Introduction

From the development of alcoholism to the transformation of a healthy cell into tumor cell, on a molecular basis, every state of the human body arises from changes in proteins produced, which, in turn, is affected by the genes that are being expressed in a given cell under certain conditions. By understanding the genes that are affected by a given event, researchers can find target proteins for inhibition or activation by a drug to change the way in which the body reacts to the event.

Summer Summary

An increasingly common way to find answers to gene expression is by microarray. Oligonucleotide microarrays allow researchers to examine the relative expression level of thousands of genes simultaneously. If this technology is reliable, it can provide a huge insight to the regulation of genes for a given situation.

Specifically, the lab in which I was used microarrays to look at differences in gene expression between two inbred mouse strains (namely, C57 Black 6 or C57/BL6 and DBA) that have different behavioral reactions when treated with ethanol. Microarrays measure the amount of gene expressed by the amount of complimentary DNA (cDNA) that anneals to a complimentary sequence on the chip. But the different inbred strains are not genetically identical. In addition to other possible genetic differences, they may have one nucleotide mismatches. These are called single nucleotide polymorphisms or SNPs. If a SNP were to occur in the target sequence (the part of the gene that anneals to the microarray) would that affect the annealing of the cDNA to the array enough to alter the perceived relative expression of the gene?

In order to answer this question I found genes that were differentially expressed in a manner that suggested that the difference could be a false positive caused by SNPs: genes for which a difference in expression was reported regardless of cellular environment. I selected twenty-one candidate genes from genes that were differentially expressed in a consistent manner in saline-treated animals of the two strains across all brain regions examined (the prefrontal cortex, the nucleus accumbens and the ventral tegmental area.)

For each of the candidate genes I took the sequence that Affymetrix was using for probes, and the sequence of each of the two mice strains. I then aligned the three sequences. Of the eight genes for which I could find sequence data for the entire target sequence, five had at least one SNP that affected at least one probe. In addition, of the seven genes for which at least partial sequence data was available for the target sequence in each strain, three had at least one probe affected by a SNP. I selected three of the five genes for which there was complete sequence information in both strains and that were affected by SNPs. The genes I selected were Aldehyde Dehydrogenase family 9, subfamily a1 (ALDH9a1), Kinesin Associated Protein 3 (KAP3) and Valosin Containing Protein (VCP).

For each selected gene, I designed polymerase chain reaction (PCR) primers that bound outside of the microarray target sequence for the gene to an area that had no SNPs. I then ran real-time reverse transcriptase PCR on complementary DNA (cDNA) for samples of each gene from three saline-treated animals of each strain. For Real time PCR a chemical that fluoresces when bound to DNA is added to the reaction (I used SYBR-green, which binds non-specifically to the minor groove of double stranded DNA). As

the DNA is amplified, more fluorescence is emitted. Pictures of the samples taken every cycle by a digital camera inside of the thermocycler are analyzed and used to calculate the quantity of DNA in each sample[2]. I compared the fold-change in the starting quantity (SQ) between the two strains to the fold change between the strains in average difference from Affymetrix oligionucleotide arrays. I used cDNA made from the same animals and the same brain region as those used for the microarrays.

For all three genes, my data suggests that the arrays were affected by the SNPs. For ALDH9a1, for which all but 4 of the Affymetrix probes had polymorphisms with respect to DBA mice, microarray data suggests that C57BL/6 mice express the gene 9.9 times more than DBA mice, whereas qPCR showed a fold change of only 1.31 (figure 1). For VCP, which had SNPs in 4 of its Affymetrix probes with respect to the C57BL/6 sequence, microarrays suggested that DBA was expressed at a level 1.18 times higher than C57BL/6 was, whereas PCR suggests that the difference is closer to 73 times (figure 2). Although that data seems strange in that it suggests that the SNPs actually increased annealing to the probes, it is predicted that in some cases a mismatch actually does make binding more favorable by increasing the stacking energy[3]. But perhaps the most striking result is that for KAP3, which has two probes affected by SNPs for DBA, microarrays reported that the gene was expressed in C57BL/6 1.34 times more highly than in DBA, whereas PCR data suggests that the gene was actually expressed 1.30 times more highly in DBA (figure 3).

