teaching (GCAT)



A. Malcolm Campbell,[1,2*†] Todd T. Eckdahl,[2,4] Edison Fowlks,[2,5] Laurie J. Heyer,[2,3] Laura L. Mays Hoopes,[2,6] Mary Lee Ledbetter,[2,7] Anne G. Rosenwald[2,8]

A supportive network of scientists and faculty brings sophisticated microarray experiments to the undergraduate lab and classroom.

Biological research has been transformed in recent years by substantial advances in efficient data accumulation. The transcription output for every gene in a genome now can be measured in an afternoon; before it might have taken years. However, the recent advances in technology have yet to be incorporated into many biology classrooms (*1*). Most undergraduates are taught the same way their instructors were taught, which seldom reflects leading-edge research practices. Training faculty in the latest research methods is not well supported on most campuses (*2*). Worse yet, when students with outdated undergraduate science experiences become primary and secondary school teachers, they condemn future generations to inadequate preparation for college. Today's teachers may also neglect the more quantitative aspects and increased interdisciplinary involvement of modern biology (*3*–*5*). Educational options that reflect quantitative, interdisciplinary, and technological trends would provide students with experiences that mirror today's scholarship.

We have developed the Genome Consortium for Active Teaching (GCAT) (*6*) to engage undergraduates in genomics experimental design and data analysis. GCAT faculty use DNA microarrays to bring the excitement of interdisciplinary research to students. Students discover the importance of quantitative data analysis, and the faculty are reinvigorated by the opportunity to learn new technology.

GCAT in the lab. Undergraduates prepare samples and scan microarrays as part of their research at Davidson College.

CREDIT: DAVIDSON COLLEGE

Origins of GCAT

GCAT was formed in 1999 with the intent of bringing genomics into undergraduate curricula, primarily through student research (*7*, *8*). Leading scientists donated materials and equipment. Undergraduates designed and performed experiments (see photograph above), mailed their microarrays for scanning, and then downloaded and analyzed their data (*9*).

Two limiting factors, long-term scanner access and a growing appetite for microarrays, were addressed by grant support and further donations from scientists (*10*–*12*). GCAT thus grew in size and expertise. GCAT supports free access to information and results through its Web site (*6*) and a listserv of more than 200 subscribers.

GCAT projects replaced student laboratory methods less prevalent in today's research, such as cloning and sequencing a gene and Northern blotting.

Rapid Growth

GCAT is committed to enabling any institution to adopt the use of microarrays in its undergraduate curriculum at affordable prices. To date, about 5000 undergraduates from 120 schools have used about 3400 microarrays. For the 2005–2006 academic year, GCAT provided more than 750 microarrays of nine plant, animal, and microbial species to students on 64 different campuses (*6*, *9*). Tested protocols and teaching aids are available from GCAT. Continued grant support (*11*) covers the cost of microarrays.

Schools pay a nominal fee to GCAT for microarrays and scanning. Students produce and hybridize their own probes. Other than the scanners, only standard molecular biology equipment is required; the software is free. The summer workshop costs, which are currently covered by grant support, are about $2300 per participant.

The number of interested faculty continues to grow. Although this enthusiasm is more a measure of the importance of the microarray method in molecular biology today than of GCAT itself, it also serves as a testament to GCAT's user-friendly format.

GCAT faculty use the microarrays in various ways. Some analyze existing data sets, such as the yeast diauxic shift data (*13*) that shows how yeast switch from one metabolic route to another. Other faculty members offer courses in which students collect their own microarray data. Students have studied the effects of environmental conditions on growth, aging in yeast, chromatin structure, and the cellular side effects of chemotherapy (*6*). Microarrays offer a view of the connections between different pathways in a cell in ways that are hidden by many other methods. For example, one student project looked for expression changes in DNA replication mutants and found cell wall assembly changes, thus linking cytokinesis to mitosis.

Dissemination Through Faculty Development

GCAT has sponsored data generation (wet lab) and data analysis (dry lab) workshops in various settings (*14*). Wet and dry lab sessions work best when they run 2 and 3 days, respectively. Participants learn data analysis using MAGIC Tool freeware (*15*). MAGIC Tool works on any computer platform and is designed to enhance student understanding of

[1]Department of Biology, [2]Genome Consortium for Active Teaching, [3]Department of Mathematics, Davidson College, Davidson, NC 28035, USA. [4]Department of Biology, Missouri Western State University, St. Joseph, MO 64507, USA. [5]Department of Biology, Hampton University, Hampton, VA 23668, USA. [6]Department of Biology, Pomona College, Claremont, CA 91711, USA. [7]Department of Biology, College of the Holy Cross, Worcester, MA 01610–2395, USA. [8]Department of Biology, Georgetown University, Washington, DC 20057, USA.

*Authors are listed alphabetically.

†Author for correspondence. E-mail: macampbell@davidson.edu

microarray and data analysis techniques.

In 2004, 35 faculty attended NSF-funded data analysis or combined data generation and analysis workshops at Georgetown University. Assessments demonstrated that combined training had a greater impact on undergraduate courses than the analysis workshop alone. The 23 who participated in the combined workshops reported that 800 undergraduates subsequently used microarrays (~35 students per teacher). In 2005, 64 faculty received microarrays. With similar rates, the microarrays might reach as many as 2200 undergraduates.

Diversity

Historically black colleges and universities (HBCUs) are often left behind the technology curve. Two-thirds of attendees at the 2005 GCAT workshop at Morehouse College represented schools with substantial populations of underrepresented students, including African Americans, Native Americans, Hispanics, and nontraditional students attending community colleges. These populations are critical for diversifying the population of scientists in the United States. Faculty from biology, chemistry, mathematics, and computer science have attended GCAT workshops. GCAT activities have attracted diverse populations of students: 21% of GCAT students are non-Caucasian, 64% are female, 21% are majoring in a discipline other than biology, and 44% are interested in pursuing research careers in biology. GCAT implements BIO2010 recommendations (*1*) by teaching genomics through student research, which excites students across disciplines and ethnicities.

Keys to Success

GCAT's success is due to the people involved. The early GCAT faculty took a collective leap of faith by teaching genomics while simultaneously learning it themselves. Today's GCAT users can avoid much of the risk by taking a workshop before beginning with microarray analysis. GCAT faculty demonstrate their dedication by voluntarily leading the consortium's efforts (*16*). Working as a community maximizes efficiency and produces a sense of belonging to a larger effort that transcends a single campus.

Faculty and students participate in assessments of student comprehension, attitudes toward research, and demographic information. Anonymous, open-ended responses from students have been very enthusiastic. Selected comments from students include, "Microarray: GREAT! I am amazed that we can do this! Such an interesting concept yet simple enough to perform" and "What a powerful concept, microarrays. I greatly appreciated the opportunity to use what is quite possibly the most important tool in current analyses of gene expression."

Pre- and posttest results showed that GCAT courses produced significant improvement ($P < 0.001$) in students' abilities to design experiments and interpret data, areas often neglected in traditional teaching laboratories (see table). For example, students learned that whole-pathway changes are more reliable than individual gene changes. Students saw how spot identification must be quantitatively guided and how ratios are more informative than intensities. When faculty explained their learning goals, how they use GCAT resources, and the impact GCAT had on their ability to use microarray technology, they overwhelmingly indicated that they would not be able to do this work without GCAT resources and will continue to participate in GCAT activities.

GCAT Students Participate in Various Aspects of the Scientific Process

85% hybridize probes to microarrays
78% produce cDNA probes
58% analyze their own data
53% design their experiments
25% analyze published microarray data
63% write a paper for course credit
35% present a poster of findings

When participants of the 2004 workshops were surveyed 1 year later, 80% (64% response rate) rated their experiences with the highest category on the survey. Sixty-one percent indicated networking with other faculty was very valuable. Faculty who had attended the combined data generation and analysis workshop altered an average of 1.6 courses to include the new content, whereas those who had attended only the data analysis workshop modified half as many courses (average 0.86). Faculty reported that their students showed an increased interest in mathematics as a result of microarray experiences. Faculty felt their teaching had improved and their classes were more interesting. One faculty member wrote, "… the presentation of this subject makes [students] realize and practice the close interaction biology/genetics has with other fields like mathematics. They enjoyed [being] introduced to a novel genetic technique. They said they can understand better, and can relate their class more to real life…when they watch [news about] health and advances in science." Another faculty member reported, "…many students have come back and said they got jobs or were assigned or allowed to do special projects in graduate schools because of their experience with microarrays."

Future Directions

GCAT wants to reach more faculty, especially at HBCUs, tribal colleges, Hispanic-serving institutions, community colleges, and small institutions. Regional workshops are being developed. GCAT is also working with high school teachers to develop a classroom and laboratory module on DNA microarrays (*9*).

Worrisome data suggest that students in the United States are falling behind students in other countries in the sciences. The National Assessment of Educational Progress "national report card" indicates only 18% of high school seniors were proficient or advanced in science in 2000 (*17*). Our educational system must prepare both future scientists and science-literate citizens for success in a world of continuing scientific and technological advances. The GCAT approach encourages faculty who focus on undergraduate teaching to become pioneers in incorporating the technological innovations of molecular biology. The GCAT community empowers faculty and students alike to solve educational problems (*1*–*5*) that seemed too big to tackle individually but were too important to ignore.

References and Notes

1. National Research Council, *Bio2010: Transforming Undergraduate Education for Future Research Biologists* (National Academies Press, Washington, DC, 2003).
2. Project Kaleidoscope, *Investing in Faculty* (Project Kaleidoscope, Washington, DC, 2001); (www.pkal.org/documents/index.cfm?page=3080).
3. L. H. Hartwell *et al*., *Nature* **402** (suppl.), C47 (1999).
4. L. A. Steen, Ed., *Math & Bio 2010: Linking Undergraduate Disciplines* (The Mathematical Association of America, Washington, DC, 2005).
5. National Research Council, *Facilitating Interdisciplinary Research* (National Academies Press, Washington, DC, 2005).
6. Genome Consortium for Active Teaching (www.bio.davidson.edu/GCAT).
7. A. M. Campbell, *Cell Biol. Educ.* **1**, 70 (2002).
8. J. L. Brewster *et al*., *Biochem. Mol. Biol. Educ.* **32**, 217 (2004).
9. Further discussion is available on *Science* Online.
10. Funded by NSF Multiple User Equipment grant no. DBI-0099720, awarded to A.M.C., L.L.M.H., T.T.E., and L.J.H.
11. Grinnell College, Pomona College, Swarthmore College, and Davidson College contribute funds equally from their 2004 to 2008 Howard Hughes Medical Institute (HHMI) educational grants to support GCAT activities.
12. GCAT members include P. Brown, B. Dunn, and D. Botstein (Stanford University), L. Hood (Institute for Systems Biology), R. Bookman (University of Miami Medical School), F. Blattner (University of Wisconsin–Madison), and E. Johnson (University of Oregon).
13. J. L. DeRisi *et al*., *Science* **278**, 680 (1997).
14. NSF workshop grants: 2003 (DBI-0305176 and DBI-0408386) and 2005 (DBI-0520908). In 2003, L. Hood, K. Dimitrov, and J. Aitchison (Institute for Systems Biology) and M. Katze (University of Washington) gave talks and shared expert advice.
15. L. J. Heyer *et al*., *Bioinformatics* **21,** 2114 (2005); (www.bio.davidson.edu/MAGIC).
16. HHMI and NSF funding have funded personnel for assessment and logistical support of scanning, shipping, and bookkeeping. Summer workshops provide honoraria for instructors.
17. National Center for Education Statistics (http://nces.ed.gov/nationsreportcard/).
18. P. Brown (Stanford University) provided chips and L. Hood (Institute for Systems Biology) donated chips and scanner use. GCAT has been funded by the Waksman Foundation for Microbiology, NSF, the Duke Endowment, and HHMI. We thank to B. Lom for help in improving this manuscript and S. Tonidanel and G. Gottfried for help with assessment.

Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5764/1103/DC1

10.1126/science.1121955

CBE—Life Sciences Education
Vol. 6, 109–118, Summer 2007

Article

Genome Consortium for Active Teaching: Meeting the Goals of BIO2010

A. Malcolm Campbell,*† Mary Lee S. Ledbetter,†‡ Laura L.M. Hoopes,†§ Todd T. Eckdahl,†‖ Laurie J. Heyer,†¶ Anne Rosenwald,†# Edison Fowlks,†@ Scott Tonidandel, Brooke Bucholtz,** and Gail Gottfried††**

Departments of *Biology, ¶Mathematics, and **Psychology, Davidson College, Davidson, NC 28035; †Genome Consortium for Active Teaching; ‡Department of Biology, College of the Holy Cross, Worcester, MA 01610; §Department of Biology, Pomona College, Claremont, CA 91711; ††Pomona College, Claremont, CA 91711; ‖Department of Biology, Missouri Western State University, Saint Joseph, MO 64507; #Department of Biology, Georgetown University, Washington, DC 20057; and @Department of Biology, Hampton University, Hampton, VA 23668

Submitted October 5, 2006; Revised December 18, 2006; Accepted January 12, 2007
Monitoring Editor: Sarah Elgin

The Genome Consortium for Active Teaching (GCAT) facilitates the use of modern genomics methods in undergraduate education. Initially focused on microarray technology, but with an eye toward diversification, GCAT is a community working to improve the education of tomorrow's life science professionals. GCAT participants have access to affordable microarrays, microarray scanners, free software for data analysis, and faculty workshops. Microarrays provided by GCAT have been used by 141 faculty on 134 campuses, including 21 faculty that serve large numbers of underrepresented minority students. An estimated 9480 undergraduates a year will have access to microarrays by 2009 as a direct result of GCAT faculty workshops. Gains for students include significantly improved comprehension of topics in functional genomics and increased interest in research. Faculty reported improved access to new technology and gains in understanding thanks to their involvement with GCAT. GCAT's network of supportive colleagues encourages faculty to explore genomics through student research and to learn a new and complex method with their undergraduates. GCAT is meeting important goals of BIO2010 by making research methods accessible to undergraduates, training faculty in genomics and bioinformatics, integrating mathematics into the biology curriculum, and increasing participation by underrepresented minority students.

INTRODUCTION

Science and mathematics education plays a vital role in the preparation of tomorrow's scientists, teachers and parents, doctors and patients, and scientifically literate citizens. For years, many leaders in science and education have called for reform (Project Kaleidoscope [PKAL], 2001; National Research Council [NRC], 2003, 2005; Handelsman *et al*., 2004; Steen, 2005). To help guide the reform process, the NRC (2003) produced a report entitled BIO2010. The report concludes that although advances in technology have caused a dramatic transformation in biological research, undergraduate biology education has not kept pace. Among the remedies offered, four major recommendations are of critical importance: 1) integrate mathematics and physical science within cell and molecular biology courses; 2) redesign lab courses to be interdisciplinary and based on research projects, rather than canned labs with predictable outcomes; 3) provide faculty development in modern disciplines such as genomics and bioinformatics; and 4) increase the number of students from underrepresented minorities in the talent pool from which future scientists will emerge. These four recommendations present many challenges, but professional societies, institutions, departments, and forward-thinking

DOI: 10.1187/cbe.06–10–0196
Address correspondence to: A. Malcolm Campbell (macampbell@davidson.edu).

faculty throughout the country are working to address them (Kumar, 2005; Campbell *et al.*, 2006a; Kuldell, 2006; Pfund *et al.*, 2006). One of these efforts is the Genome Consortium for Active Teaching (GCAT), the only laboratory-based model curriculum mentioned in BIO2010. This report documents the first 6 yr of GCAT activity and GCAT's progress toward accomplishing BIO2010 recommendations.

GCAT's mission is to bring modern genomics to undergraduate students, primarily through student research and research-based laboratory curricula. Our primary focus has been the use of DNA microarrays (sometimes referred to as chips) as a means to address the four BIO2010 recommendations outlined above (see Supplemental Material A for an overview of microarray methodology). The annual operational cycle of GCAT is illustrated in Figure 1. In the spring, GCAT solicits requests for DNA microarrays from participating faculty. Microarrays from 11 different species are currently available to GCAT members. GCAT contracts for the production of microarrays during the summer and distributes the microarrays in the fall. Faculty and students design and perform their experiments and ship their hybridized chips overnight for scanning on a GCAT community scanner purchased with support from the National Science Foundation (NSF), or on backup scanners available at other locations. GCAT then delivers tiff microarray data files to the student investigators by File Transfer Protocol (FTP). Students and faculty analyze their own data, and they have access to data produced by all other GCAT members. Many investigators use MAGIC Tool (Heyer *et al.*, 2005), free software provided by GCAT (Heyer and Campbell, 2004a).

GCAT members are free to pursue their own research or research-style teaching without any limitations by GCAT. The only requirements for participation are 1) only undergraduates may use the microarrays; 2) faculty and students must participate in assessment; and 3) all data and protocols are open access for the GCAT community. Faculty training in microarray laboratory protocols and data analysis methods is provided by workshops. NSF has funded three workshops to date, and it has recently awarded GCAT a new grant to fund three more workshops during summers 2007, 2008, and 2009. New and veteran GCAT faculty alike appreciate the collective expertise and support of the GCAT community, as evidenced by the high level of activity on the GCAT-Listserv (GCAT-L) e-mail distribution list.

Based on assessment data from students and faculty, GCAT is having a significant impact. Faculty report very strong support for GCAT, and students report learning gains and attitudinal changes as a result of their GCAT experiences. Faculty self-reported substantial gains in their scholarship and teaching activities as well as overall satisfaction with GCAT. This report documents these successes and identifies new ways in which GCAT can reach a wider audience.

MATERIALS AND METHODS

DNA Microarray Resources

Currently, DNA microarrays are purchased with funds from Howard Hughes Medical Institute (HHMI) awarded to Grinnell, Pomona, Swarthmore, and Davidson. Chips were produced at numerous academic labs. The cost to participating faculty is $50 for the first microarray and $20 for each additional microarray per species to cover the costs of shipping and scanning. For the first 3 yr, before the advent of HHMI funding, academic labs donated the chips to GCAT free of charge; this service provided crucial support necessary for launching GCAT (from Patrick Brown at Stanford University [Stanford, CA] in year 1 and Leroy Hood at the Institute for Systems Biology [Seattle, WA] in years 2 and 3).

MAGIC Tool software was developed using funds from Davidson College, HHMI, and NSF. The software is written in Java, so it works on all operating systems (i.e., Macintosh OS X, Windows, and Linux), is freely available for downloading (Heyer and Campbell, 2004a), and is open source. Computers must have at least 512 MB of RAM to run MAGIC Tool, but we recommend 1–2 GB of RAM for optimal performance.

GCAT offered NSF-funded workshops during summers 2003, 2004, and 2005. In 2004, we offered one complete workshop (data analysis and wet lab components) as well as some sessions for data analysis only (Campbell *et al.*, 2006a). The workshop participants produced and analyzed two-color microarray data from the yeast diauxic shift from anaerobic to aerobic metabolism. In that way, the participants were able to obtain data directly comparable with a ground-breaking published study (DeRisi *et al.*, 1997). Professors using variations on this experiment in their classes can augment their student data by adding the public domain data before analysis (Heyer and Campbell, 2004b). The two concurrent workshops in 2005 were 5 d long, including training in both data production and data analysis (GCAT, 2005). NSF funding has now been obtained for additional workshops in 2007, 2008, and 2009 to cover both aspects of microarray work. All workshop participants receive materials to take home, including a CD containing MAGIC Tool software, and all raw and analyzed data produced at the workshop; the MAGIC Tool user's guide; a data analysis exercise for further practice; guided activities in comparing and clustering gene expression profiles; a reading quiz on the DeRisi paper; laboratory protocols with annotations; guidelines for faculty timing of weekly labs; and notes on reagents and suppliers. The hands-on, collaborative nature of the workshop ensures that participants have experienced the microar-

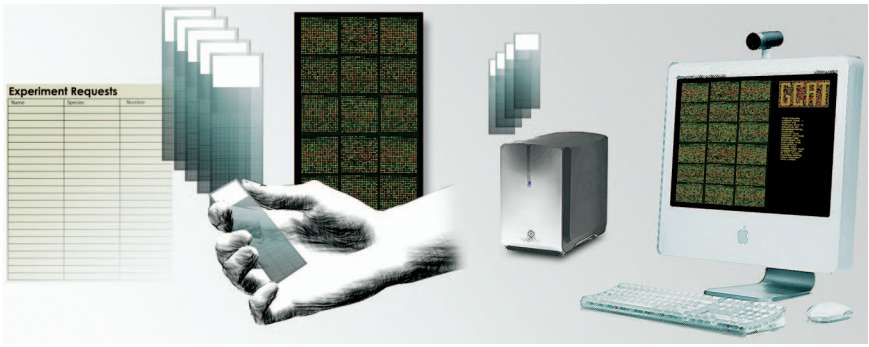


Figure 1. Outline of GCAT microarray distribution plan. Faculty who teach undergraduates submit their requests for microarrays, which are produced by several academic labs. Students perform the experiments, then the chips are scanned and data are posted to an FTP server for students to analyze.

ray process from beginning to end and have had a chance to learn from mistakes in a supportive environment. Continued support for faculty participants is provided after the workshops via e-mail (GCAT, 2003) and by a helpdesk (staffed by NSF-funded undergraduates), through 2009. In addition, some of us have led other less comprehensive workshops at a variety of locations.

Surveys and Statistics

All student and faculty surveys were conducted online (Tonidandel, 2004). A pre- and postterm design was used for the student assessment, whereas faculty completed an assessment only at the end of the term. The preterm assessment (see Supplemental Material B) was completed by students at the beginning of a semester. The assessment asks students to respond with basic demographic information and to complete an 11-item test of prior knowledge (see Supplemental Material D). When the semester was over, students' knowledge was again assessed along with their attitudes about using the GCAT materials (see Supplemental Material C). The online survey for faculty asked whether they used the GCAT materials in their classes, and it assessed their attitudes regarding the GCAT experience in their particular course. A mixed factorial analysis of variance was used to evaluate statistical significance.

Of 52 professors identified by students as supervising the use of GCAT materials at their home institutions, 43 professors completed the faculty survey at the end of the program. Three faculty members responded to the survey twice because they used GCAT materials in both semesters of the 2005–2006 school year. In July 2004, GCAT offered a series of NSF-sponsored hands-on workshops for faculty interested in curricular innovations to include gene expression analysis via microarrays. Thirty-seven participants attended one of two 1.5-d dry lab workshops that introduced the microarray method and covered data analysis by using public domain data. Participants learned to work with the open-source MAGIC Tool spot-finding and analysis software, along with other free packages, to analyze public domain data sets in short projects. Twenty-three additional participants continued with a 2.5-d hybridization workshop, which involved hands-on preparation of fluorescently labeled probes for yeast expression microarrays, their hybridization, data acquisition, and data analysis by using the methods presented in the earlier workshop. In June 2005, the GCAT team conducted a follow-up evaluation to assess the degree to which the participants met their goals. An e-mail invitation was sent to 39 faculty members in late May 2005, requesting that they respond to a 10-item online survey assessing participants' retrospective evaluation of the workshops and their use of the microarray tools during the 2004–2005 academic year. Twenty-five people (64%) returned surveys by June 10; of those, 10 participated in the wet and dry labs, 14 only in the dry labs, and 1 was unable to attend but gathered and implemented the workshop materials.

RESULTS

Origin and Growth of GCAT

The concept of GCAT was inspired by a 1999 presentation given by Dr. Pat Brown of Stanford University. Two of us (A.M.C. and M.L.L.) realized that this technology embodied the power of genome-wide strategies and could be affordable for undergraduate institutions if we pooled our resources. Brown agreed to provide us with 144 yeast DNA microarrays. With this promise, we used the annual meetings of PKAL and the Council on Undergraduate Research (CUR) to recruit faculty who would be willing to take a collective leap of faith and learn how to conduct microarray experiments together but on different campuses. None of us had ever performed such an experiment, but the procedure was conceptually accessible and seemed relatively straightforward. Twenty-three faculty agreed to participate in the inaugural year (2000–2001) of GCAT.

During GCAT's first year, we realized that two potential limitations might prevent widespread adoption of microarray strategies: the cost of microarrays themselves and access to a microarray scanner. A collaborative grant from NSF allowed us to purchase a scanner in fall 2001 with additional funding from Missouri Western State University, Pomona College, and Davidson College. User fees of $20 per microarray were collected to cover the expense of its service contract. After the first year of demonstrated success using the microarrays with undergraduates, other academic labs donated additional microarrays. Dr. Hood, for example, donated 400 yeast DNA microarrays over 2 yr. We used Michael Eisen's ScanAlyze (Eisen, 2006) and commercial GeneSpring software (Agilent Technology, Santa Clara, CA) to analyze the data, because they were offered to GCAT members free of charge for educational purposes.

Because GCAT relied on donated microarrrays, we were hesitant to advertise in any formal manner. However, a number of our institutions were invited to participate in the HHMI competition for undergraduate institutions in 2003. Working with guidance from Stephen Barkanic of HHMI, the proposed budgets in applications from 24 GCAT member institutions included funds to support direct costs of the consortium. We hoped that 20% of those institutions would be successful, allowing extended stable support. In fall 2004, Pomona, Grinnell, Swarthmore, and Davidson were each awarded 4-yr educational grants from HHMI, which included funding for GCAT. With this funding, we could purchase microarrays to meet the growing demand of the consortium. For yeast microarrays, we purchased our own whole-genome oligonucleotide sets and contracted with the microarray core facility at Washington University in St. Louis, MO, to produce microarrays for GCAT. Membership in GCAT has continued to grow in two dimensions (Figure 2), beginning with exclusively yeast microarrays in 2001 to 11 different types of microarrays in 2006. In the first 7 yr, GCAT provided ~5000 DNA microarrays for use by approximately 6000 undergraduates.

