Genomic Neuroscience Tools and Methods

I. Introduction

Two factors are driving a significant change in the way biological and biomedical research is done: the massive increase in the amount of DNA sequence information and the development of technologies to exploit its use. Among the most useful and versatile tools developed for molecular and cellular studies are high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a highly parallel and quantitative fashion. DNA arrays can be employed for many different purposes, most prominently to measure gene expression levels (messenger RNA abundance) for tens of thousands of genes simultaneously.

Obviously, the brain is a complex and inhomogeneous organ containing a large number of different regions and cell types. This does not mean, however, that the brain is too complex to be studied using these new tools. Instead, what is clear is that extra care must be taken, experiments need to be designed with the unique features of the brain in mind, and that array-based measurements need to be applied in combination with other methods. There has been a flood of papers describing the use of genomic technologies to interrogate the brain, demonstrating the feasibility of these approaches. However, a significant obstacle for many wet bench biologists, is access to user friendly tools that help one "mine" data. In the session I will discuss briefly how to apply array-based methods to the study of cells and complex tissue, and describe some special considerations for applying these methods to the study of the brain. I will then mainly focus on describing tools and methods that we and others have developed that make it easy for wet bench biologists to analyze their own data. This chapter is designed to provide you more detail than I will cover in the session. I hope this will aid you understanding the important steps required to perform a neurogenomic experiment and provide you with reference material and helpful tips for use in the laboratory.

II. Global gene expression experiments and DNA arrays - an overview

Dr. Geshwind has described many of the experimental details regarding the use of DNA microarrays. My laboratory mainly uses Affymetrix arrays and so I will briefly describe some of the unique aspects of DNA microarrays that are manufactured by Affymetrix. The most important difference between Affymetrix oligonucleotide arrays and all other arrays is the use of multiple independent "probes" to interrogate a sample.

Basic Definitions Pertaining to Affymetrix GeneChip Microarrays

Probe – A single stranded DNA oligonucleotide designed to be complementary to a specific sequence. Affymetrix GeneChip arrays use oligo probes that are up to 25 bases long. The probes are synthesized directly on the surface of the array using photolithography and combinatorial chemistry.

Probe Cell – A single square-shaped feature on an array containing one type of probe. The size can vary depending on the array type, typically 21 to 100 □m. Each probe cell contains millions of probe molecules so that the sample can be detected quantitatively over a dynamic range.

Perfect Match (PM) – Probes that are designed to be complementary to a reference sequence.

Mismatch (MM) – Probes that are designed to be complementary to the reference sequence except for a homomeric base mismatch at the central position (for example, 13th of a 25 base length probe array). Mismatch probes serve as controls for cross-hybridization.

Probe Pair – Two probe cells, a PM and its corresponding MM. On the probe array, a probe pair is arranged with a PM cell directly above the MM cell.

Probe Set – A set of probe pairs designed to detect a transcript. A probe set usually consists of 11-20 probe pairs. For example, a 20 probe pair set is made up of 20 PMs and 20 MMs for a total of 40 probe cells. In this way, there are 20 independent probes of unique complementary sequence to the gene or EST of interest (PMs) with 20 control probes (MMs) all attempting to detect the same gene.

This figure below provides background on the design of Affymetrix GeneChips (**Figure 1**) and how gene expression measurement data are extracted using Affymetrix software (**Figure 2**).

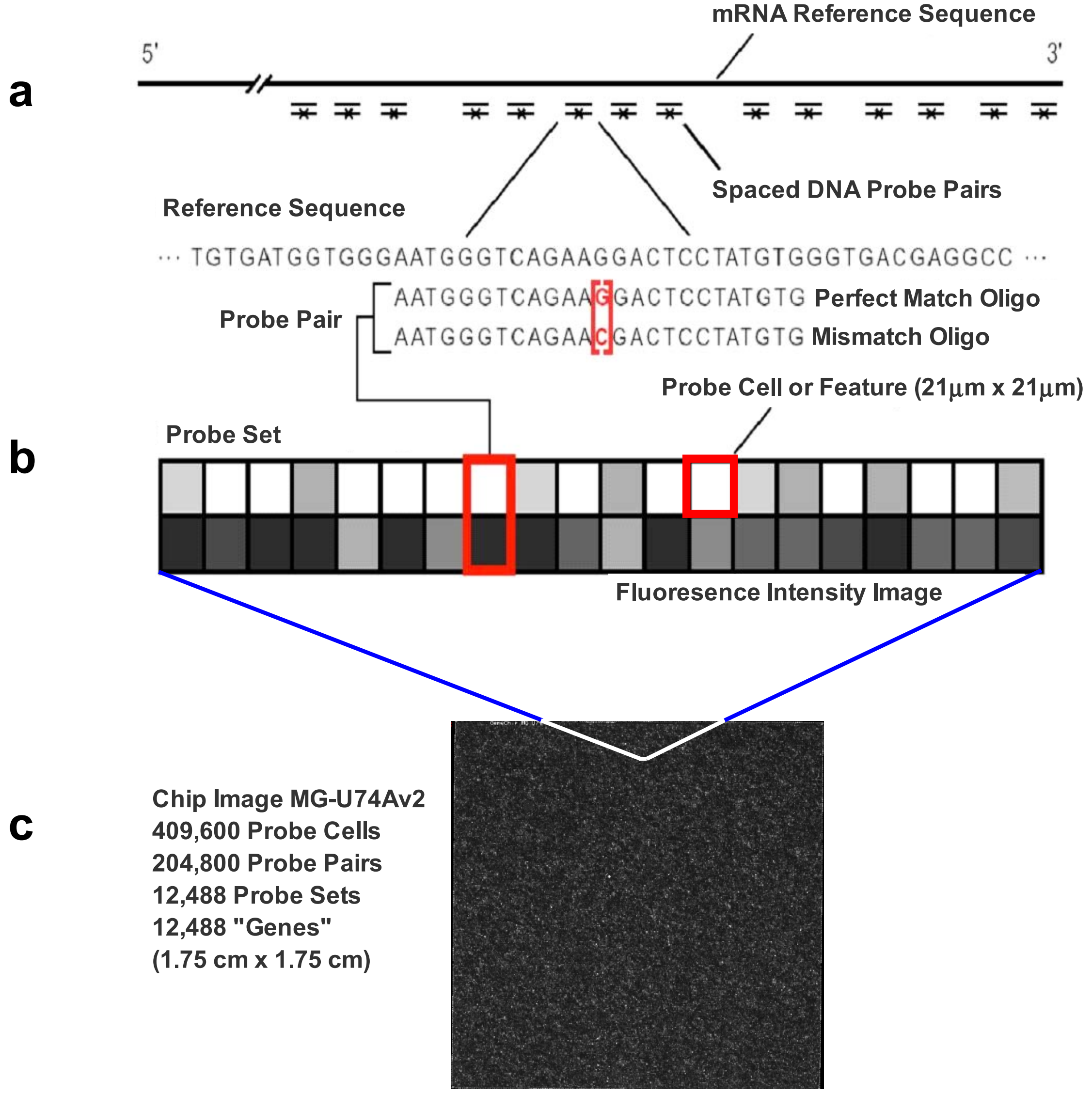
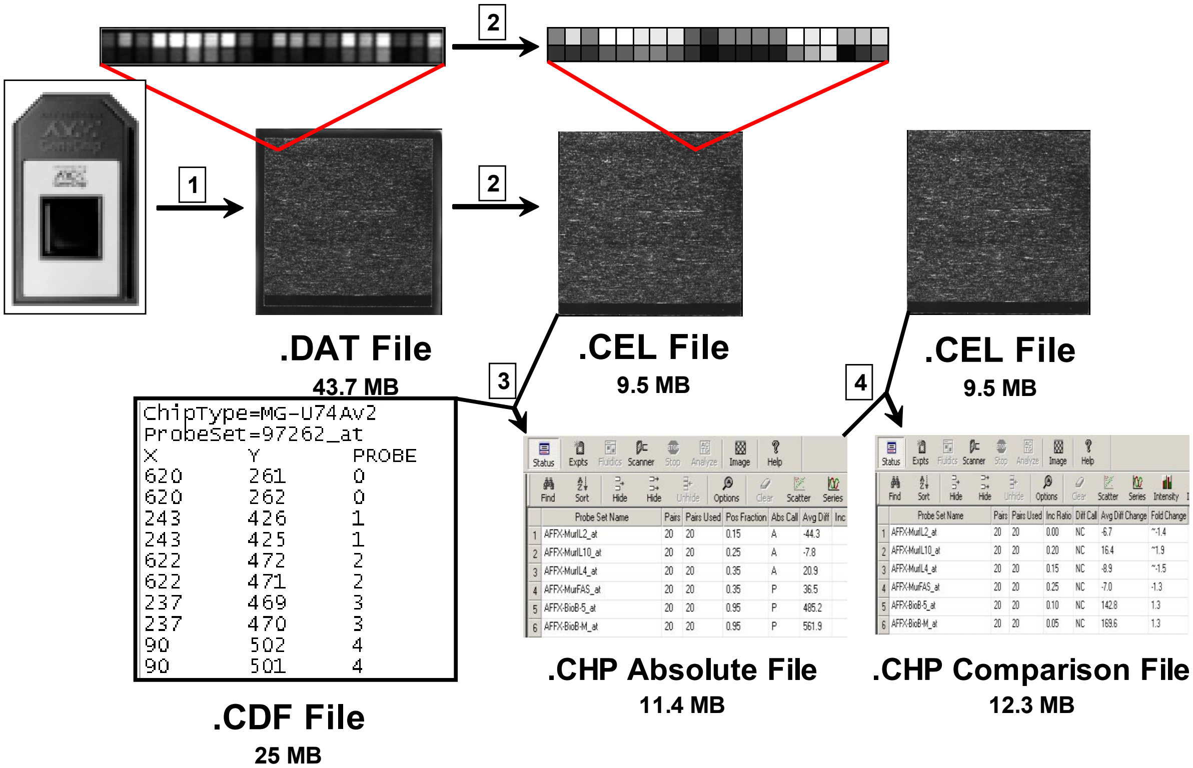