To confirm the wet lab experiment, I compared the measured intensity of each probe for the three candidate genes from experiments run on saline treated C57 and DBA animals. For ALDH9a1 (figure 4) and KCP (figure 6), the probes for which there were

SNPs in DBA showed a higher C57 to DBA ratio of expression than those without SNPs, which appears to confirm that SNPs hinder the annealing to probes for those genes. For VCP (figure 5), the probes that had SNPs with regard to C57 also reported a higher C57 to DBA expression ratio than those without SNPs, which appears to confirm that the SNPs are increasing the annealing of cDNA to the probes for that gene.

My data suggest that microarray results are not trustworthy when being used to compare relative levels of gene expression between genetically different subjects. In one case, my data even suggest that the relative expression was opposite that reported by the microarray.

Academic Year Project

For the academic year, my project will be taking a very different approach to the question of the affects of gene expression. Knowing what genes are expressed in a given situation is very important, but it also leads to more questions: what do those genes do? How do they affect other genes? How are they affected by other genes? Molecular changes happen not just by the production of a single protein, but also often by a pathway of effects.

One of the ways in which pathways can begin to be predicted is by creating a graphical network that connects genes that are closely related in function and in regulation. This is partially already possible; gene expression can be measured, data on predicted functionality and correlations in expression are available on the World Wide Web[5,6], and programs such as Cytoscape[1,4] will take predicted network relations and graph them and allow the user to work with them. But collecting all of the necessary data

and manually predicting networks is still a time consuming, manual task. My academic year project seeks to attempt to automate that procedure.

I will be writing a program that accesses numerous databases and links genes on the basis of the data from all of the databases. It would then display the created network. This program also has the advantage that it can keep predicted networks as up-to-date as the databases that it accesses with no additional manual work.

There are three main types of data that I plan on integrating to complete this task. The first is microarray data. If the up-regulation of a gene strongly correlates with the up- or down-regulation of a second gene on microarray chips, that would be considered a linkage between them. For this data, I plan on using microarray data collected by the Miles lab. The second type of data is the predicted function of the protein produced by the gene. If two genes are thought to be involved in the same broad category of functionality that would be considered a link. This data would be accessed from EASE[5]. The third type of data is Quantitative Trait Loci (QTLs). If one gene is found to have its expression highly correlated with another gene the genes are said to have a QTL. This data will be accessed from WebQTL[6].

Algorithmically, the program would be very simply. Based on graph theory, it would consider each gene to be a vertex or a "node." Two nodes would be connected by an edge if they were linked by microarray data, predicted functionality or a QTL. The edge would be "weighted" – given a number that, functionally, relates to the probability and strength of a relationship between the genes. The weight would be directly related to the number of different data sources that indicate a connection between the genes.

Once the data structure of nodes and edges is created the program will be able to create a graphic using a similar system to that of the open source program Cytoscape. That is, the program will display a graph in which the vertices are genes (nodes), which are connected by lines whose width is proportional to the weight of the corresponding edges. Code from Cytoscape, which is licensed under the Lesser Gnu Public License (which makes it legal to modify and redistribute the code) may be used in helping to build the graphical interface.