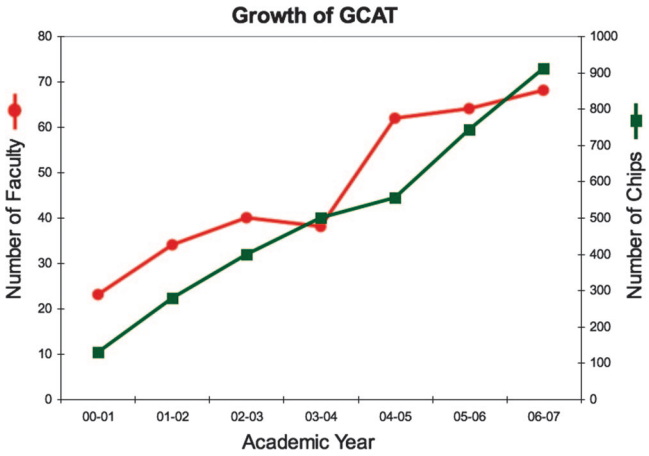


Figure 2. GCAT growth over seven years. GCAT has expanded the number of microarrays distributed (right Y-axis) and the number of faculty (left Y-axis) participating each year.

As GCAT grew, we recognized two new limiting factors: faculty training and appropriate software. Very few GCAT faculty had formal training with microarray techniques or analysis of the data. This need was expressed very clearly by the numerous attendees at an American Society for Microbiology (AMC) symposium chaired by AMC in 2002 (Campbell, 2002). In response, a core group decided to develop student-friendly lab protocols and to offer workshops for faculty training. Free software programs such as ScanAlyze and Cluster (Eisen, 2006) were restricted to the Windows platform, whereas many commercial packages were cumbersome and prohibitively expensive. Therefore, one of us (L.J.H.) worked with several undergraduates to write MAGIC Tool (Heyer *et al.*, 2005) for data analysis. MAGIC Tool is written in Java, and so runs on all major computer platforms; it is freely available, and is open source (Heyer and Campbell, 2004a).

From the outset, GCAT has been guided by a few simple principles:

1. bring genomic methods into the undergraduate curriculum, primarily through student research;
2. share resources to make experiments affordable;
3. be as inclusive as possible so all schools can participate;
4. create a clearinghouse of information for faculty;
5. provide all data freely to anyone for pedagogical use;
6. develop a distributed community to help each other trouble-shoot and develop curriculum;
7. make assessment a fundamental requirement for participation; and
8. encourage participants to set their own educational and research goals.

By following these principles, GCAT has reached many campuses, some of which have been overlooked in national educational reform efforts (e.g., small campuses and community colleges), have student populations who are underrepresented in science, or both (Figure 3). Microarrays provided by GCAT have been used by 141 faculty on 134 campuses in 36 states as well as two universities in Canada and one university in Australia. Of the 134 U.S. campuses, 21 (16%) serve large numbers of students from underrepresented minorities (1 in Hawaii, 3 that serve a mixture of ethnic groups, 7 historically black colleges or universities, and 10 Hispanic-serving institutions; U.S. Department of Education, 2006). A majority of the GCAT campuses are small, 4-yr, liberal arts colleges, but GCAT membership includes faculty from three community colleges as well as large universities such as the University of Georgia (25,000 undergraduates), California State University at Sacramento (23,000 undergraduates), University of Louisville (22,000 undergraduates), Boston College (9000 undergraduates), University of Southern Maine (8600 undergraduates), and Georgetown University (7000 undergraduates). In addition to direct support for these schools, GCAT has provided student-friendly protocols, curriculum, and pedagogical advice to research powerhouses such as Massachusetts Institute of Technology (MIT, Cambridge, MA) and University of California at San Diego as they began using microarrays in undergraduate laboratories. GCAT helps faculty overcome some of the common barriers to the introduction of new technologies into undergraduate curricula, and it offers faculty the freedom to adapt the materials to their own research interests and institutions.

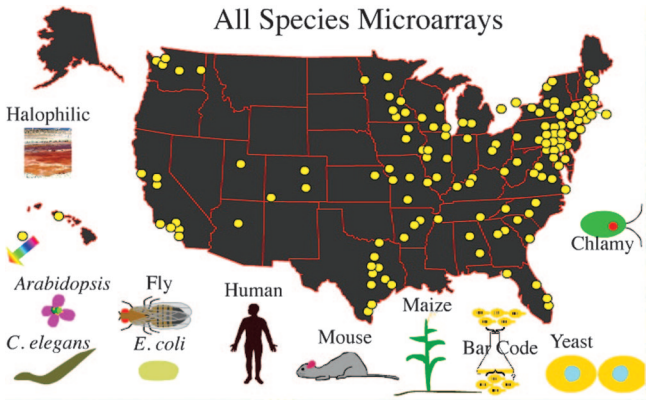


Figure 3. Map of GCAT-participating schools. GCAT is composed of 141 faculty on 134 campuses in 36 states, including two universities in Canada and one in Australia (colored arrow), with nodes serving as hyperlinks to the appropriate departments (GCAT, 2006b). This screen shot is from an interactive map that allows viewers to see the geographical distribution of users for each type of microarray. Contact information for GCAT faculty is available and organized by academic year.

Student Outcomes

Students who work with DNA microarrays tend to be juniors (29%) or seniors (57%), and either biology (74%) or chemistry (14%) majors. Men and women are equally represented and 18% are from underrepresented ethnic groups. Nearly 60% want to pursue medical careers and 33% want to pursue a Ph.D. in cell/molecular biology. Based on these career goals, it is not surprising that most students had already completed introductory biology (93%), organic chemistry (77%), calculus (72%) and first-year physics (67%). Only 5% had completed a course in genomics or bioinformatics before working with the GCAT materials.

Nearly 80% of the students were able to progress through the experimental procedure far enough to have their microarrays scanned, although only 54% reported that they obtained usable data. Of faculty surveyed, 70% reported that some of their students obtained usable data. Among these 34 faculty, the average success rate (scanned microarrays with usable data) of their students was 81%, with 22 faculty reporting that 100% of their students produced usable data. Therefore, the overall student success rate as reported by faculty is estimated to be ~56% (0.70×0.81). By comparison, 212 (76%) of the 277 microarrays scanned at Davidson College between June 17, 2005, and July 28, 2006, produced usable results. This is a very impressive success rate for a method that early skeptics thought was impossible for undergraduates to use. Notably, some faculty whose students at first obtained no usable data are now getting very good results due to improved methods for RNA isolation and cDNA production. As faculty determine the best way to produce labeled

Table 1. Scores on 11 knowledge questions for pre- and postsurveys[a] (n = 409)

Question	Subject matter	Correct pre-GCAT (%)	Correct post-GCAT (%)	Increase (%)
1	Microarray experimental error–dye bias	23.1	59.3	36.2
2	Microarrary experimental error–gradient	32.7	43.2	10.5
3	Microarray negative controls	28.9	39.2	10.3
4	Microarray experimental design	33.9	72.1	38.2
5	Gene expression ratios using a graph	5.0	10.8	5.8
6	Gene expression–probability	20.9	21.1	0.2
7	Gene expression–gene clusters	30.0	52.3	22.3
8	Gene expression–regulatory cascade	28.4	43.3	14.9
9	Gene expression–gene circuit graphs	37.9	49.7	11.8
10	Interpreting microarray results	40.0	59.0	19.0
11	Diagnosis with microarrays	54.9	67.4	12.5

[a] Performance increased significantly ($p < 0.05$) on all questions except item 6.

probes for their system, the success rate should continue to climb above 70%.

Knowledge Gains. Eleven questions to test knowledge were presented in identical form on the two surveys (pre- and postsurveys) taken many weeks apart; in total, 409 students responded to both sets of questions (see Supplemental Material B–D for the 2 surveys and 11 knowledge questions). Students were instructed to answer questions without the use of notes or consultation with friends. Those questions presented hypothetical scenarios pertaining to gene expression and microarray experimentation techniques. The questions were not focused on details or specific facts, but they were designed to be very challenging and to emphasize problem solving and data analysis. With the exception of the final question, correct response rates for each question in the presurvey were below 50%. On average, students were not knowledgeable about microarray experimentation relating to either DNA or RNA at the outset of their GCAT experiences. The average percentage of correct responses across all test items before GCAT training was 30.5%. Item 5 was particularly difficult for student participants; only 5.0% of students answered it correctly on the preprogram survey. Correct response rates for each item and students' knowledge gains are found in Table 1.

Knowledge scores improved substantially after the GCAT program; the average percentage of correct responses on the post-GCAT survey was 47.1%. Correct responses for each item increased on average by 16.5%. All gains were statistically significant, with the exception of item 6. Questions 1 and 4 showed particularly large improvements, and both specifically pertain to microarray experimentation. Knowledge gains and final performance were lowest on items 5 (10.8% correct) and 6 (21.1% correct); the subject matter for these two questions relates to gene expression ratios and probability.

Although the pre- and postsurveys showed significant gains for students who worked with GCAT materials, it would be interesting to know whether similar gains were possible for students who learned about microarrays in a lecture-only course. Fortunately, one GCAT faculty member volunteered to have her genomics lecture course of 18 students take the pre- and postsurveys as a control group. Lectures and reading assignments were congruent with other classes that used GCAT materials, but the control class did not conduct laboratory experiments. Students in the control group gained an average of 3.5% correct responses at the end of the semester, and this increase was not statistically significant. Conversely, the remaining students, who implemented GCAT materials in their laboratories (n = 377), showed significant increases on knowledge questions ($p < 0.01$); the average student increased by 16.4%. This improvement is roughly equivalent to two additional correct answers on the 11-item quiz. There was significant correlation with time spent working with the microarrays and use of GCAT materials ($p < 0.05$). Students who conducted microarray experiments improved significantly in knowledge assessments over the course of a semester, whereas students who did not participate in laboratory activities did not show significant knowledge gains over the same amount of time.

Attitude Changes. After their GCAT experiences, students rated their change in interest and understanding of genomics, biology, and research on a 7-point scale where 1 is decreased a lot and 7 is increased a lot. On average, students' interest and understanding of all three areas increased over the course of the GCAT program (Table 2). Students also rated the effectiveness of various GCAT activities on a 7-point scale, where 1 is not effective at all and 7 is highly effective. Table 3 presents the percentage of students who rated these activities at least 4.00 and at least 5.00 on the 7-point scale. The average effectiveness value students assigned to all of the activities was 5.20, and mean scores on individual activities ranged from 5.06 to 5.32. On average, students did not judge any activity to be drastically more or

Table 2. Student attitude change (on a 7-point scale where 1 is decreased a lot and 7 is increased a lot) in interest and understanding of subject areas (n = 409)

Area	Mean	SD
Genomics	5.5	1.1
Biology	5.5	1.1
Research	5.4	1.2

Table 3. Student responses (on a 7-point scale where 1 is not effective at all and 7 is highly effective) measuring satisfaction with methods used in lab

GCAT activity	% of students who rated the activity with at least 4.00 or 5.00			
	Mean	≥ 4.00	≥ 5.00	N[a]
Practicing data analysis before I began analyzing my own data	5.25	93.6	67.1	313
Isolating RNA or genomic DNA to produce probe	5.32	94.1	70.0	323
Producing the fluorescently labeled probe	5.22	94.4	68.9	306
Hybridizing the probe with the spotted DNA	5.20	92.8	70.1	334
Designing my own experiment	5.13	87.3	64.3	244
Analyzing data from public domain source	5.22	94.7	65.8	325
Reading papers that used DNA microarrays	5.06	88.9	62.4	343

[a] Number of students who did not rate the activity "not applicable."

less effective than others. All average ratings are above 5.0 on the 7-point scale, indicating that students judged all of the activities to be effective.

Faculty Outcomes

2005–2006 Academic Year Responses. Faculty estimated the number of weeks allocated for each of the activities performed by their students (i.e., isolate mRNA, make cDNA probes, make total genomic DNA probes, hybridize probes to microarray, analyze students' own data, analyze data from public sources, and (students) design their own experiments). We evaluated both the time devoted to each activity (e.g., 2.6 wk) and the frequency with which faculty members reported a particular activity (e.g., 80% of the faculty may have reported doing a particular activity) but not the frequency with which each activity was done at each institution (i.e., 80% does not mean an activity was done 80% of the time at an institution). Hybridizing probes to a microarray (80.0%) was the activity reported most often by the faculty, whereas only a small percentage asked that students make total genomic DNA probes (12.2%). Excluding the lecture-only control group, 90.6% of the professors reported that their students performed at least three of the GCAT activities during the semester, using an average of 1.8 wk per task. Students were given the most time to make total genomic DNA probes (2.6 wk) and to analyze their own data (2.5 wk). The least amount of time was allotted for making cDNA probes and analyzing public domain data (1.12 wk for each). Professors were asked how they measured student performance when they used GCAT materials (Table 4). The most common assessment tool used by GCAT professors was informal feedback (62.2%), but term papers and lab reports were nearly as popular (51.1%). Other methods for assessment included tests (42.2%) and poster presentations (33.3%). About 24% of the professors reported "other" techniques (e.g., three faculty used lab notebooks, whereas single responses were recorded for honors thesis, constant discussion with the student, constructive participation in course discussion [graded daily], laboratory work, and quizzes). A small number of faculty (8.8%) assessed students through preparation of a manuscript for publication.