a
mRNA Reference Sequence
5'
3'
Spaced DNA Probe Pairs
Reference Sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
Probe Pair
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Probe Cell or Feature (21µm x 21µm)
b
Probe Set
Fluoresence Intensity Image
c
Chip Image MG-U74Av2
409,600 Probe Cells
204,800 Probe Pairs
12,488 Probe Sets
12,488 "Genes"
(1.75 cm x 1.75 cm)

Figure 1 Affymetrix Chip Design

Shown in **a** is a schematic of a gene or EST from the 5' end to the 3' end. The lines under the gene (——) indicate the regions chosen to design the **perfect match probes**. The perfect match probes are 25 oligonucleotides (oligos) that are designed to be complementary to the mRNA reference sequence. In general, there are 11 to 20 unique perfect match probes chosen to represent a single gene. Shown below the perfect match probes are the **mismatch probes** (—x—). The mismatches are also 25 oligos designed to be complementary to the reference sequence except for a homomeric base mismatch at the central position (for example, the 13th oligonucleotide of a 25 length probe shown in red). The mismatch probes are used as controls for cross-hybridization. Shown in **b** is an idealized **probe set**. Each square represents a **probe cell** on the array. Each probe cell is 21µm by 21µm and contains millions of copies of the 25 oligo probe (i.e. the perfect match or mismatch). Each perfect match probe cell is spatially situated directly above its' partner mismatch and together they constitute a **probe pair**. A probe set is designed with multiple probes pairs per gene. These multiple probe pairs are used for sensitivity, confidence, quantitative precision, and reduction of false positives. **c** is a chip image and the relative scale of a single probe set. The image contains over 400,000 data points (probe cells) that allow one to interrogate over 12,000 genes (probe sets).

In its raw form, the data from one array constitutes over 20 million data points (approximately 40MB per array).

In order to assess mRNA presence and its quantitative abundance, the raw data must be processed through a series of computational steps as shown in **Figure 2**.



```
ChipType=MG-U74Av2
ProbeSet=97262_at
X        Y        PROBE
620      261      0
620      262      0
243      426      1
243      425      1
622      472      2
622      471      2
237      469      3
237      470      3
90       502      4
90       501      4
```

	Probe Set Name	Pairs	Pairs Used	Pos Fraction	Abs Call	Avg Diff	Inc
1	AFFX-MurIL2_at	20	20	0.15	A	-44.3	
2	AFFX-MurIL10_at	20	20	0.25	A	-7.8	
3	AFFX-MurIL4_at	20	20	0.35	A	20.9	
4	AFFX-MurFAS_at	20	20	0.35	P	36.5	
5	AFFX-BioB-5_at	20	20	0.95	P	485.2	
6	AFFX-BioB-M_at	20	20	0.95	P	561.9	

	Probe Set Name	Pairs	Pairs Used	Inc Ratio	Diff Call	Avg Diff Change	Fold Change
1	AFFX-MurIL2_at	20	20	0.00	NC	-6.7	~-1.4
2	AFFX-MurIL10_at	20	20	0.20	NC	16.4	~1.9
3	AFFX-MurIL4_at	20	20	0.15	NC	-8.9	~-1.5
4	AFFX-MurFAS_at	20	20	0.25	NC	-7.0	-1.3
5	AFFX-BioB-5_at	20	20	0.10	NC	142.8	1.3
6	AFFX-BioB-M_at	20	20	0.05	NC	169.6	1.3

Figure 2 Data Analysis Flow using Commercial Affymetrix Software

Shown is the current data analysis flow using the Affymetrix MAS software, starting with scanning the chip and ending with determining differential gene expression. Step **1**| The Affymetrix arrays are scanned with an Agilent GeneArray argon-ion laser scanner. As the surface of the array is scanned, a photomultiplier tube collects and converts the fluorescence emissions into electrical current. This electrical current is converted into numeric values through an analog to digital converter to create the multi-pixilated raw image (.DAT file). Shown above the .DAT file is an enlarged example of a probe set after hybridization. Step **2**| The MAS software converts each multi-pixilated probe cell to a single intensity value thus transforming the raw image file (.DAT file) into a feature by feature flat file (.CEL file). The probe cell feature is scanned at a resolution of 3µm per pixel resulting in 7 pixels by 7 pixels for every probe cell for a total of approximately 49 pixels per probe cell. Taking the 75th percentile of the signal distribution for these 49 pixels creates a single intensity value for every probe cell. The single intensity value is representative of the number of targets (messenger RNA) hybridizing to multiple copies of a particular probe (Lockhart, 1997). Shown above the .CEL file is an enlarged example of a probe set after conversion to single intensity values. The feature by feature flat file (.CEL file) is now composed of X and Y coordinates and a single intensity value for each probe cell Step **3**| Once the raw image (.DAT file) has been converted into a feature by feature flat file (.CEL file) with a single intensity value for each probe cell, it is now possible to compute on the file and determine the qualitative and quantitative information with an analysis algorithm. The MAS software uses a map file (.CDF) to determine the X and Y location of each probe cell and its corresponding probe set. As shown in the .CDF file in the figure, the probe set 97262_at from chip MG-U74Av2 has 8 probe cells located at various XY locations. Using the map file to find where certain genes and ESTs are represented on the array, the software creates an absolute analysis output file (.CHP) with the results for a single array automatically displayed in a specific window. The most important data provided in the absolute analysis file is the qualitative call of "present" or "absent" and the signal intensity for each probe set Step **4**| In order to determine differential gene expression between two arrays, the MAS software performs a comparison analysis. The software uses the .CEL file of one array as the experimental file and the .CHP file of another array as the baseline to determine relative signal differences and signal ratios (fold changes) between the two arrays. Again, the results are placed into a comparison output file (.CHP) automatically displayed in a specific window.