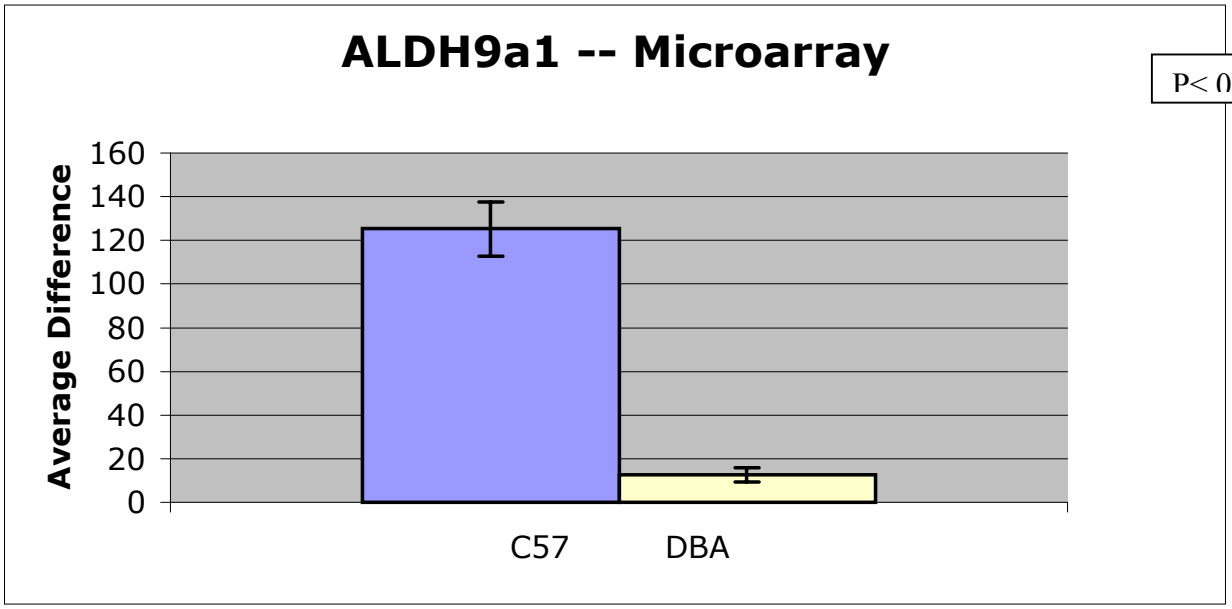
Required Resources

No formal coursework is necessary for this project, although independent study of Perl and consultation with faculty of Oberlin College, specifically on databases and graphics are expected. Additionally, no funding is necessary for this project.

Figures

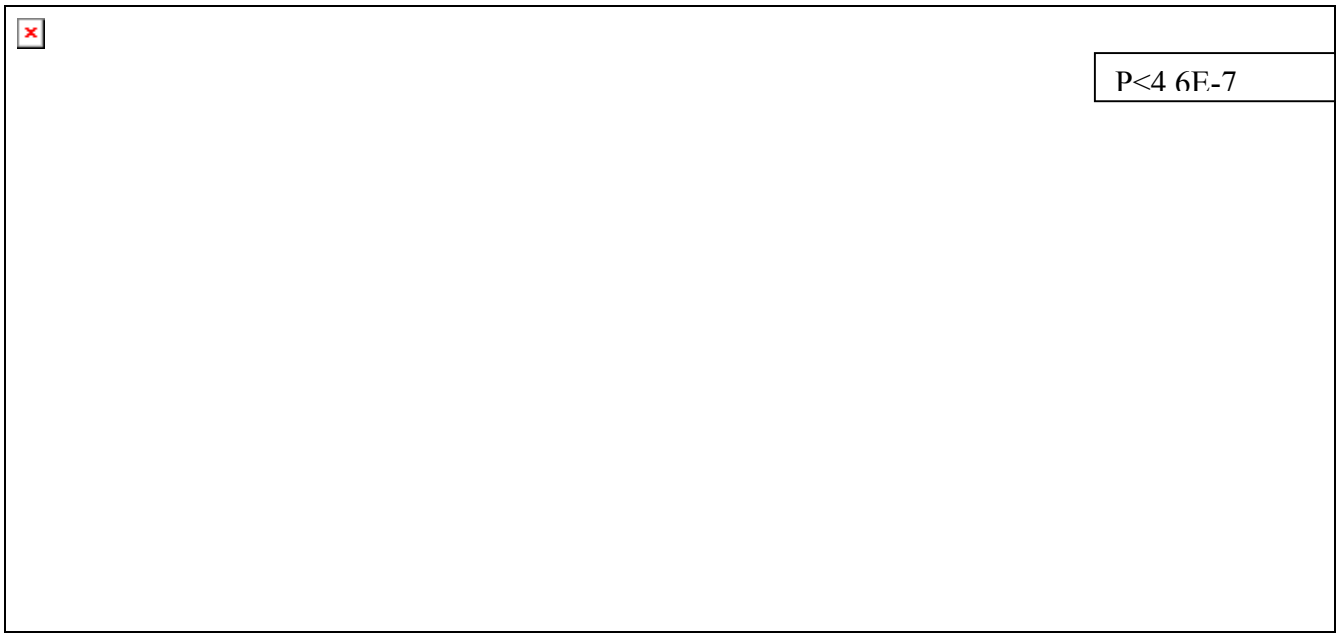
Fig 1.