Faculty received funding to support use of GCAT resources from a variety of sources, with most support coming from departmental funds (89.0%). Institutional and extramural funds each supported 20% of the participating faculty. Only 4.4% of professors indicated that they received no funding for using the materials provided by GCAT. Although most professors (61.7%) did not feel that their implementation of GCAT materials was limited by computer resources, 38.3% indicated that they experienced such limitations.

GCAT faculty rated their agreement with statements describing their experiences with GCAT (Table 5). Most faculty responded that they would not have access to microarray technology without GCAT, and they reported a positive overall GCAT experience. Faculty participants generally agreed that the online protocols and e-mail distribution list (GCAT-L) were helpful. Working with DNA microarrays is inherently an interdisciplinary effort, as illustrated by two unsolicited faculty comments. A biology faculty member commented about GCAT,

> "You have awakened parts of my brain that have been dormant since my last stats course. The only reason I have gone over the manual so carefully is that this is my first time teaching microarrays, or even using them, for that matter. GCAT has been remarkably helpful to me. In fact I don't think I would have undertaken this new module in my lab course without the tools GCAT makes available."

Table 4. Faculty assessment methods from 2005 to 2006 academic year

Assessment method	Professors who used each assessment method (%)
Test	42.2
Term paper/lab report	51.1
Poster presentation	33.3
Oral presentation	26.6
Manuscript for publication	8.8
Course evaluation	33.3
Informal feedback	62.2
Other	24.4

Table 5. Faculty responses from 2005 to 2006 academic year by using a 5-point scale, where 1 is strongly disagree and 5 is strongly agree

	Mean	SD
I would have access to microarray technology without GCAT.	1.5	0.75
The online protocols available on the GCAT website were useful.	4.4	0.69
The GCAT-Listserv was helpful.	4.2	1.0
The collection of other GCAT members as a support network was a significant factor in launching microarray technology on my campus.	4.2	0.79
Overall, I had a positive experience using GCAT.	4.6	0.60
I would use GCAT again in the future.	4.7	0.63

Conversely, a mathematics professor remarked (with identifiers removed for anonymity),

> "I am working with a student who is trying to do some serious data analysis on [Dr. X's] chips – we are having great fun learning and thinking about how to understand and analyze all of this data – we are going back to basics – and have already found some interesting things – we are excited that our mathematical results seem to be synching up with [Dr. X's] biological results/insights. I hope we are not the first on board with the GCAT project that are primarily data analysis oriented folks – but I daresay, if we are, we won't be the last! This project provides a great area of study for undergrad students interested in data analysis but not necessarily the actually bench work (but of course they need to understand what happened on the bench to understand the data!). Also, it is a fantastic opportunity for math/stats and bio majors (and professors!) to interact! Hmmm... looks like your project may be expanding to us lab phobic (but data loving!) types!"

These quotes illustrate the power of providing stimulating opportunities to faculty who otherwise would not venture out of their comfort zones.

2004 GCAT Workshop Outcomes. Immediately following the 2004 workshop, all faculty indicated the workshop was very good. One year later, 67% of the respondents said that, overall, the workshop they attended was excellent (80% of the wet lab attendees and 57% of the dry lab attendees). The remainder reported the workshop was very good (29%) or good (4%); none reported that it was fair or poor. When asked to select the aspects of the workshops that, in retrospect, were most valuable in preparing for and teaching during the 2004–2005 academic year, participants consistently indicated that the handouts and notebook were critical (70%). Additionally, 48% found the protocols for data analysis valuable, and 39% found the protocols for hybridization valuable. In open-ended responses, two participants wrote that gaining confidence to use the tools was important, and one wrote that doing the data analysis in the workshop was useful. Importantly, 61% of the respondents indicated that networking with instructors and other participants via GCAT was among the most important aspects of the workshop. This finding is consistent with the 2004 onsite evaluation, which indicated that participants felt the collaborative nature of the workshops was among the most valuable aspect of the workshop.

Upon completion of the workshop, respondents to the 2004 evaluation indicated that they intended to alter existing courses to include data analysis with MAGIC Tool, expected to add a wet lab in upper-division courses, and that they planned to emphasize microarrays in several courses across the curriculum. The 2005 survey asked the 2004 workshop participants whether these courses had in fact been altered to accommodate what they had learned (Table 6). Nineteen of the 24 respondents (79%) who attended the workshop used the materials in at least one course (including independent study) during the 2004–2005 academic year, as did the one instructor who could not attend the workshop but received the written materials. The others indicated that they were still in the curricular planning phases or had committed to using the materials in a class scheduled for the 2005–2006 academic year. Of the 20 respondents who reported using the workshop-supplied information during 2004–2005, 18 said they met at least one goal that they had proposed before taking the workshop. Many participants used the materials in more than one class; two participants indicated they altered three courses to use what they learned in the workshop. Two respondents added a dry lab, and two added a wet lab. The average number of courses modified by wet lab workshop attendees was 1.6 (SD = 0.84); the average number of modified courses for dry lab attendees was 0.86 (SD = 0.77). These values differ significantly ($t = -2.2$, $p < 0.05$), suggesting that attendance at the wet lab workshop may yield better preparation or more confidence for using the microarray tools.

The GCAT workshop materials were used in 10 different types of courses. MAGIC Tool was used in genetics classes by 25% of respondents. The software was used by other respondents in Biochemistry (2), Introductory Biology (2), Bioinformatics (2), Molecular Biology (1), Advanced Molecular/Cell Biology (1), Data Analysis (1), Biotechnology (1), Cell Physiology (1), and Microbiology (1). Twenty-five percent of the respondents indicated that they used the microarray tools for independent research with students. Interestingly, only one faculty member developed a wet lab component but did not use MAGIC Tool software, which reveals the intense need for free software that is student

Table 6. Faculty goals prior to 2004 workshop and percentage who accomplished these goals

Proposed change	% Participants (n = 20)
Proposed to use in specific lectures and used the material in those lectures	35 (7)
Proposed to use in specific labs and used the material in those labs	60 (12)
Proposed to use in research and did so	15 (3)
Proposed to use in specific lectures but used the materials in other ways	20 (4)
Proposed to use in specific labs but used the materials in other ways	15 (3)

friendly. An additional benefit faculty identified was their increased collaboration as a result of the workshop (Table 7). This result addresses important recommendations in the BIO2010 report that call for increasing faculty development opportunities and building communities with a shared commitment to educational reform.

Twenty faculty reported that a total of approximately 800 students participated in a course or in research that used workshop materials in some way. Individual faculty reports ranged from engagement of 2–220 students, with an average of 39.5 and median of 20 per faculty member. Sixteen students were involved in advanced tutorials or independent research using the microarray tools; most had successful experiences. Five students made presentations at their respective colleges or universities or at the regional Sigma Xi conference; one received a grant for an honors proposal using the microarray technique. Four students were conducting research for the first time. However, one respondent indicated that his two students had a less than optimal experience because the data were not readable and the term allowed no time for replication.

Open-ended faculty comments included the following:

> ". . . the presentation of this subject makes [students] realize and practice the close interaction biology/genetics has with other fields like mathematics. They enjoyed [being] introduced to a novel genetic technique. They said they can understand better and related more [of] their class to real life, like when they watch health news and advances in science."

> "Because [this course] was an absolutely introductory exposure to using microarrays for faculty and students, exposure was limited. I anticipate a strong uptick in activity in the next year as new molecular faculty become involved."

> "Many students have come back and said they got jobs or were assigned or allowed to do special projects in graduate schools because of their experience with microarrays. Many others come back and tell how helpful what they learned in the class was with job experiences or graduate school and how they feel ahead of many others attending classes."

> "I think students were extremely excited to have exposure to microarray technology and data analysis."

> "The students said this made them think about what they were doing more critically and it made the whole process seem less 'magical.'"

> "I didn't have quite as much time as I had hoped for data analysis. I found that it took longer than I anticipated for students to grasp the analysis."

> ". . . none of our arrays worked. Unfortunately, I think a lot of it was lost on the students. Negative results tend to confuse them, they are not yet appreciative of the fact that experiments don't always work."

Table 7. Faculty-perceived collaborative benefits from attending 2004 GCAT workshop

Collaborations after returning to home campus	% Participants (n = 23)
Talked with GCAT faculty via GCAT-L	48
Collaborated with faculty at my home institution to assist in curriculum/course development	48
Worked with teaching assistants	9
Discussed material at department/faculty meetings	48
Shared with colleagues at other institutions	9
Other	13

DISCUSSION

Students

The main purpose of GCAT has been to use DNA microarrays as a vehicle to bring genomics into the undergraduate curriculum. The NRC recommends undergraduate curricula should blend mathematics with cell/molecular biology and laboratory experiences that are research-based and interdisciplinary (NRC, 2003, 2005). GCAT provides ready access to an exciting area of interdisciplinary research that is moving into clinical applications—DNA microarrays. Analyzing real microarray data requires students to understand the complexities of genomics and use quantitative methods such as bioinformatics to understand their data and statistical analysis to interpret their results. Students enjoy working with cutting-edge techniques, and they see the value of an integrative approach to science. GCAT helps teachers provide students with valuable skills and train them to think in ways that are critical to the future success of research scientists (Hartwell *et al*., 1999).

Based on the knowledge surveys, students have made significant gains in many areas (Table 1). Although we provide here some preliminary evidence that GCAT offers learning benefits over a control group, the conclusions one can draw from these data are limited by the small size of the control group. As a result, we are expanding our evaluation efforts to include more control classes from a variety of institutions in an attempt to determine more concretely the learning gains associated with the wet lab portion of GCAT. The high percentage of microarrays with usable data is a tribute to the student-friendly protocols and faculty support network. Students attending a wide range of institutions have been able to perform microarray experiments, because the costs of microarrays are low and the software is free. Only 25% of GCAT faculty have access to extramural funding of some kind (including HHMI educational grants), which explains why affordability is so critical to GCAT's success. Furthermore, student interest and understanding in genomics and appreciation of research increased (Table 2) in part because they felt the methods were beneficial (Table 3).

Based on student learning gains, GCAT faculty and students should devote more time to gene expression ratios and probability, because these topics are essential to understanding gene expression data. Student weakness in these topics reflects the traditional lack of effective integration of mathematics in biology programs. Working with microarrays creates an opportunity for faculty to integrate math and biology, as recommended in BIO2010. The next area of concern is to make sure students fully understand the experimental method and how to troubleshoot. Fewer than half of student participants were able to answer items 2, 3, 8, and 9

correctly after their GCAT experience, and all of these items pertain to microarray experimentation methods. Although significant gains were observed for these questions, there is room for additional improvement. Professors might want to emphasize a wider range of microarray techniques in their future implementations of GCAT activities.

Faculty

Faculty development is an ongoing concern for every campus, and BIO2010 recognized this as a critical issue (NRC, 2003). GCAT provides an easy way for faculty to learn a new method with their students. The GCAT protocols and MAGIC Tool software minimize the risk to faculty of trying this method for the first time. Often, faculty do not have colleagues on campus who can help them. GCAT's network of supportive colleagues encourages faculty to learn a new and intimidating method. The workshops are efficient and effective, based on the number of courses altered and the number of students affected after 1 yr. NSF has provided funding for three more summers of workshops, projected to involve a total of 120 faculty. If we multiply the number of faculty trained each year (40) by the number of students affected based on the 2004 workshop (39.5), then by 2009, ~9480 undergraduates will have been provided with access to microarrays as a direct result of future GCAT workshops. This number does not include the current number of GCAT faculty (141 to date) and all the students they will reach. Furthermore, because many GCAT faculty teach at minority-serving institutions, another BIO2010 goal is being supported—diversification of future researchers.

No program is perfect, and there are areas where GCAT could improve. Because faculty indicated that interactions with other GCAT members were very significant factors when they launched microarray technology on their own campuses, additional networking resources such as online curriculum workshops or electronic communication could potentially enhance GCAT faculty training and success rates. Workshop participants from 2004 indicated that some additional information or materials would have facilitated increased use of microarray data analysis in the curriculum. Four primary suggestions were clear from open-ended comments:

1. Help with course planning. Faculty particularly sought additional instructor guidelines (perhaps lesson plans), especially focused on how best to explain and present the experimental design, and, critically, requested information regarding the prep time needed to incorporate the materials into the course with confidence. The need for a course guide for teaching the analysis component was noted by several respondents.
Response: In the future, GCAT may sponsor curriculum development workshops, but currently there is no funding for this. At this time, the best option for new faculty is comparing notes with other GCAT faculty on GCAT-L.