III. Data analysis algorithms from absolute chip analysis and comparison chip analysis

It will always be the case that there will be room for improvement and redesign about the specific implementations of any analysis method (algorithm). Regardless of the specific details, an obvious point is that it is very important to determine not just signal strengths, but whether the signals (or signal changes) are due to the gene for which the probes were designed. This type of analysis is made possible by the use of multiple independent probes for each gene (feature by feature detail). And although the use of simple average signals and fold-changes (ratios) will work in some instances, this common approach is often inadequate and highly discouraged because it can result in an increase in the false positive rate, while at the same time sacrificing sensitivity. What is most important for the biologist to consider is whether or not the method has been PROVEN to work emprically by confirmation of the ENTIRE data set (e.g. not simply the 10 best candidates) using independent methods such as northern blotting or PCR.

In the following section, I will briefly describe some of the steps involved in analyzing array data that takes into consideration analysis methods used in the original papers describing the Affymetrix arrays as well as some of the new statistics that are used in the updated algorithms in MAS 5.0. In addition, several groups have also written public domain packages that use some but not all of these parameters. It is important to understand however that these analysis methods are separate entirely from how one then "filters" analyzed data to determine what gene or genes are changed in a specific experiment where multiple arrays have been analyzed. For this purpose, we have designed an easy to use tool (BullFrog) that makes it easy to further analyze data and to perform data triage so that you as the biologist can assign meaning to the numbers derived from the analysis algorithm and this will be the subject of a later section.

There are three key questions that are asked from any absolute analysis of gene expression data:

Is there a signal from the hybridization of sample?
Is the signal due to the mRNA that the probe set on the array was designed for?
What is the signal strength of the hybridized mRNA?

The first two questions are answered using a **qualitative call that measures whether the signal strengths across the multiple probes in a probe set are consistently greater than background and cross hybridization**. The last question is answered by calculating a weighted average of the signal over the multiple probe pairs in a probe set. **The use of both a qualitative call and quantitative measurement reduces the risk of erroneously assigning a gene as detectable or "present" while maintaining sensitivity to rare mRNAs.**.

Similar logic is used to determine if a gene is differentially expressed between two samples. Once the array has been analyzed and scaled, it is possible to compare the array data to any other array data of the same type (i.e. MG-U74Av2 arrays) to look for differential gene expression. The three questions that are asked for any comparison analysis are:

Are the levels of gene expression on Array 1 statistically different than on Array 2?
What are differences in gene expression levels between Array 1 and Array 2?
What are the ratios of expression levels between Array 1 and Array 2?

A. Background Calculation

The first step to analyze array data is to calculate the background. The background calculation is a measurement of the signal intensity caused by auto-fluorescence of the array surface and nonspecific binding of target or stain molecules. One of the primary factors used to check data quality is the array background.

B. Noise Calculation

Another primary factor used to check data quality is the array noise. The noise is a measure of the small variations in the digitized signal observed by the scanner as it samples the probe array's surface. A high level of noise can indicate a low quality hybridization or a manufacturing problem with the chip.

C. Number of PM and MM Saturated

Some of the other factors used to assess data quality are the number of PMs and MMs that have saturated signals.

D. Scaling

Even if researchers have utilized standard experimental protocols, it is necessary to normalize the data because non-biological factors can contribute to the variability of the data. Differences of non-biological origin consist of variations in the amount and quality of the sample hybridized on the array, the amount of stain applied, or other experimental variables that may contribute to an overall variability in hybridization intensities. In order to reliably compare data from multiple arrays these differences of non-biological origin must be minimized through a scaling factor. In our lab a scaling factor is multiplied to the experimental output to make the experimental output's average intensity equal to an arbitrary target intensity (default = 200). Scaling allows a number of experiments to become normalized to one target intensity, allowing comparison between any two experiments.

E. Qualitative and Quantitative Metrics

To further analyze the array data several qualitative and quantitative metrics are performed on the data to examine whether the gene is detectable by the array and at what level the gene is being expressed. These are the metrics provided by the analysis program that calls a gene Present, Absent, Increased, Decreased. In the next section I will briefly describe in general terms the types of statistics that can be used to make the "calls".

E1: Positive Fraction

To determine if a gene is detectable, quantitative measures and statistics are used. The positive fraction is a measure of the fraction of probe pairs in which the PM probe cells have hybridized with a specific target to a greater level than the corresponding MM control for a particular probe set.

E2: Binomial Distribution p-value

The binomial distribution is used as one of the four statistical tests to assess whether the population of PM signals is greater than the population of MM signals in a probe set. The binomial distribution describes the possible number of times that a particular event will occur in a sequence of observations. The binomial distribution is used when there are a fixed number of tests or trials, when the outcomes of any trial are only success or failure, when trials are independent, and when the probability of success is constant throughout the experiment.

E3: Student's Paired Two-Tailed T-Test p-value

The student's t-test is another statistic used to assess the difference between the PM population and MM population in a probe set. It is one of the most commonly used techniques for testing a hypothesis on the basis of a difference between sample means.

E4: Wilcoxon Signed-Rank Test p-value using Absolute Differences

Like the t-test for correlated samples, the Wilcoxon signed-ranks test applies to two sample designs involving repeated measures and matched pairs. Beginning with a set of paired values (PM and MM), the Wilcoxon signed rank test performs the following: takes the absolute difference $|PM – MM|$ for each pair; omits from consideration those cases where $|PM – MM| = 0$; ranks the remaining absolute differences, from smallest to largest, employing tied ranks where appropriate; assigns to each such rank a "+" sign when $PM – MM > 0$ and a "-" sign when $PM – MM < 0$; calculates the sum of the "+" ranks and the sum of the "-" ranks. Using the larger of the summed ranks the z-ratio along of the associated two-tailed probability (p-value) is calculated.

E5: Wilcoxon Signed-Rank Test p-value using Relative Differences

The Wilcoxon signed-rank using the relative difference uses the same statistical logic as above except instead of using the absolute difference, it uses a relative difference.

E6: The Absolute Call

The absolute call is a qualitative determination of whether the gene or EST for the probe set is detectable. Default metrics and thresholds have established to determine the call through empirical testing for the Affymetrix software.

E7: Signal

In order to calculate an average signal across an entire probe set, several methods have been used including a trimmed mean in the MAS 4.0 program and more recently is based on the 65th percentile

E8: Difference Call
The difference call is part of the comparison analysis that compares data from two probe arrays to determine whether the expression level of each gene or EST has changed. The difference call is a qualitative determination of whether the signal difference between a probe set on chip 1 is consistently different than the signal on chip 2. The difference call can be "Increase", "Marginal Increase", "No Change", "Decrease" and "Marginal Decrease" in chip 1 with reference to chip 2. Default metrics and thresholds have been established to determine the call through empirical testing.

