

# Genome-based analysis of expression microarray data

Ann Loraine (1), Xiangqin Cui (1) and Gary Churchill (2)

(1) Section on Statistical Genetics, Dept. of Biostatistics, University of Alabama at Birmingham; (2) The Jackson Laboratory, Bar Harbor, ME

## Abstract

We developed a genome-based analysis method that uses data from expression microarray experiments (Affymetrix platform) to investigate alternative mRNA processing. This approach revealed numerous examples of genes whose expression patterns suggest strain-dependent differential mRNA processing in mouse.

## Introduction

Current Affymetrix expression microarray designs typically include numerous redundant probe sets that interrogate different regions of the same gene. This is a side effect of a design process which creates probe sets matching alternative splicing or other types of mRNA variants. Thus, when redundant probe sets produce discordant results, such as differential expression in opposite directions, the most logical explanation is that the condition under investigation affects mRNA processing pathways, such as alternative splicing and alternative polyadenylation site choice.

Both types of alternative mRNA processing can have profound impacts on gene function when important functional motifs, such as conserved protein motifs or mRNA stability determinants, are affected. Discordant behavior of redundant probe sets therefore may represent an opportunity to study how alternative mRNA processing pathways operate. Here we investigate this idea using expression data collected from two mouse strains.

## Methods

### Data collection and pre-processing.

Labeled mRNA samples were prepared from the livers of six mice, three of strain AJ and three of strain B6. Labeled samples were hybridized to two Affymetrix 430\_2 microarrays per mouse. The array data were processed using RMA and normalized using quantile-quantile normalization.

### Redundant probe sets.

Affymetrix provides annotation data files that map probe set ids to Entrez Gene ids as well as 'psl' files that map probe set design sequences onto the May 2004 mouse genome. We used these files to create and then screen a provisional list of redundant probe sets for use in our study.

## Methods

Probe sets per Entrez Gene ID	# Entrez Gene IDS
1	11,511
2	4,703
3	2,309
4	1,062
5	454
6	257
7	109
>7	87
TOTAL	20,492

The 430\_2 mouse array contains 37,449 probe set ids that map onto 20,492 Entrez Gene ids. Thus, according to the Affymetrix annotations, over 9,000 Entrez Gene ids are associated with two or more probe sets.

These annotation files contain a number of inconsistencies, such as probe sets mapping to an unexpected chromosome or strand. To remove these, we compared genomic alignments of the probe set design sequences with alignments for Refseq mRNAs associated with the annotated Entrez Gene ids. Probe sets whose design sequence alignments were inconsistent with a reliable RefSeq alignment were eliminated.

Screening step	# Probe Sets
One genomic location, one Entrez Gene id	34,196
Inconsistent with Refseq genomic alignment	-2036
Reliable Refseq alignment not available	-7,895
TOTAL remaining	24,265

### Intensity Screening.

We eliminated all probe sets from the analysis that produced low-intensity signals. For this study, we consider *active* probe sets only - probe sets that detected a target mRNA in at least one of the experimental conditions.

### Statistical Methodology

We used mixed effect ANOVA model to identify genes in which the redundant probe sets produced discordant differential expression readings across strains. For each gene, we fitted a mixed effects linear model:

$$y_{ijk} = \mu + S_i + P_k + S_i * P_k + M_{j(i)} + \epsilon_{jk}$$

where  $y_{ijk}$  is a measure of gene expression produced by an individual probe set;  $\mu$  is the grand mean of all probe set readings for the gene being considered;

## Methods

$S_i$ ,  $M_{j(i)}$  and  $\epsilon_{jk}$  are the strain, mouse and residual effects;  $P_k$  is the probe set effect; and  $S_i * P_k$  is the interaction between strain and probe set.

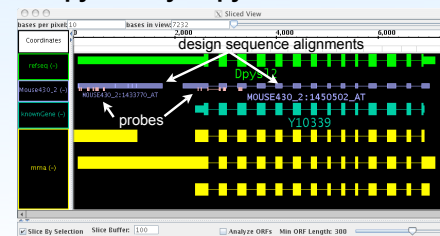
In this setting, different probe sets can provide expression measurements for a single gene, and these measurements may vary across different mouse strains (fixed effects) as well as across individual mice (random effects). However, some probe sets - because they interrogate different regions of the same gene - may respond differently in the different strains.

The interaction term  $S_i * P_k$  captures any differential expression arising from the combined effect of probe set and strain. Our null hypothesis was that this strain by probe set interaction term is zero. We tested this using an F test for mixed effect ANOVA models and used an adaptive FDR method to adjust p-values for a desired FDR of 0.05. Note that this test does not identify the individual probe sets which are discordant but instead identifies genes in which not all probe sets respond to the strain difference in the same way.

## Results

The analysis identified 1,400 genes with adjusted p-values less than or equal to 0.05. We examined a number of these using the Integrated Genome Browser - see below [1]. In some cases, alternative mRNA processing alters the coding region and modifies conserved functional motifs, a common occurrence among alternatively spliced genes [2,3].

### Dpys12 dihydropyrimidinase-like 2

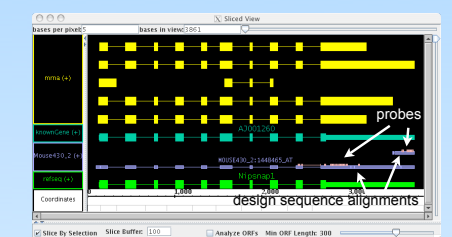


#### probe set average intensity readings

probe set	AJ	B6	orientation
1450502_at	8.05	7.77	-0.28 proximal
1433770_at	9.32	9.58	+0.26 distal

## Results

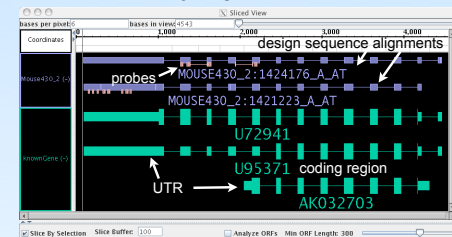
### Nipsnap1 4-nitrophenylphosphatase domain and non-neuronal SNAP25-like



#### probe set average intensity readings

probe set	AJ	B6	orientation
1448465_at	10.4	10.6	+0.2 proximal
1450184_s_at	11.8	11.4	-0.4 distal

### annexin A4



#### probe set average intensity readings

probe set	AJ	B6	orientation
1424176_a_at	8.76	8.43	-0.33 proximal
1421223_a_at	9.01	9.62	+0.61 distal

## Conclusion

Based on these preliminary results, we conclude that analyzing the relative expression of redundant probe sets from Affymetrix expression microarray designs has the potential to reveal how diverse experimental conditions affect differential mRNA processing and degradation pathways.

## References

- [1] The Integrated Genome Browser is open source software available from [www.genoviz.org](http://www.genoviz.org).
- [2] Loraine, et al. (2002) Protein-based analysis of alternative splicing in the human genome. IEEE Comput So Bioinform Conf. pp. 321-326
- [3] Loraine, et al. (2003) Exploring alternative transcript structure in the human genome using BLOCKS and InterPro. J Bioinform Comput Biol. pp. 289-306.