2. Help with analysis. Faculty requested comparisons with other software packages and more instruction on related analysis programs (e.g., ScanAlyze).
Response: The workshops do help with data analysis but due to the high costs of commercial software, we support only free programs. ScanAlyze is free, but does not work on Macintosh, and does only part of the data analysis, whereas MAGIC Tool does the full analysis in a single program. MAGIC Tool was designed to be student-friendly and to help users understand the consequences of various actions, such as background subtraction. Most other programs were designed for researchers and do not readily lend themselves to instructional applications. Faculty may decide to use other programs, and GCAT does not place any constraints on the software programs its members use. The former company Silicon Genetics provided free access to GeneSpring, a Windows-compatible analysis program, to GCAT members. Agilent Technologies is continuing to provide free access to GeneSpring during research-style classes, after reviewing the laboratory syllabus, but no longer permits publications to use their graphics, even undergraduate research projects.

3. Documentation. Suggestions included developing help files, distributing slide sets (PowerPoint) of the lectures, creating a detailed handout for MAGIC Tool explaining *why* certain tasks are performed, and publishing a troubleshooting guide.
Response: NSF has provided funds for a helpdesk staffed by students as well as for the production of tutorials that contain movies to teach users how to use MAGIC Tool. We hope this will address the needs of many faculty and students.

4. Networking. Faculty want contact with others using similar protocols.
Response: In addition to electronic community building via GCAT-L, GCAT faculty attend many professional workshops and may seek each other out during these meetings. As stated in number 1 above, GCAT may organize curriculum workshops, but currently does not have funding to do this.

No matter how much the GCAT community offers to faculty, some problems are institutional and cannot be solved by GCAT. Faculty challenges include the following:

- Faculty may not be able to predict their teaching assignments into future years, so long-term planning for curricular change can be difficult.
- Although many faculty have a great desire to use these new materials, they require extensive time to prepare. In some cases, admirable goals cannot be met within the existing time constraints.
- The need for unusually large amounts of computer RAM must be considered in advance; some labs are not adequately equipped.
- Faculty found it hard to imagine and develop productive lesson plans to incorporate the tools.

Future Directions for GCAT

By focusing on microarrays as a tool for understanding gene expression and functional genomics, GCAT has accomplished many of the BIO2010 goals; but not all faculty want to work with microarrays. Exploration of other routes may help faculty bring genomics into their courses while remaining consistent with our goals. Two additional efforts are underway currently. The first is a collaboration with Dr. Sarah Elgin at Washington University who is working with college faculty on actual, authentic DNA sequencing projects. She has collected resources and protocols so that undergraduates can learn to finish and annotate genome sequences (Elgin, 2005). The other project is a collaboration with Randy Rettburg and Drew Endy at MIT, working in the field of synthetic biology (Rettburg and Endy, 2006). Synthetic biology blends mathematics, computer science, and engineering with molecular and cell biology (SyntheticBiol-

ogy, 2006). Furthermore, to extend the pipeline of students who can work in genomics as undergraduates, we have developed microarray wet lab simulations and paper activities for high school students (GCAT, 2006a; Campbell *et al.*, 2006b). These tools allow teachers to use hands-on learning activities to blend mathematics with biology in a way that students enjoy and retain.

Accessing Materials

Any faculty member may join GCAT and there is no fee for joining. All you need to do is sign up for GCAT-L to receive e-mail announcements (http://www.bio.davidson.edu/projects/GCAT/GCAT-L.html), including the free summer workshops for faculty (http://www.bio.davidson.edu/projects/GCAT/gcat.html#workshops). Only undergraduates can use the DNA microarrays, though anyone can analyze data with MAGIC Tool or use any of the other resources on the GCAT or MAGIC Tool websites. The microarray simulation kit is available for anyone (http://www.bio.davidson.edu/projects/GCAT/HSChips/HSchips.html). GCAT invites faculty who teach undergraduates to participate in synthetic biology (http://www.bio.davidson.edu/projects/GCAT/Synthetic/synthetic.html) or to contact The International Genetically Engineered Machine (iGEM) leaders directly (http://parts2.mit.edu/wiki/index.php/Main_Page).

ACKNOWLEDGMENTS

GCAT gratefully thanks several agencies for financial contribution: The Waksman Foundation for Microbiology, National Science Foundation, Howard Hughes Medical Institute (grants 52005120, 52005137, 52005202, and 52005328), Associated Colleges of the South, Davidson College, Pomona College, and Missouri Western State University. Many investigators have donated time and resources to GCAT: Patrick Brown, Barbara Dunn, Leroy Hood, Michael Katz, Krassen Dimitrov, John Aitchison, Steve Proper, Richard Bookman, Jef Boeke, Anna Ballew, Corey Nislow, Fred Blatner, Patrick Schnable, David Galbraith, Elaine Mardis, Sarah C. R. Elgin, Seth Crosby, and Chris Sawyer. We thank Sally O'Connor, Diane Okamuro, Mary Clutter, Machi Dilworth, Gerald Selzer, Rob Last, and Angela Klaus at NSF for their vision, guidance, and encouragement. We thank the Institute for Systems Biology, Georgetown University, and Morehouse College for hosting our first three workshops. We also thank Peggy Maiorano for invaluable support at Davidson and all the GCAT faculty and students who have willingly ventured into uncharted educational territory. GCAT faculty workshops were funded by four NSF grants: DBI-0627478 (2006–2010), DBI-0520908 (2005), DBI-0408386 (2004), and DBI-0305176 (2003). NSF also funded the first microarray scanner for GCAT via grant DBI-0099720.

REFERENCES

Campbell, A. M. (2002). Genomics in the Undergraduate Curriculum: Rocket Science or Basic Science? www.bio.davidson.edu/people/macampbell/ASM/ASM.html (accessed 3 October 2006).

Campbell, A. M., Eckdahl, T. T., Fowlks, E., Heyer, L. J., Hoopes, L.L.M., Ledbetter, M. L., and Rosenwald, A. G. (2006a). Collaborative programs. Genome Consortium for Active Teaching (GCAT). Science *311*, 1103–1104.

Campbell, A. M., Zanta, C. A., Heyer, L. J., Kittinger, B., Gabric, K. M., and Adler, L. (2006b). DNA microarray wet lab simulation brings genomics into the high school curriculum. CBE Life Sci. Educ. *5*, 108–115.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680–686.

Eisen, M. B. (2006). Eisen Lab Software. http://rana.lbl.gov/EisenSoftware.htm (accessed 10 September 2006).

Elgin, S.C.R. (2005). Genomics in Education. http://www.nslc.wustl.edu/elgin/genomics/index.html (accessed 10 August 2006).

Genome Consortium for Active Teaching (2003). GCAT-Listserv (GCAT-L). www.bio.davidson.edu/projects/GCAT/GCAT-L.htm (accessed 10 August 2006).

GCAT (2005). Two GCAT Best Practices Workshops. www.bio.davidson.edu/projects/GCAT/workshop3.html (accessed 10 August 2006).

GCAT (2006a). GCAT DNA Chip Simulations: Dry Lab and Wet Lab Curricula. http://www.bio.davidson.edu/projects/GCAT/HSChips/HSchips.html (accessed 10 August 2006).

GCAT (2006b). GCAT Faculty Members Using DNA Microarrays with Undergraduate Students. http://www.bio.davidson.edu/projects/GCAT/members/main.html (accessed 12 August 2006).

Handelsman, J. *et al*. (2004). Scientific teaching. Science *304*, 521–522.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. Nature *402* (suppl), C47–C52.

Heyer, L. J., and Campbell, A. M. (2004a). MicroArray Genome Imaging and Clustering Tool. www.bio.davidson.edu/MAGIC (accessed 10 August 2006).

Heyer, L. J., and Campbell, A. M. (2004b). Exploring Diauxic Shift Microarray Data with MAGIC Tool. http://gcat.davidson.edu/GCAT/workshop2/derisi_lab.html (accessed 4 September 2006).

Heyer, L. J., Moskowitz, D. Z., Abele, J. A., Karnik, P., Choi, D., Campbell, A. M., Oldham, E. E., and Akin, B. K. (2005). MAGIC tool: integrated microarray data analysis. Bioinformatics *21*, 2114–2115.

Kuldell, N. H. (2006). How golden is silence? Teaching undergraduates the power and limits of RNA interference. CBE Life Sci. Educ. *5*, 247–254.

Kumar, A. (2005). Teaching systems biology: an active-learning approach. Cell Biol. Educ. *4*, 323–329.

National Research Council (2003). BIO 2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

NRC (2005). Facilitating Interdisciplinary Research, Washington, DC: National Academies Press.

Pfund, C. *et al*. (2006). The merits of training mentors. Science *311*, 473–474.

Project Kaleidoscope (2001). Investing in Faculty. http://www.pkal.org/documents/index.cfm?page=3080 (accessed 10 August 2006).

Rettburg, R., and Endy, D. (2006). iGEM—The International Genetically Engineered Machine competition. http://parts2.mit.edu/wiki/index.php/Main_Page (accessed 10 August 2006).

Steen, L. A. (ed.) (2005). Math & Bio 2010, Linking Undergraduate Disciplines, Washington, DC: The Mathematical Association of America.

SyntheticBiology(2006).SyntheticBiologyis.http://syntheticbiology.org/FAQ.html (accessed 10 August 2006).

Tonidandel, S. (2004). GCAT Assessment Online. www.bio.davidson.edu/projects/GCAT/assessment/assess.html (accessed 10 August 2006).

U.S. Department of Education (2006). United States Department of Education List of Postsecondary Minority Institutions. www.ed.gov/about/offices/list/ocr/edlite-minorityinst.html (accessed 20 July 2006).

Teachers' group brings genomics revolution to minority colleges

When the human genome sequence was released in 1999, it meant two things to Edison Fowlks, a biology professor at Hampton University in Virginia.

First, genomics technologies were about to revolutionize science. And second, students and faculty of so-called minority-serving institutions such as Hampton, a historically black college, needed to be part of the revolution.

But where were such institutions going to come up with the funds to train faculty in the new technologies—much less buy microarrays and the scanners needed to read them?

In 2004, Fowlks found an answer when he met fellow biologist A. Malcolm Campbell, who since 2000 had been organizing a program called Genome Consortium for Active Teaching (GCAT) for faculty at small undergraduate institutions. Campbell is himself a researcher at Davidson College in North Carolina, a liberal arts college with 1,700 students.

Campbell had convinced genomics pioneer Pat Brown of Stanford University to donate microarrays, which Campbell then mailed to dozens of other professors. These professors taught students how to do experiments with the chips and then mailed them back to him. Campbell then read data from the chips using a single scanner and sent it back to the professors, who analyzed it with free software written by one of Campbell's colleagues. The only charges for chip users were shipping fees and the cost of the reagents for their experiments—no more than $500.

Fowlks saw the power of the model immediately. "GCAT essentially democratizes genomics," he says. "It allows a consortium of small colleges and universities to do informatics and genomics without all the powerful equipment that major universities have."

Fowlks joined forces with Campbell to expand GCAT's reach. The pair wrote a grant, awarded by the US National Science Foundation, to support a GCAT workshop at Morehouse College in Atlanta in 2005. The agency has committed to funding yearly workshops through 2009; the most recent of these, held this July, trained 40 teachers.

Genome Consortium for Active Teaching

Bring on the revolution: Using donated microarrays and a single scanner, minority faculty and students are jumping into genomics research.

The workshops are open to anyone who teaches undergraduates, with an emphasis on faculty teaching minority students. Since 2003, the Howard Hughes Medical Institute has spent $100,000 each year to buy microchips for the program. This year, GCAT distributed 1,200 chips to 72 teachers at institutions across the nation, from Alaska to Hawaii and Puerto Rico.

This fulfills not only Fowlks's and Campbell's goals, but also those set out by numerous reports on American competitiveness, such as a 2005 National Academies manifesto that calls for the nation to train more minority scientists and engineers.

"The National Academies and so many other groups have said we need to increase diversity in science, and I don't know how that's supposed to happen if we don't reach out to schools that serve large numbers of minorities," Campbell says.

GCAT is already showing results. Scientists such as Consuelo Alvarez at Longwood University in Farmville, Virginia, are publishing genomics research, and students such as Hampton University senior Sabriya Rosemond are getting swept into the genomics revolution.

Rosemond, one of Fowlks's former students, has worked in biology labs for the past two summers and is determined to go into science after she graduates next year. "I want to make science a little browner, like Dr. Campbell and Dr. Fowlks are doing," she says. For the GCAT leaders, that's an even more satisfying benchmark than the growing list of grants and papers that they are helping to produce every year.

Erika Check, San Francisco

Rakai Health Sciences Program

Simple solution: Women and children made swabs needed to test vaginal pH for a clinical study in Uganda.