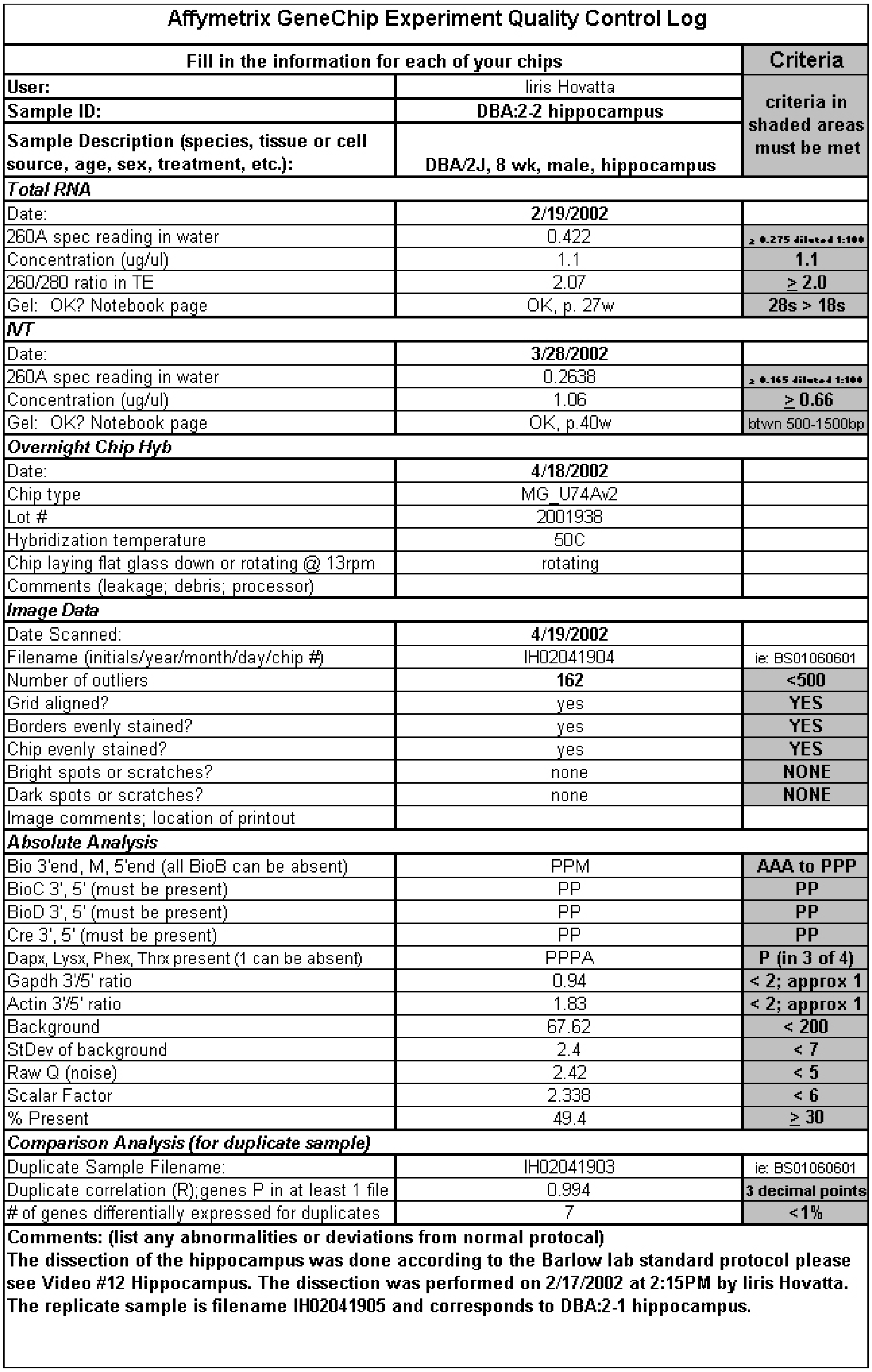
E9: Signal Difference
In addition to the qualitative difference call discussed above there are also quantitative data calculated in the comparison analysis. The first quantitative data calculated is a simple difference between the probe set signals.

E10: Fold Change (Ratio)
The second quantitative data calculated in the comparison analysis is a fold change or ratio of the probe set signals. Both the difference and the ratio are informative as to the level of expression difference between to different sample chips.

Once you have generated your sample, done your microarray and analyzed the array using MAS software or other types of freeware, it is critical to then look at the data BEFORE pursuing any further data analysis. We call this step, DATA TRIAGE as described in the next section.

IV. Data Quality Control-Do this BEFORE filtering your data to get a list of candidate genes!
An important part of any experiment is to be sure to track all the information associated with a single array experiment. Shown in **Figure 3** is the current form used to track experimental information called the Experiment Quality Control Log.