The kids received their wages in gourmet gummy bears and M&Ms. On a good night, the team would make 200 swabs.

The researchers used the swabs to follow weekly changes in vaginal pH in 311 women over two years. "It's a poor woman's way of doing it," says John Thorp, a gynecologist at the University of North Carolina, who was not involved in the project. "I think that taking the [vaginal] speculum out of it greatly diminishes the cost."

Sullivan and her colleagues used to joke about patenting the swab. But it's too late. In July, New York–based company Vagisil launched its own over-the-counter version, a spatula-shaped 'wand' that measures pH.

The study ended in 2003, and the handmade pH swabs are no longer being used. But the researchers plan to revive their approach if necessary: a tube of pH strips and a box of tongue depressors are still cheaper than the $15 tab for a Vagisil kit.

Cassandra Willyard, New York

Fischer-Vize, *Science* **270**, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol*. **6**, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* **13**, 3822 (1994); M. T. Madireddi *et al.*, *Cell* **87**, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* **266**, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* **58**, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* **174**, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* **47**, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* **178**, 3015 (1996).
37. M. Ho *et al.*, *Cell* **77**, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* **59**, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3*–*6*).

Saccharomyces cerevisiae is an especially favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl–coenzyme A(CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

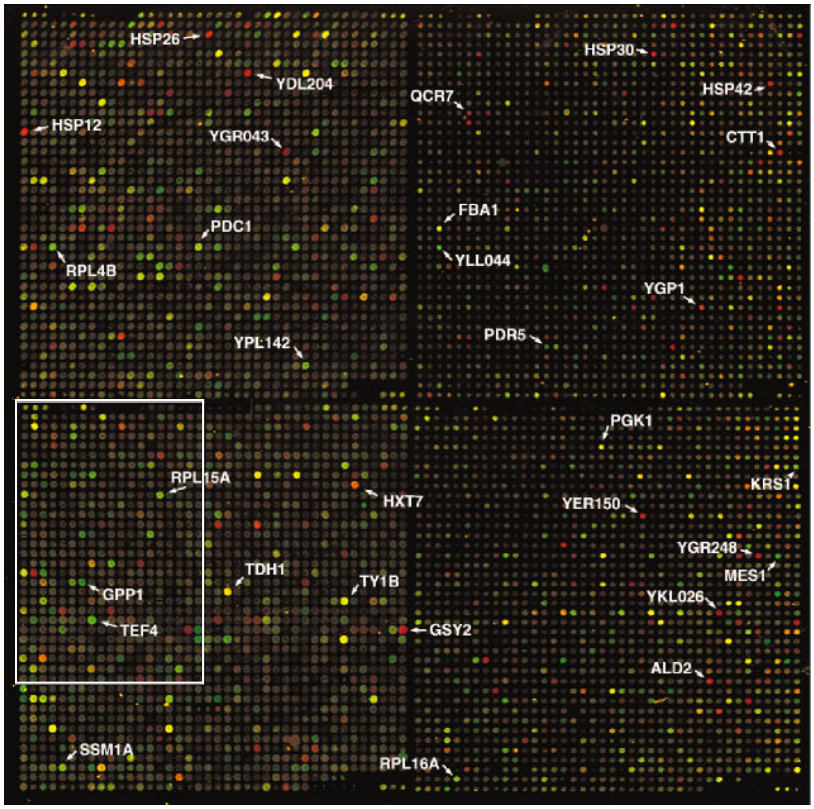


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (*21*–*24*). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex *HAP2,3,4* has been shown to be responsible for induction of several genes important for respiration (*26*–*28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of *HAP2,3,4* (*30*). Indeed, a putative *HAP2,3,4* binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, *HAP2,3,4* binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rpg}) that is recognized by the Rap1 DNA-binding protein (*31*, *32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expression ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37*, *38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

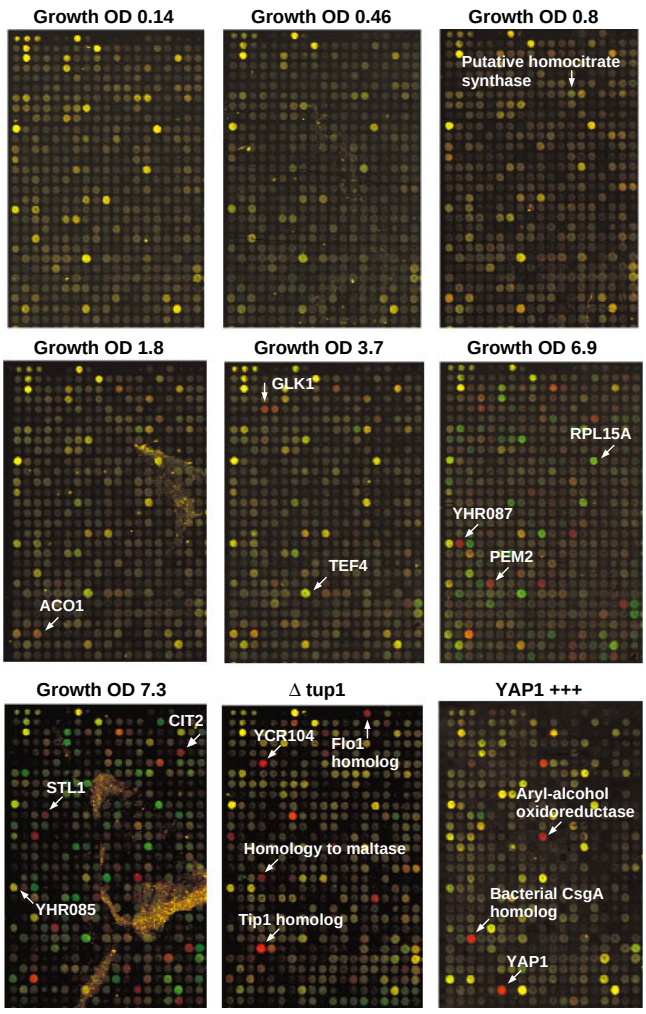


Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (*39*). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (*40*).

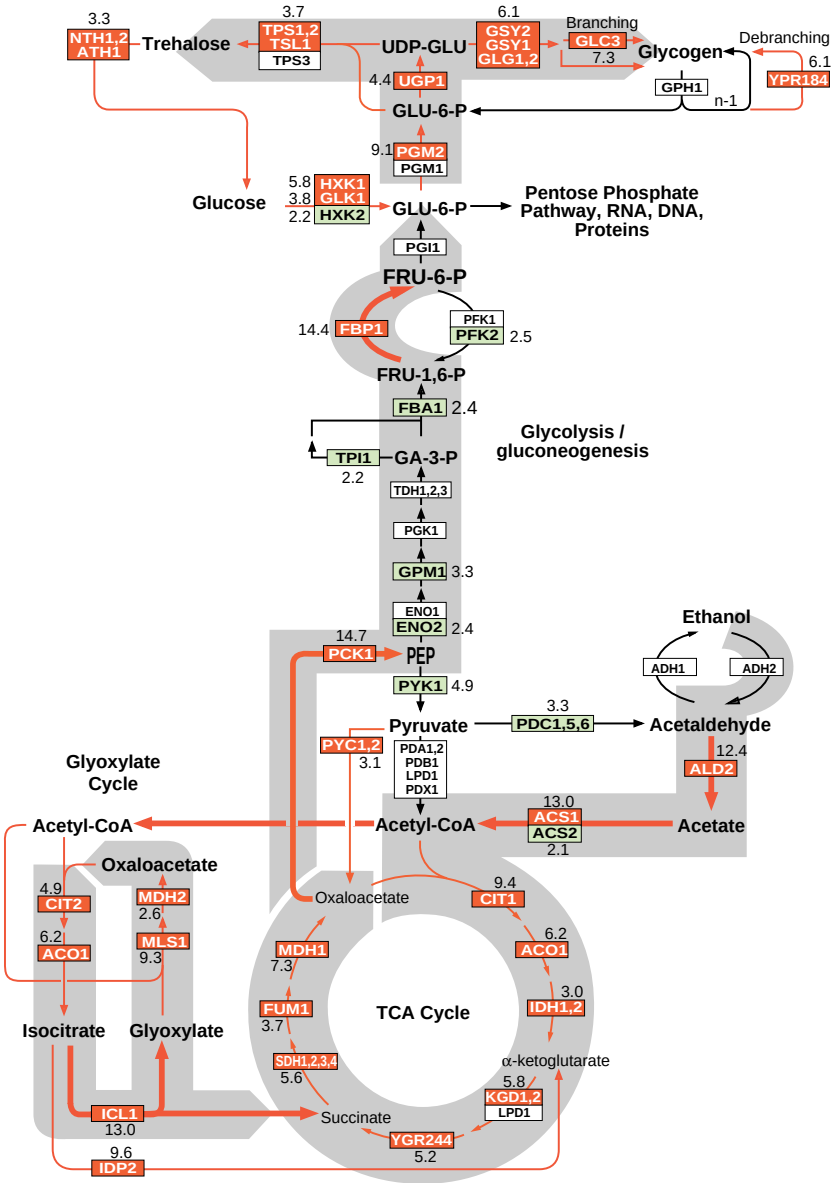


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1*Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1*Δ strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1*Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1*Δ strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1*Δ

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1*Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MATA* strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when *YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46*, *47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

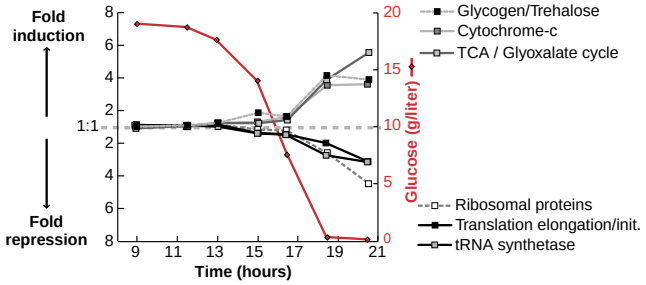


Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and networks. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

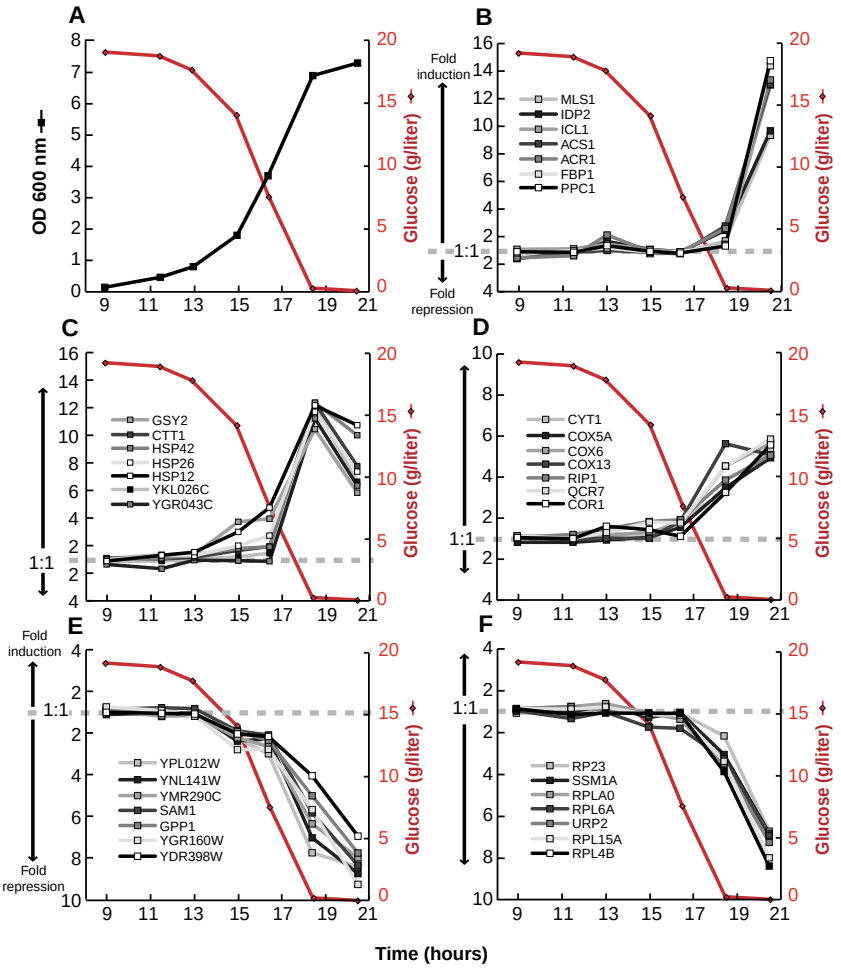


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (**C**) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (**D**) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (**E**) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (**F**) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
5. D. J. Lockhart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996).
6. M. Chee *et al.*, *Science* **274**, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.
11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)$^+$] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).
12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).
16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* **14**, 3613 (1994).
17. S. Kratzer and H. J. Schuller, *Gene* **161**, 75 (1995).
18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* **268**, 12116 (1993).
19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* **242**, 727 (1994).
20. A. Hartig *et al.*, *Nucleic Acids Res.* **20**, 5677 (1992).
21. P. M. Martinez *et al.*, *EMBO J.* **15**, 2227 (1996).
22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* **15**, 6232 (1995).
23. H. Ruis and C. Schuller, *Bioessays* **17**, 959 (1995).
24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* **143**, 1891 (1997).
25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
26. S. L. Forsburg and L. Guarente, *Genes Dev.* **3**, 1166 (1989).
27. J. T. Olesen and L. Guarente, *ibid.* **4**, 1714 (1990).
28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* **13**, 119 (1994).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.
30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* **12**, 363 (1996).
31. D. Shore, *Trends Genet.* **10**, 408 (1994).
32. R. J. Planta and H. A. Raue, *ibid.* **4**, 64 (1988).
33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (*30*), is WACAYCCRTACATYW, with up to three differences allowed.
34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* **15**, 3187 (1995).
35. P. Lesage, X. Yang, M. Carlson, *ibid.* **16**, 1921 (1996).
36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin *et al.*, *Mol. Microbiol.* **20**, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* **23**, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg *et al.*, *J. Biol. Chem.* **271**, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore *et al.*, *Mol. Cell. Biol.* **11**, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T *CTT1* [P. H. Bissinger *et al.*, *ibid.* **9**, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas *et al.*, *J. Biol. Chem.* **266**, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* **223**, 97 (1990); D. Wotton *et al.*, *J. Biol. Chem.* **271**, 2717 (1996)].
37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (*46*).
38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero *et al.*, *Yeast* **11**, 137 (1995)], *MLS1/DAL7* (73% identical) (*20*), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* **10**, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* **11**, 3307 (1991).
40. D. Tzamarias and K. Struhl, *Nature* **369**, 758 (1994).
41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco R I site 124 base pairs before the stop codon of the *TUP1* gene.
43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* **15**, 341 (1995).
44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* **148**, 149 (1994).
45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* **242**, 250 (1994).
46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* **60**, 1783 (1994).
47. A. Muheim *et al.*, *Eur. J. Biochem.* **195**, 369 (1991).
48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* **269**, 32592 (1994).
49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)
51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Cell Biology Education
Vol. 4, 19–23, Spring 2005

Feature

Meeting Report: Synthetic Biology Jamboree for Undergraduates

A. Malcolm Campbell

Davidson College, Biology Department, P.O. Box 7118, Davidson, NC 28035

While most of us have been following news in the fields of genomics, proteomics, bioinformatics, and maybe even systems biology, a new field may have escaped our attention. The field of synthetic biology (the name is derived from an analogy to synthetic chemistry) has recognized itself as a "field" only since about 2002. Synthetic biology has gotten some high-profile attention recently (e.g., Ferber, 2004; Hasty *et al.*, 2002; Hopkin, 2004; Nature Staff, 2004a, b; Pennisi, 2003; Zak *et al.*, 2003), but most people are not aware the field even exists. Synthetic biologists apply engineering principles to genomic circuits to construct small biological devices. The Jamboree, as it was affectionately called, was the culmination of a summer of undergraduate research on five campuses across the United States.[1] The participants shared data, frustrations, lessons learned, and plans for the future. The entire weekend was, to some extent, a pyramid turned upside down. Normally, new fields in biology are explored first by postdocs and graduate students under the watchful eyes of their Primary Investigator (PI) mentors. This National Science Foundation (NSF)–supported Jamboree featured undergraduates (some having just completed one year of college) who were pushing the boundaries of a field so new, its name is subject to debate. This report will highlight some of the interesting research conducted by undergraduates during the summer and early fall of 2004.

Teams of undergraduates spent 10 weeks of their summers blending biology with computer science, engineering, and chemistry (Figure 1). As is often true of young students, many were oblivious of the significance of their efforts before the Jamboree. Only after sharing their stories did they begin to appreciate the magnitude of their summer's efforts. Each group of students had been given a one-phrase directive (design and build a genetically encoded finite state machine), and over the summer, they designed, modeled, built, and tested their constructions. The most interesting presentations were those made by undergraduates. One team had more senior people present, and you could tell they were less candid and less enthusiastic. When the undergraduates spoke, they had a sheen of freshness and personal investment that was infectious and exhilarating.

The teams were composed of diverse sets of students, with only two self-identified as biology majors with previous lab experience. The educational goals of this NSF-funded program were varied and vague: to introduce students to a new field; to encourage them to stay in this field; to increase excitement about research; and to foster cross-disciplinary education and collaboration. Although these goals are difficult to define and assess, they are exactly what the National Research Council's publication *Bio2010* stated the future of biomedical research requires to bring success in the future and a more diverse population to biology (National Research Council, 2003).

BACKGROUND FOR SYNTHETIC BIOLOGY

Any new field evolves from the work of pre-existing fields, but a few seminal papers can be cited as the foundation for synthetic biology. In one such paper, Gardner *et al.* (2000) report the design and construction of a genetic bistable toggle switch in *Escherichia coli* (Figure 2A). The design is simple: two promoters and three genes. When the black gene is active, the gray gene and the reporter gene are silenced. Conversely, when the gray gene is active, so too is the reporter gene, but the black gene is repressed. The gray inducer (IPTG) leads to the production of the reporter protein, green fluorescent protein (GFP), whereas the black inducer (tetracycline) halts production of the reporter GFP (Figure 2B). This simple biological machine might seem like a widget that does nothing in particular, but imagine if the reporter gene were exchanged with a biologically functional gene. Then a production facility could turn the secretion of a biomedical product on and off that otherwise would be toxic to the cells. Or, perhaps the cells could monitor waste sewage from a factory to detect violations of environmental laws.

The "repressilator" by Elowitz and Leibler (2000) set a precedent for naming (fill in the blank–alator) and sophistication. The repressilator is composed of two plasmids (Figure 3A). The larger plasmid contains the oscillatory circuit of three repressors. Each repressor is induced in turn, so the circuit rotates around the plasmid as the previous repressors are degraded by the cell. When TetR is produced, the production of GFP is silenced. Activity of the repressilator

DOI: 10.1187/cbe.04-11-0047
Address correspondence to: A. Malcolm Campbell (macampbell@davidson.edu).

[1] Participating schools: Massachusetts Institute of Technology, California Institute of Technology (Caltech), Boston University, Princeton, University of Texas at Austin.

Figure 1. Participants and mentors at the 2004 Synthetic Biology Jamboree, held on the grounds of the American Academy of Arts and Sciences in Cambridge, MA.

is monitored by observing GFP, which oscillates at a regular interval (Figure 3B). It is worth noting that the periodicity of the GFP cycle was much longer than the periodicity of cell division by the bacteria, which indicates the signaling mechanism outlived the lifetime of any given cell.

THEIR AMAZING MACHINES

Now that you have an idea what synthetic biologists do, I want to share two student constructions with you. The first was produced by a Princeton team that wanted to build a biological equivalent of the children's game called Simon (see http://www.begent.net/games/simon/simongame.htm for an online version). The object of the game is for the user to repeat a pattern of signals that grow in complexity at each successful iteration. What the Princeton team wanted to produce was a set of three bacterial strains that could correctly detect the

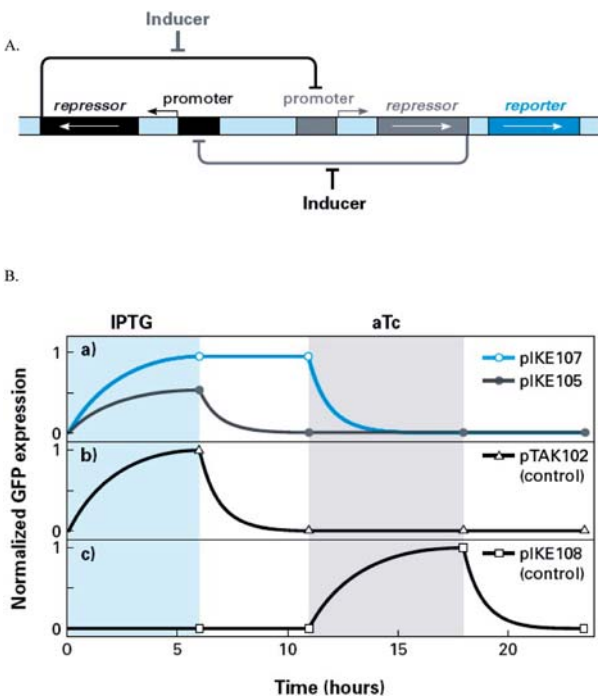


Figure 2. Bistable toggle switch. A. Generic design of a bistable switch that can be flipped one of two ways depending on which inducer is applied. B. Data produced by the final bistable toggle switch (panel a, blue trace) as well as several control constructs (black traces).

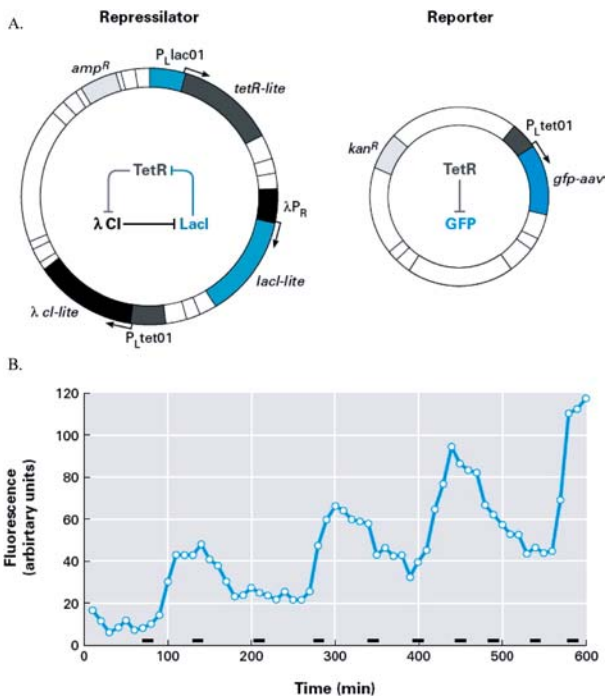


Figure 3. Repressilator. A. The repressilator was designed to produce three repressors in succession, each degrading over time and repressing a different promoter. The progress of the cycle was monitored by the production of GFP, encoded on a separate plasmid and repressed by one of the three repressors on the repressilator plasmid. B. Production of GFP was monitored over time. The black bars at the bottom indicate the time of cell division for a collection of *E. coli* cells monitored through a microscope.

input of three environmental stimuli that had to be delivered in a particular sequence. In addition, like the game, their biological Simon had the capacity to be reset at will (Figure 4).

With the use of BioBricks from the MIT database (http://parts.mit.edu/), the students designed three strains of cells that had three different circuits. The first cell type accepted the input of anhydrotetracycline (aTc) and secreted a molecule in response. Cell type 2 accepted the input of homoserine lactone (HSL) and the secreted molecule from cell type 1 and responded by secreting a new molecule, which was half of the signal required by cell type 3. When the user applied arabinose to cell type 3, which had been signaled by cell type 2, the third and final cell in the chain responded by producing yellow fluorescent protein (YFP). At any step in the process, the students could reset their biological Simon by applying a heat shock, which would destabilize a temperature-sensitive component (cI^{ts}) shared by the last two cell types. The team was not able to build their biological Simon because of problems they had in the construction phase and the YFP gene in particular. The students used parts from the BioBricks library and offered suggestions for ways the BioBricks repository could be improved.

One of the CalTech teams designed and built a strain of yeast that was capable of detecting three concentrations of caffeine. For their design, this team relied on small noncoding RNA switches composed of two domains: an aptamer domain and an antisense domain. Aptamers are nucleic acid molecules that can bind to small ligands with a high degree of specificity. Depending on how the RNA switches were designed, they could activate or inactivate sequence-specific mitochondrial RNAs (mRNAs). The students designed and constructed RNA switches that could detect the ligand caffeine at different concentrations and built two types of switches. One switch destroyed GFP mRNA at high doses of caffeine, whereas the other switch activated YFP mRNA beginning at medium doses of caffeine. The combination of switches produced a cell that glowed green in the presence of low caffeine, green and yellow in medium caffeine, and yellow only in high caffeine. Having proven their device worked under laboratory conditions, the students headed out to their favorite campus source of coffee and tested their device on real-world samples (Figure 5). To everyone's delight, their modified yeast could distinguish decaf, regular, and espresso coffees. As one Jamboree participant noted, combining coffee and yeast metabolites is the dream of every student.

There were additional presentations by students. Some emphasized computer modeling of behaviors and others focused on biological output. For example, some cells were designed and modeled to swarm toward a chemoattractant, signal each other, diffuse away, signal each other, and reswarm. Another team produced cells that were photosensitive and produced a color product. The photosensitive results culminated in the world's first biological photograph of the oft used phrase in computer science, "Hello World."