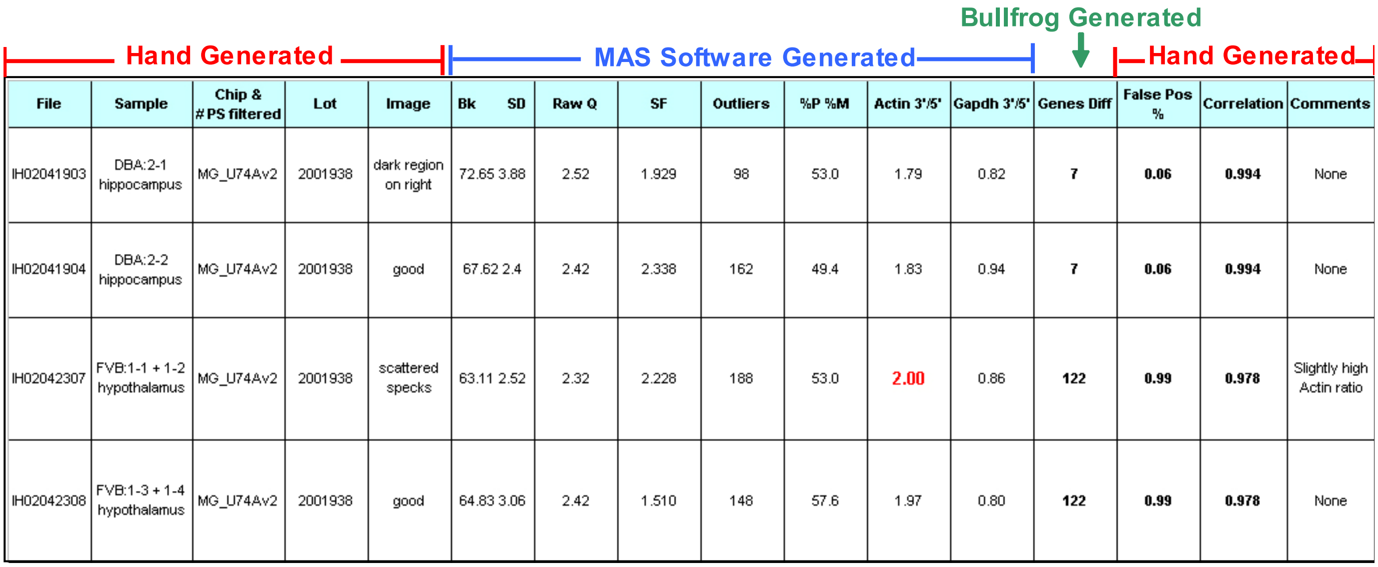
Figure 3 Experiment Quality Control Log



Affymetrix GeneChip Experiment Quality Control Log		
Fill in the information for each of your chips		**Criteria**
User:	liris Hovatta	**criteria in shaded areas must be met**
Sample ID:	**DBA:2-2 hippocampus**	
Sample Description (species, tissue or cell source, age, sex, treatment, etc.):	**DBA/2J, 8 wk, male, hippocampus**	
Total RNA		
Date:	**2/19/2002**	
260A spec reading in water	0.422	≥ 0.275 diluted 1:100
Concentration (ug/ul)	1.1	**1.1**
260/280 ratio in TE	2.07	**≥ 2.0**
Gel: OK? Notebook page	OK, p. 27w	**28s > 18s**
IVT		
Date:	**3/28/2002**	
260A spec reading in water	0.2638	≥ 0.165 diluted 1:100
Concentration (ug/ul)	1.06	**≥ 0.66**
Gel: OK? Notebook page	OK, p.40w	btwn 500-1500bp
Overnight Chip Hyb		
Date:	**4/18/2002**	
Chip type	MG_U74Av2	
Lot #	2001938	
Hybridization temperature	50C	
Chip laying flat glass down or rotating @ 13rpm	rotating	
Comments (leakage; debris; processor)		
Image Data		
Date Scanned:	**4/19/2002**	
Filename (initials/year/month/day/chip #)	IH02041904	ie: BS01060601
Number of outliers	**162**	**<500**
Grid aligned?	yes	**YES**
Borders evenly stained?	yes	**YES**
Chip evenly stained?	yes	**YES**
Bright spots or scratches?	none	**NONE**
Dark spots or scratches?	none	**NONE**
Image comments; location of printout		
Absolute Analysis		
Bio 3'end, M, 5'end (all BioB can be absent)	PPM	**AAA to PPP**
BioC 3', 5' (must be present)	PP	**PP**
BioD 3', 5' (must be present)	PP	**PP**
Cre 3', 5' (must be present)	PP	**PP**
Dapx, Lysx, Phex, Thrx present (1 can be absent)	PPPA	**P (in 3 of 4)**
Gapdh 3'/5' ratio	0.94	**< 2; approx 1**
Actin 3'/5' ratio	1.83	**< 2; approx 1**
Background	67.62	**< 200**
StDev of background	2.4	**< 7**
Raw Q (noise)	2.42	**< 5**
Scalar Factor	2.338	**< 6**
% Present	49.4	**≥ 30**
Comparison Analysis (for duplicate sample)		
Duplicate Sample Filename:	IH02041903	ie: BS01060601
Duplicate correlation (R);genes P in at least 1 file	0.994	**3 decimal points**
# of genes differentially expressed for duplicates	7	**<1%**
Comments: (list any abnormalities or deviations from normal protocal) **The dissection of the hippocampus was done according to the Barlow lab standard protocol please see Video #12 Hippocampus. The dissection was performed on 2/17/2002 at 2:15PM by liris Hovatta. The replicate sample is filename IH02041905 and corresponds to DBA:2-1 hippocampus.**		

Present in the logs are the data parameters that allow for the quality of the chip to be assessed. Shown in the left hand column are the criteria that must be met for a chip to be considered high quality. If the data in the middle column is outside of the ranges displayed in the left hand column, the chip would be of questionable quality. If total RNA parameters are questionable, the sample may need to be completely regenerated. If IVT (in-vitro transcription) parameters are questionable, cRNA may need to be regenerated from stored total RNA. If chip parameters are questionable, the sample may need to be reused and hybridized to a new chip or the same chip may need to be rescanned. In addition to quality control information, other associated information is kept in this log such as the page numbers for gel photos, the reference for the dissection (this example has a video dissection), the scientist that performed the dissection, the time of the dissection, and the replicate sample filename. Lastly, the sample name DBA:2-2 is also cross-referenced with a sample tracking book that has information on how all sample animals were housed and maintained.

In order to assess, in an experimental framework, if all the arrays are of high quality and worthy of further analysis, data triage sheets have been used. **Importantly, the ability to view multiple metrics for quality control purposes over a large number of chips is not readily available in commercial packages so you must do this on your own.** We built a macro that allows us to take information from various sources and generate a **Data Triage Sheet (see Figure 4)**. This type of data triage essentially places pertinent information from the experimental quality QC logs into a single location for all the chips associated with an experiment.


Bullfrog Generated
Hand Generated
MAS Software Generated
Hand Generated

File	Sample	Chip & # PS filtered	Lot	Image	Bk SD	Raw Q	SF	Outliers	%P %M	Actin 3'/5'	Gapdh 3'/5'	Genes Diff	False Pos %	Correlation	Comments
IH02041903	DBA:2-1 hippocampus	MG_U74Av2	2001938	dark region on right	72.65 3.88	2.52	1.929	98	53.0	1.79	0.82	**7**	**0.06**	**0.994**	None
IH02041904	DBA:2-2 hippocampus	MG_U74Av2	2001938	good	67.62 2.4	2.42	2.338	162	49.4	1.83	0.94	**7**	**0.06**	**0.994**	None
IH02042307	FVB:1-1 + 1-2 hypothalamus	MG_U74Av2	2001938	scattered specks	63.11 2.52	2.32	2.228	188	53.0	2.00	0.86	**122**	**0.99**	**0.978**	Slightly high Actin ratio
IH02042308	FVB:1-3 + 1-4 hypothalamus	MG_U74Av2	2001938	good	64.83 3.06	2.42	1.510	148	57.6	1.97	0.80	**122**	**0.99**	**0.978**	None

Figure 4 Data Triage Sheet

Shown in Figure 4 is a data triage sheet for an experiment designed to analyze differential gene expression between mouse hippocampus and mouse hypothalamus. A macro in Excel was made to extract all of this information from a series of Experiment QC logs (see Figure 5) and place them into the table format. Shown are four Affymetrix arrays from two sets of replicates (hippocampus 1, hippocampus 2, hypothalamus 1, and hypothalamus 2), and the associated analysis information to determine the quality of the data. The hippocampus replicates show data that are all within an acceptable range and the correlation coefficients between replicates is high with few genes passing the filtering criteria (7 genes out of 12,488 or a false positive rate of 0.05%). In contrast, the hypothalamus replicates are not as good. For example, there are higher Actin 3'/5' ratios potentially indicative of poor RNA quality (2.00 in red) and lower correlation coefficients; this results in a higher false positive rate (122 genes out of 12,488 (1.0%)). The data triage sheet allows the researcher to quickly scan over a table and assess the data quality for all the arrays in a certain experiment. However, in order to assess data quality, you as the researchers must manually assemble this data using several different programs (shown in red and green). As displayed above, only 7 of the 16 metrics are generated automatically by the Affymetrix MAS software (shown in blue). However, it is well worth the effort.

Examples of QC Criteria for Various Array Types

The following is a list of data quality measurements that were determined based on data from the Barlow laboratory for human, chimpanzee, macaque and mouse and from the CRPF for rat.