MEASURING SUCCESS

One goal of the Jamboree was to foster interdisciplinary collaborations. The selection process assured the goal of mixing students from different disciplines. Chemistry and computer science were the two most common majors after engineering. Some of the students had taken a previous course at their home institutions that prepared them for synthetic biology, but this was true for only a small percentage of the summer research students. Therefore, many participants were exposed to a new field during their summer research.

As the summer began, it was uncertain whether the students would enjoy their experience and be influenced to stay in the field of synthetic biology. During the breaks, I talked to several students informally and heard some say how the summer had affected their career interests. A couple of their comments were: "I had some prior research but now I'm more interested

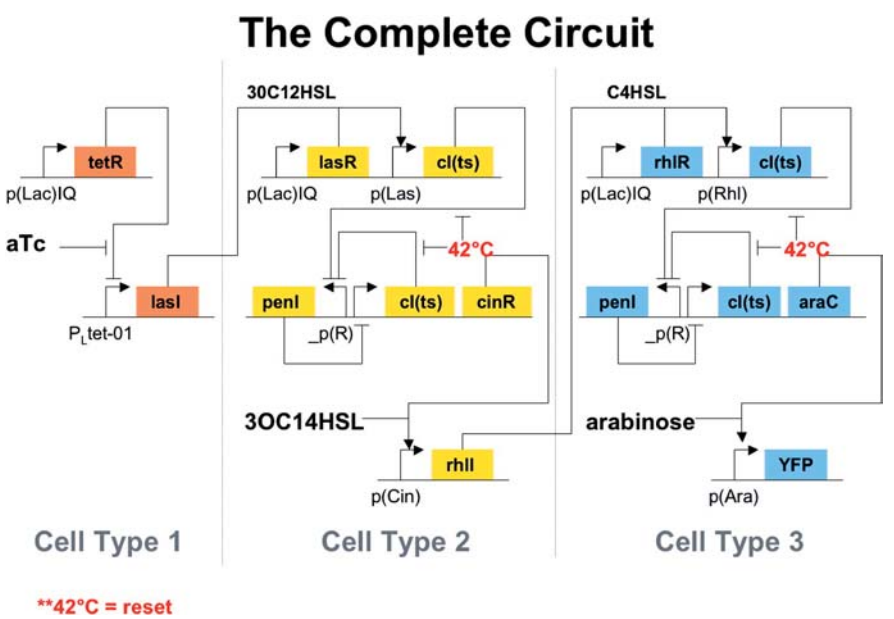


Figure 4. Circuitry for Simon 1.0, designed by a team of undergraduates from Princeton University. The three cell types were part of a pattern of inputs that had to be produced in the correct order for the reporter protein, yellow fluorescent protein (YFP), to be produced.

Figure 5. The team from Caltech constructed yeast cells that were able to distinguish low, medium, and high levels of caffeine. Two reporter proteins, green fluorescent protein (GFP) and yellow fluorescent protein (YFP), signaled which level of caffeine was detected. Shown here are the four student investigators comparing their skills against the caffeine-sensing fungi. Clockwise, from top left: Travis Bayer, Maung Nyan Win, Brandon Rawlings, and Jack Lee.

in biology, specifically in engineering circuits. I am continuing to do research this semester and am working now to make our machine function." and "I gained an appreciation for CS [computer science] and will take some CS courses during the last semester of my senior year as a biology major."

Every one of the summer groups has some students that continue to work on their constructions. This shows real commitment, excitement, and the spark of a researcher in the making. Approximately one-third of all students at the Jamboree were still working on their projects in November. Rather than seeing this as a sign of failed summer work, the students saw continuation on their projects as a challenge worthy of their time. Research is not easy and they know it.

When asked whether the weekend gathering was useful, everyone at my lunch table said absolutely. Before the weekend, they did not realize others were interested in their efforts. They had assumed none of the other groups were having problems and something must be wrong with them for the frustrations and setbacks they faced. Hearing the troubles experience by each group helped individuals gain a better understanding of the expression, "if it were easy, no one would interested." They enjoyed hearing the diverse plans and outcomes from the other groups.

Some lessons learned include the need for clear and ongoing communication. The participants learned that a community is more productive than an individual, uncoordinated effort. They took pride in their work and enjoyed sharing with their peers. Although an electronic discussion board was available, it was not used much, which probably says more about the negative side of electronic communication compared with personal contacts.

CONCLUSION

It is a rare treat to watch the birth of a new island when a volcano rises from the ocean. The Jamboree felt like the intellectual equivalent, with burgeoning students creating fantastic designs and finite state machines. The future of synthetic biology could be very bright. These undergraduates personified the recommendations of *Bio2010* (National Research Council, 2003). They did world-class work, yet their level of training was embryonic. Imagine where they may lead the field in 20 years. I was so impressed with their work that this summer, I too will have my students design, model, and produce simple biological machines. We will begin by reading and designing, but the students will need to settle on a design quickly enough to have time to build their devices.

During the final session of the Jamboree, the group discussed the ethical, legal, and social implications (ELSI) of synthetic biology (see Ferber, 2004; Hopkin, 2004). Considering the ELSI of synthetic biology was new for the undergraduates, although it was a familiar topic for their PIs. The perception of a self-contained, insulated group of scientists is what could put synthetic biology in the same politically charged boat as stem cells, somatic cell cloning, and GMOs; knowledge is trumped by fear every time. All

investigators should link ELSI and education with synthetic biology research if we want it to be funded by the U.S. government.

The Jamboree leaders also need to place a bit more effort in measuring educational outcomes. Educational assessment is awkward and sometimes abhorrent to scientists, but why treat our teaching any less seriously than our science? Would you accept a claim in science without data? If not, then why trust your instinct when data are available? A short survey at the end would provide "summative" data. How many of you will take additional courses in this area? How many of you will take courses in different departments as a result of your experience? How many of you would like to continue your work beyond the summer? How many of you would like to pursue this type of research in graduate school? Would you like to use this as a foundation for an honors thesis? Would you recommend your friends get involved in future summers?

In the end, the students seemed unanimous that the Jamboree should become a national and annual event. It is impressive that students could design cells from BioBricks parts to perform new functions. Perhaps next year, my students can share their results, and more schools will join the fun of the 2005 Jamboree.

ACKNOWLEDGMENTS

I thank Drew Endy for his invitation to join the Jamboree and all the participants for their willingness to share with me. Ron Weiss and Christina Smolke were generous for sharing Figures 4 and 5, respectively.

REFERENCES

BioBricks. (2004). Parts List. MIT Registry of Standard Biological Parts. http://parts.mit.edu/ (accessed 23 November, 2004).

Elowitz, M.B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature *403,* 335–338.

Ferber, D. (2004). Microbes made to order. Science *303,* 158–161.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. Nature *403,* 339–342.

Hasty, J., McMillen, D., and Collins, J.J. (2002). Engineered gene circuits. Nature *240,* 224–230.

Hopkin, K. (2004). Reverse transcript life: the next generation. Engineers and biologists team up to create synthetic biological systems. Scientist *18*(19), http://www.the-scientist.com/yr2004/oct/upfront1_041011.html (accessed 26 November 2004).

Nature Staff. (2004a). Futures of artificial life. Nature *431,* 613.

Nature Staff. (2004b). Starting from scratch. Nature *431,* 624–626.

National Research Council. (2003). Bio2010*:* Transforming Undergraduate Education for Future Research Biologists. Washington, DC: National Academies of Science. http://books.nap.edu/catalog/10497.html (accessed 26 November 2004).

Pennisi, E. (2003). Tracing life's circuitry. Science *302,* 1646–1649.

Zak, D.E., Gonye, G.E., Schwaber, J.S., and Doyle, F.J. III. (2003). Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an *in silico* network. Genome Res. *13*(11), 2396–2405. http://www.genome.org/cgi/content/full/13/11/2396 (accessed 26 November 2004).

DeRisi Paper Analysis Pre-Workshop Problem Set

This assignment can be done during the workhop. We will discuss the questions and answers during the workshop. The material that you need can be found in the article by DeRisi et al. referred to below. A copy is included in your notebook. DeRisi, J. , Iyer, V, and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a global scale. *Science*. 278:680-686 (1997).

1) Please answer the following with regard to your plans to use microarrays in an undergraduate class.

a. How will microarray use relate to the overall goals of your class?

b. What concerns do you have about using this type of experiment in your class?

c. If you have chosen a type of microarray experiment that you would like your class to perform, what kind?

d. How much time are you planning to devote to microarray experiments? To microarray data analysis?

e. What contingencies do you have in mind for problems that may arise with your microarray experiments?

2) We will be preparing total RNA and then we will copy the mRNA (a small minority of the RNA molecules present, about 1-3%) into cDNA.

a. Why won't the noise from the other RNAs drown out the signal from the mRNA? (Hint: we will be using a short molecule of oligo dT (16 T nucleotides in a row) as the primer for reverse transcription, so you may want to consider how eukaryotic mRNAs are processed in the nucleus.) Explain why this method would copy the mRNA but not the rRNA, tRNA, or other RNA molecules present.

b. What would be the effect of contaminating DNA in our RNA preparations upon the cDNA synthesis?

c. Would this method, using oligo dT primed cDNA synthesis, be suitable for use with prokaryotic total RNA preparations?

3) In the formation of the cDNA, we will be incorporating molecules of fluorescent dyes called Cy3 and Cy5. Cy3 fluoresces green and Cy5 fluoresces red (that's not the colors they look like, but the colors of fluorescent light they emit when excited). Each of these dyes will be added to the reaction mixture coupled to dUTP.

a. What base will each of these dye-coupled nucleotides pair with?

b. Since the bulky dye slows the cDNA synthesis, what do you think might be done to incorporate dye but still keep the speed of synthesis up?

c. Diagram how you would set up the experiment so that the Cy3 dye will be attached to the cDNA from the glucose/0.7 cells and the Cy5 dye attached to the cDNA from the ethanol/6.0 cells?

d. If you started with the SAME RNA for the two fluorescently labeled preparations, mixed together the two cDNA preparations, and hybridized them with the same microarray slide containing all the yeast genes, can you think of any reasons why they should not hybridize to each spot with exactly the same green and red fluorescence intensities (ratio of 1.000000)? (No fair answering 'experimental error"; be specific!)

4) Read the 1997 paper by DeRisi on gene expression changes in yeast diauxie. The diauxie means 'two foods' and refers to the use of glucose first, producing ethanol by non-oxygen requiring pathways, and then the aerobic utilization of ethanol. There is no need to change the medium; it naturally happens over time. We will be using total RNA prepared from yeast early in the logarithmic growth period (A_{600} = 0.7) and total RNA prepared from yeast late in growth near stationary phase (A_{600} = 6.0).

a. After you have read the DeRisi paper, using the last set of graphs in the paper and, if you wish the Saccharomyces Genome

Database, at the web site: http://www.yeastgenome.org/ write down three genes you predict would be high and three genes you would predict would be low in expression when using glucose (0.7) and when using ethanol aerobically (6.0). Use the three letter gene name and the yeast gene identifier for each one. (Example: ENO1, YGR254W)

0.7 A_{600}; glucose use	**6.0 A_{600}; aerobic ethanol use**
predicted high expression:	predicted high expression:
predicted low expression:	predicted low expression:

b. For each of these genes given above, briefly describe the molecular function and the biological process in which the gene is involved. You may use the SGD shorthand versions or read the longer descriptive paragraphs and give a longer summary.

5) In microarray data, genes from the same pathway often are co-regulated.

What kinds of mechanisms could result in coordinate transcriptional control of all the genes in the same pathway?

6) In microarray data, duplicate samples do not always provide the same green/red ratios in the output data.

a. What are some of the reasons why they might not? How could these reasons be best addressed in figuring out the meaning of apparent changes in gene expression?

b. What controls can be used to address these problems?

c. How do these problems confound the interpretation of apparent changes in gene expression?

7) Our microarray expression technique generates data in the form of ratios of mRNA signals from cells grown under two conditions. People talk informally about the results as if they show that mRNA is induced under some condition and repressed under some other condition. However, this method does not measure the absolute

amount of mRNA present, nor does it measure how fast it is being synthesized or broken down. Rather, it measures the ratio of the 'standing crop' of the total mRNA present under two different conditions.

a. In your own words, explain what we are actually measuring with this method. Explain why this insight makes the choice of a control sample absolutely critical to the interpretation of the results.

b. What factors besides mRNA concentration affect the level of the functional protein product of a gene?

8) The investigators at Institute for Systems Biology in Seattle have found that only around 60% of the changes they see in mRNA hybridization on microarray chips correspond to changes in the cellular concentration of the same protein encoded by the mRNA. Given what is discussed in question 6 and question 7, why do you think people are still using this technique; i.e. what can it contribute to our understanding compared with other methods for examining regulatory events?

Return to the 2007 GCAT Workshop Page

NOTES, SPECIAL INSTRUCTIONS AND DIRECTIONS