RG_U74A (Rat sample on Rat arrays)

RG_U74A Range for good chips 1 round IVT:

Outliers	0 - 400
Background	50 - 150
Standard Deviation of Background	1 – 4.75
Raw Q	1.77 – 4.49
Scaling Factor	1.012 – 2.043
Percent Present	47% – 52%
3'/5' Actin ratio	1.03 - 1.41
3'/5' Gapdh ratio	1.07 - 1.31

RG_U74A Range for questionable chips 1 round IVT:

Outliers	> 300
Background	> 175
Standard Deviation of Background	3 - 7
Raw Q	> 3.6
Scaling Factor	1.884 – 3.771
Percent Present	44% – 48%
3'/5' Actin ratio	1.15 - 1.71
3'/5' Gapdh ratio	1.17 - 1.29

RG_U74A Range for unusable chips 1 round IVT:

Outliers	> 400
Background	> 225
Standard Deviation of Background	> 7
Raw Q	2.92 - 4.32
Scaling Factor	> 2.368
Percent Present	< 44%
3'/5' Actin ratio	1.2 - 5.63
3'/5' Gapdh ratio	1.34 - 8.41

MG_U74Av2 (Mouse sample on Mouse arrays)

MG_U74Av2 Range for good chips 1 round IVT:

Outliers	0 - 500
Background	50 - 100
Standard Deviation of Background	<7
Raw Q	<3.4
Scale Factor	<3.4
Scaling Factor	Depends on the tissue, generally >45%
Percent Present	<2
3'/5' Actin ratio	<2

MG_U74Av2 Range for questionable chips 1 round IVT:

Outliers	400-600
Background	100-300
Standard Deviation of Background	7-10
Raw Q	3.6-8
Scaling Factor	3.4-5
Percent Present	40-48%
3'/5' Actin ratio	2.0-2.5
3'/5' Gapdh ratio	2.0-2.5

MG_U74Av2 Range for bad chips 1 round IVT:

Outliers	>600
Background	>300
Standard Deviation of Background	>10
Raw Q	>8
Scaling Factor	>5
Percent Present	<40%
3'/5' Actin ratio	>2.5
3'/5' Gapdh ratio	>2.5

HG_U95Av2 (Human sample on Human Arrays)

HG_U95Av2 Range for good chips 1 round IVT (Human):

Outliers	0 - 500
Background	50 - 100
Standard Deviation of Background	1-5
Raw Q	2-3.5
Scaling Factor	0.5-3.5
Percent Present	50-60%
3'/5' Actin ratio	1-2.5
3'/5' Gapdh ratio	0.9-1.5

HG_U95Av2 Range for questionable chips 1 round IVT (Human):

Outliers	500-1000
Background	100-150
Standard Deviation of Background	5-10
Raw Q	3.5-7
Scaling Factor	4-6
Percent Present	45-50%
3'/5' Actin ratio	2.5-3.5
3'/5' Gapdh ratio	1.5-2.5

HG_U95Av2 Range for bad chips 1 round IVT (Human):

Outliers	>1000
Background	>150
Standard Deviation of Background	>10
Raw Q	>7
Scaling Factor	>6
Percent Present	<45%
3'/5' Actin ratio	>3.5
3'/5' Gapdh ratio	>2.5

HG U95Av2 (Chimp sample on Human arrays)

HG_U95Av2 Range for good chips 1 round IVT (Chimp):

Outliers	0 - 500
Background	50 - 100
Standard Deviation of Background	1-5
Raw Q	2-3.5
Scaling Factor	0.5-4
Percent Present	45-55%
3'/5' Actin ratio	1-2.5
3'/5' Gapdh ratio	0.9-1.5

HG_U95Av2 Range for questionable chips 1 round IVT (Chimp):

Outliers	500-1000
Background	100-150
Standard Deviation of Background	5-10
Raw Q	3.5-7
Scaling Factor	4-6
Percent Present	40-45%
3'/5' Actin ratio	2.5-3.5
3'/5' Gapdh ratio	1.5-2.5

HG_U95Av2 Range for bad chips 1 round IVT (Chimp):

Outliers	>1000
Background	>150
Standard Deviation of Background	>10
Raw Q	>7
Scaling Factor	>6
Percent Present	<40%
3'/5' Actin ratio	>3.5
3'/5' Gapdh ratio	>2.5

HG_U95Av2 (Macaque sample on Human arrays)

HG_U95Av2 Range for good chips 1 round IVT (Macaque):

Outliers	0 - 500
Background	50 - 100
Standard Deviation of Background	1-5
Raw Q	2-3.5
Scaling Factor	0.5-4
Percent Present	35-45%
3'/5' Actin ratio	1-2.5
3'/5' Gapdh ratio	0.9-1.5

HG_U95Av2 Range for questionable chips 1 round IVT (Macaque):

Outliers	500-1000
Background	100-150
Standard Deviation of Background	5-10
Raw Q	3.5-7
Scaling Factor	4-6
Percent Present	30-35%
3'/5' Actin ratio	2.5-3.5
3'/5' Gapdh ratio	1.5-2.5

HG_U95Av2 Range for bad chips 1 round IVT (Macaque):

Outliers	>1000
Background	>150
Standard Deviation of Background	>10
Raw Q	>7
Scaling Factor	>6
Percent Present	<30%
3'/5' Actin ratio	>3.5
3'/5' Gapdh ratio	>2.5

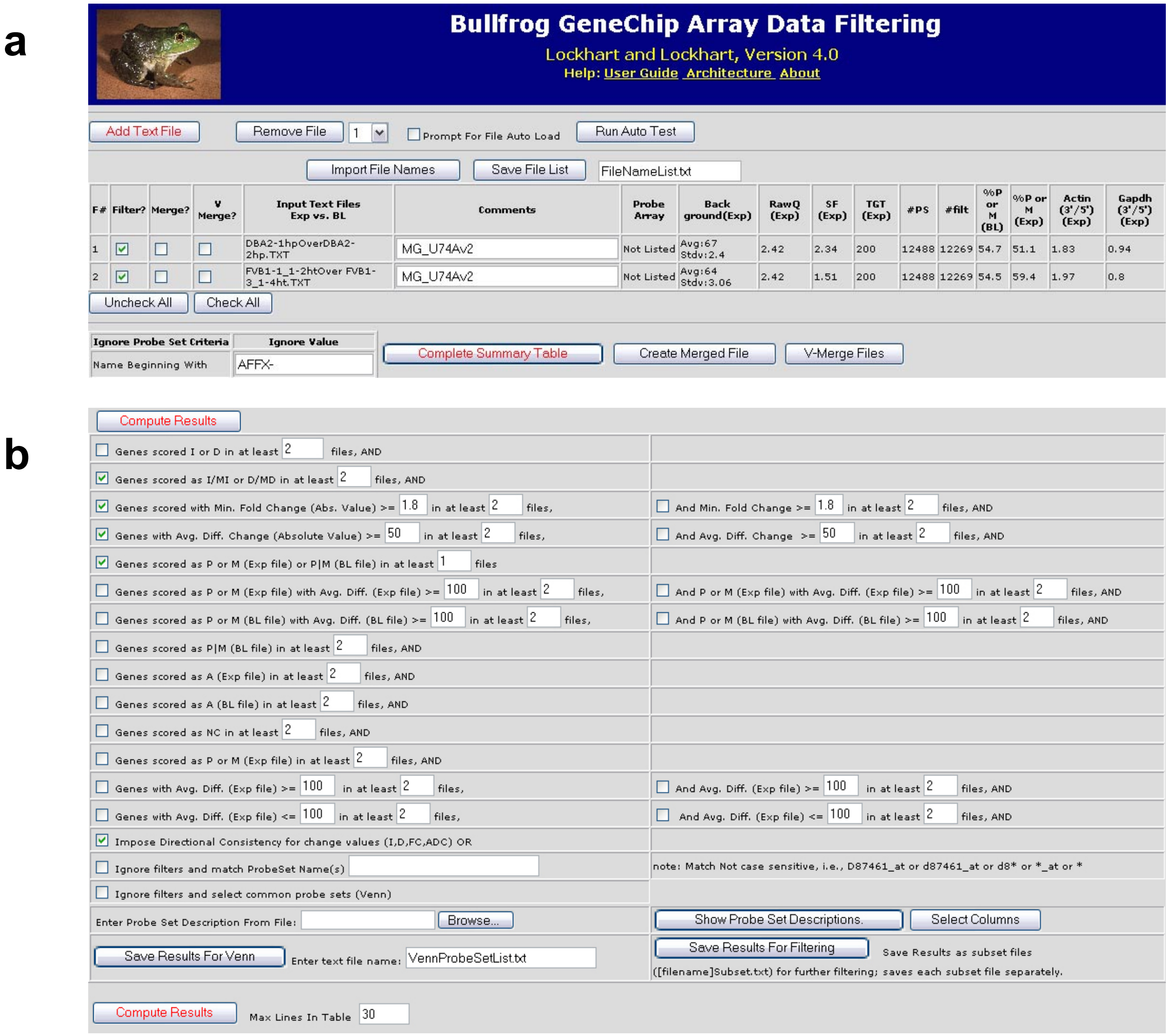
V. Data mining-Bullfrog tool

There are some complex analytical tasks that are impossible without the assistance of computers, but it seems unlikely that computational tools will ever replace the trained human brain when it comes to making biological sense of new results. The greatest progress will come from the ability to bring all the necessary computations, information and relationships to scientists' fingertips so that the most insightful questions can be asked, and the most informed, complete and meaningful interpretations can be made.

Two HTML-based programs were developed to analyze and filter gene-expression data: 'Bullfrog' for Affymetrix oligonucleotide arrays and 'Spot' for custom cDNA arrays. A background subtraction and normalization program for cDNA arrays was also built that provides an informative summary report with data-quality assessments. I will not describe the SPOT program but readers are referred to the paper for more details. Importantly, both programs come complete with user guides. The programs provide intuitive data-filtering tools through an easy-to-use interface. These programs are freeware to aid in the analysis of gene-expression results and facilitate the search for genes responsible for interesting biological processes and phenotypes.

The current Bullfrog version supports files from MAS 4.0 and an updated version is being built that will support files from MAS 5.0 as well as our own analysis algorithms from Teragenomics. In this section I will briefly describe some of the important points that should be considered when defining a gene list and how tools such as Bullfrog can facilitate your work.

Bullfrog was built with an easy to navigate user interface and adjustable analysis criteria. It was written to run quickly, allowing multiple microarray experiments to be filtered in several seconds **(see Figure 5)**.



Bullfrog GeneChip Array Data Filtering

Lockhart and Lockhart, Version 4.0

Help: User Guide Architecture About

□ Prompt For File Auto Load

FileNameList.txt

F#	Filter?	Merge?	V Merge?	Input Text Files Exp vs. BL	Comments	Probe Array	Back ground(Exp)	RawQ (Exp)	SF (Exp)	TGT (Exp)	#PS	#filt	%P or M (BL)	%P or M (Exp)	Actin (3'/5') (Exp)	Gapdh (3'/5') (Exp)
1	☑	□	□	DBA2-1hpOverDBA2-2hp.TXT	MG_U74Av2	Not Listed	Avg:67 Stdv:2.4	2.42	2.34	200	12488	12269	54.7	51.1	1.83	0.94
2	☑	□	□	FVB1-1_1-2htOver FVB1-3_1-4ht.TXT	MG_U74Av2	Not Listed	Avg:64 Stdv:3.06	2.42	1.51	200	12488	12269	54.5	59.4	1.97	0.8

Ignore Probe Set Criteria	Ignore Value
Name Beginning With	AFFX-

□ Genes scored I or D in at least 2 files, AND	
☑ Genes scored as I/MI or D/MD in at least 2 files, AND	
☑ Genes scored with Min. Fold Change (Abs. Value) >= 1.8 in at least 2 files,	□ And Min. Fold Change >= 1.8 in at least 2 files, AND
☑ Genes with Avg. Diff. Change (Absolute Value) >= 50 in at least 2 files,	□ And Avg. Diff. Change >= 50 in at least 2 files, AND
☑ Genes scored as P or M (Exp file) or P|M (BL file) in at least 1 files	
□ Genes scored as P or M (Exp file) with Avg. Diff. (Exp file) >= 100 in at least 2 files,	□ And P or M (Exp file) with Avg. Diff. (Exp file) >= 100 in at least 2 files, AND
□ Genes scored as P or M (BL file) with Avg. Diff. (BL file) >= 100 in at least 2 files,	□ And P or M (BL file) with Avg. Diff. (BL file) >= 100 in at least 2 files, AND
□ Genes scored as P|M (BL file) in at least 2 files, AND	
□ Genes scored as A (Exp file) in at least 2 files, AND	
□ Genes scored as A (BL file) in at least 2 files, AND	
□ Genes scored as NC in at least 2 files, AND	
□ Genes scored as P or M (Exp file) in at least 2 files, AND	
□ Genes with Avg. Diff. (Exp file) >= 100 in at least 2 files,	□ And Avg. Diff. (Exp file) >= 100 in at least 2 files, AND
□ Genes with Avg. Diff. (Exp file) <= 100 in at least 2 files,	□ And Avg. Diff. (Exp file) <= 100 in at least 2 files, AND
☑ Impose Directional Consistency for change values (I,D,FC,ADC) OR	
□ Ignore filters and match ProbeSet Name(s)	note: Match Not case sensitive, i.e., D87461_at or d87461_at or d8* or *_at or *
□ Ignore filters and select common probe sets (Venn)	
Enter Probe Set Description From File:	
Enter text file name: VennProbeSetList.txt	Save Results as subset files ([filename]Subset.txt) for further filtering; saves each subset file separately.

Max Lines In Table 30

Figure 5 Bullfrog Filtering Program

Shown is a screen shot of the Bullfrog program (see http://www.genomebiology.com/2002/3/6/software/0001/ to download the program). **a**| In the example above, two comparison files (hippocampus 1 vs. hippocampus 2 and hypothalamus 1 vs. hypothalamus 2) are loaded into Bullfrog using the "Add Text File" button on the HTML-based graphical user interface. For each comparison file, array hybridization data are displayed in the summary table to the right of the file name. This summary table looks very similar to the data triage sheet shown in Figure 6 and is used to identify experiments that may be of questionable quality. Results such as background, noise (Raw Q), percentage of genes called present (%P or M), scaling factor (SF), Actin and GAPDH 3'/5' ratios are shown. Other information presented in the summary table include the scaling target value (TGT), number of probe sets on the array (#PS), and number of probe sets filtered (#filt) after ignoring control probe sets. **b**| In the lower panel are the default filter criteria applied to two comparison files using both qualitative calls and quantitative data. The default filter criteria include the following: a difference call of "Increase", "Marginal Increase", "Decrease" or "Marginal Decrease", a fold change (ratio) of greater than 1.8, an average difference change (Avg. Diff. Change) of greater than 50 for both files, and an absolute call of "Present" in either the experiment (Exp file) or baseline file (BL file) or both from one or more of the comparisons. Bullfrog is freeware with user guides, open architecture, and open source code.

A. Mining your Affymetrix data

With any microarray experiment, it is first important to determine what analysis criteria should be used to obtain the lowest false positive rate while maintaining sensitivity to subtle gene expression differences. But, how is that reliably determined? In the next section, I will give some detailed examples of how we determine the false positive rate of a data set and identify genes that are differentially expressed. However, it is important to point out that it will always be the case that there will be room for improvement and redesign about the specific implementations of any analysis method (algorithm) for extracting information from an array file. **Regardless of the specific details, an obvious point is that it is very important to determine not just signal strengths, but whether the signals (or signal changes) are due to the gene for which the probes were designed. This type of analysis is made possible by the use of multiple independent probes for each gene (feature by feature detail).** However, this is often very difficult because of the size of the datasets and the richness of the analysis that is performed. And although the use of simple average signals and fold-changes (ratios) will work in some instances, this common approach is often inadequate and highly discouraged because it can result in an increase in the false positive rate, while at the same time sacrificing sensitivity (**as shown in Figure 6 below**). Therefore, this session will focus on how to analyze and mine Affymetrix data without loosing the specificity and sensitivity that the platform offers.

A common mistake when mining Affymetrix oligo-array based gene expression data is to ignore the qualitative calls (absolute and difference calls, feature by feature detail) and focus solely on the quantitative values (e.g., the signal, fold change (ratio) and signal difference). However, the qualitative calls are important because they provide an assessment of the consistency of the behavior across the multiple probes in a probe set. The use of the qualitative calls allows one to determine not only whether there is a signal (or a signal change), but also whether the signal (or the signal change) is due to the gene for which the probe set was designed. Signals or signal changes that are not consistent across a probe set should not be interpreted with confidence. Most of the computational attention for the Affymetrix platform has been directed at the early stages of image analysis or the late stages of high ordered statistical analyses. There has been a lack of specific downstream data, middle management making it difficult to ask crucial research questions, such as **what parameters should be used to set the false positive rate? What is the false positive rate? What genes are in common or different between multiple experiments at various ratios (fold changes) and signal difference levels?** This lack of downstream data management has forced the user to manually manipulate large unwieldy data sets using Microsoft Excel or Access or to merge data sets to more manageable sizes, which results in a loss in data sensitivity. **In order to assist the researcher in asking these types of questions, we use Bullfrog to address common data analysis needs that were currently unmet in the academic and commercial sectors.** Our goals in creating this program were **to provide simple tools that allow researchers at all levels to analyze their data in multiple ways without having to use more complex software, without the help of bioinformatics experts, and without having to learn to program in scripting or database languages**. Bullfrog was built with an easy to navigate user interface and adjustable analysis criteria. It was written to run quickly, allowing multiple microarray experiments to be filtered in several seconds. Lastly, it was created to **provide the bench researcher with uncomplicated tools that help focus microarray data from thousands of genes to a relatively small number of high-confidence, differentially expressed candidates**. The programs were also designed to easily export analyzed and filtered data to other visualization and clustering programs such as GENESPRING (see http://www.genomebiology.com/2002/3/6/software/0001/ to download the program and user's guide).

To illustrate a few of the capabilities of Bullfrog, we use data obtained in gene expression studies of the adult mouse brain (Sandberg et al., 2000). A simple question to ask is, "what genes are differentially expressed between two different regions of the brain (e.g., the cerebellum and the amygdala) in a 129S6/SvEvTac (129SvEv) inbred mouse strain?" To estimate the false positive rate, a comparison file between data from independent replicate 129SvEv cerebellums is made (i.e., expression data from mouse 1 cerebellum versus expression data from mouse 2 cerebellum). Using Bullfrog, the user can test a variety of criteria and check how many genes pass the filter. The filter criteria the user selects are the minimum criteria used to assign a gene as differentially expressed as seen in **Figure 6**.

Comparison	# of Genes Different with Standard Filters and Qualitative Calls	# of Genes Different with Standard Filters and without Qualitative Calls	# of Genes Different with FC >= 10 and Difference >=175 ONLY	# of Genes Different with FC >= 10, Difference >=175 and Calls
Cb1 vs. Cb2 (replicate 1)	36	715	34	0
Cb1 vs. Ag1 (experiment 1)	348	1128	50	31

Figure 6 Increasing Specificity and Sensitivity using Multiple Filtering Criteria

Shown is the number of differentially expressed genes after filtering criteria were applied to two different comparison files using the Bullfrog software tool. Bullfrog allows users to analyze and filter their data in multiple ways. Shown here are 4 different types of filtering criteria applied to two different comparisons, a replicate comparison and an experimental comparison. The top comparison is a replicate comparison of two cerebellums from two independent mice. When the Bullfrog tool is used to analyze the replicate comparisons using the default criteria (difference call of "I", "MI", "D" or "MD", fold ≥ 1.8, signal difference ≥ 50 and an absolute call of Present) the false positive rate is just 36 out of 6,584 (0.6%). However, when the qualitative calls are ignored, and just the fold change and signal difference are used (fold change ≥ 1.8, signal difference ≥ 50) the false positive rate increases to 715 out of 6,584 (10.8%). This example displays the drastic increase (20 fold) in the number of false positives between replicates by ignoring the qualitative calls (feature by feature detail) and just focusing on the averaged quantitative data (i.e. fold change and signal difference). Just using the qualitative calls dramatically reduces **specificity**. To maintain the low false positive rate obtained with the combination of qualitative and quantitative criteria (~0.6%) using only the quantitative fold change and signal difference criteria, the thresholds would have to be set very high as shown in the third column, a fold change of greater than 10 and a signal difference of greater than 175. Using just these high quantitative thresholds produces 34 out of 6,584 (0.6%). However, fold change and signal difference thresholds this high result in a tremendous loss in **sensitivity**. Shown below the replicate comparison is the experimental comparison that looks for differential gene expression between the cerebellum and the amygdala. It can be assumed from the replicate example given above that the false positive rate for the experimental comparisons between cerebellum and amygdala significantly increases when the qualitative calls are ignored. The number of genes that are classified as consistently differentially expressed in the experimental comparisons increase from 348 genes to 1128 genes when qualitative calls are ignored. If the experimental comparison is filtered with only the high quantitative thresholds (fold change greater than 10 and signal difference greater than 175) only 50 genes pass. Of the 50 genes that pass only 31 are present in the 348 genes that pass using both qualitative calls and low quantitative criteria. This example demonstrates that an effective way to preserve **specificity** while maintaining high **sensitivity** is to use a combination of both qualitative (feature by feature detail) and quantitative filters.

Based on our own experiments, we set the default criteria for calling a gene "differentially expressed" as follows: difference call of I, MI, D or MD, a fold change (expression ratio) of greater than 1.8, an average difference change of greater than 50 and an absolute call of P for the probe set in either or both replicate cerebellums. These criteria are then applied to the independent replicates and then to the comparison of cerebellum to amygdala. The use of multiple filter criteria reduces the risk of erroneously assigning a gene as differentially expressed while maintaining sensitivity to rare mRNAs and small expression differences.

During the session I will describe some of our results using the Bullfrog tool and Genespring (for clustering) that allows the basic biologist to easily manage and mine large numbers of Affymetrix arrays with confidence and ease.

Selected readings

Barlow, C. and Lockhart, D. DNA Arrays and Neurobiology - What's New and What's Next?, *Current Opinions in Neurobiology.* September 2002 online, in press October 2002.

Lockhart, D. J., and Barlow, C. (2001). DNA Arrays and Gene Expression Analysis in the Brain. In Methods in Genomic Neuroscience, S. Moldin and H. Chin, eds.: CRC press), pp. pp 143-170.

Lockhart, D. J., and Barlow, C. (2001). Expressing what's on your mind: DNA arrays and the brain. Nature Reviews, Neuroscience *2*, 63-68.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, K. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology *14*, 1675-1680.

Sandberg, R., Yasuda, R., Pankratz, D. G., Carter, T. A., Del Rio, J. A., Wodicka, L., Mayford, M., Lockhart, D. J., and Barlow, C. (2000). From the cover: regional and strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A *97*, 11038-43. *Data from this paper are used in multiple examples in this syllabus and will also be used for the session to demonstrate functionality of the tools.*

Zapala, M.A., Lockhart, D.J., Pankratz, D.G., Garcia, A.J., **Barlow, C.** and Lockhart, D.J. Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biology*, 3, 1-9.