P< 0013


ALDH9a1 -- Microarray
Average Difference
160
140
120
100
80
60
40
20
0
C57
DBA

P< 002

Figure 2



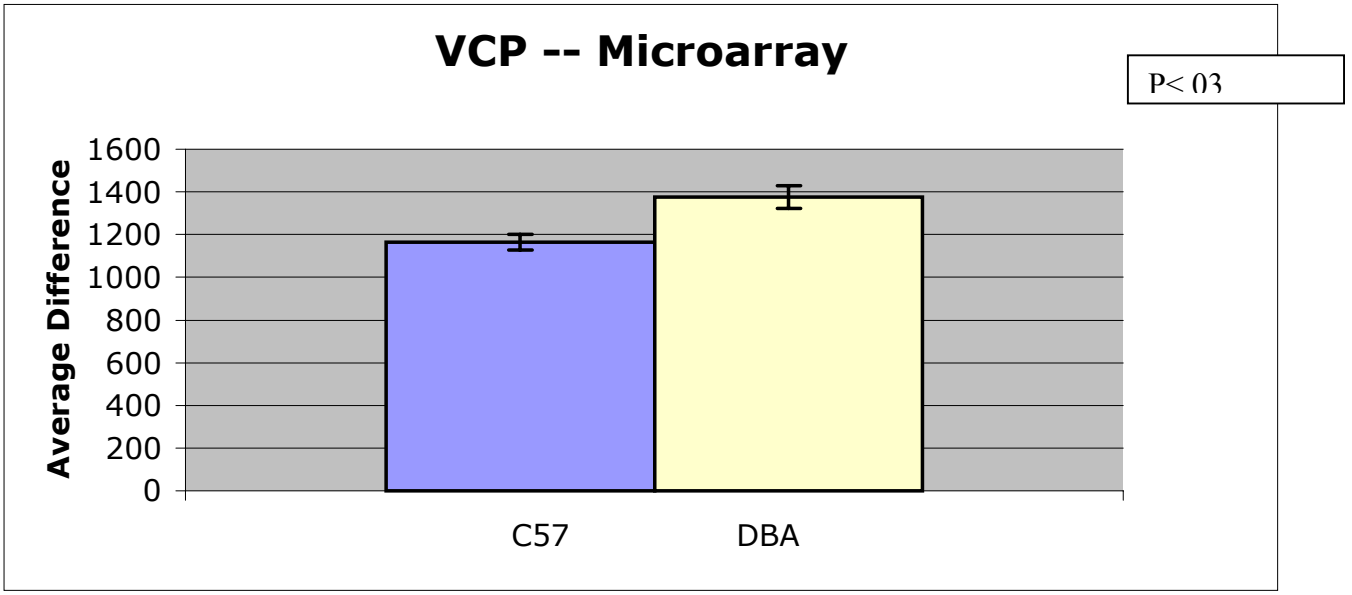
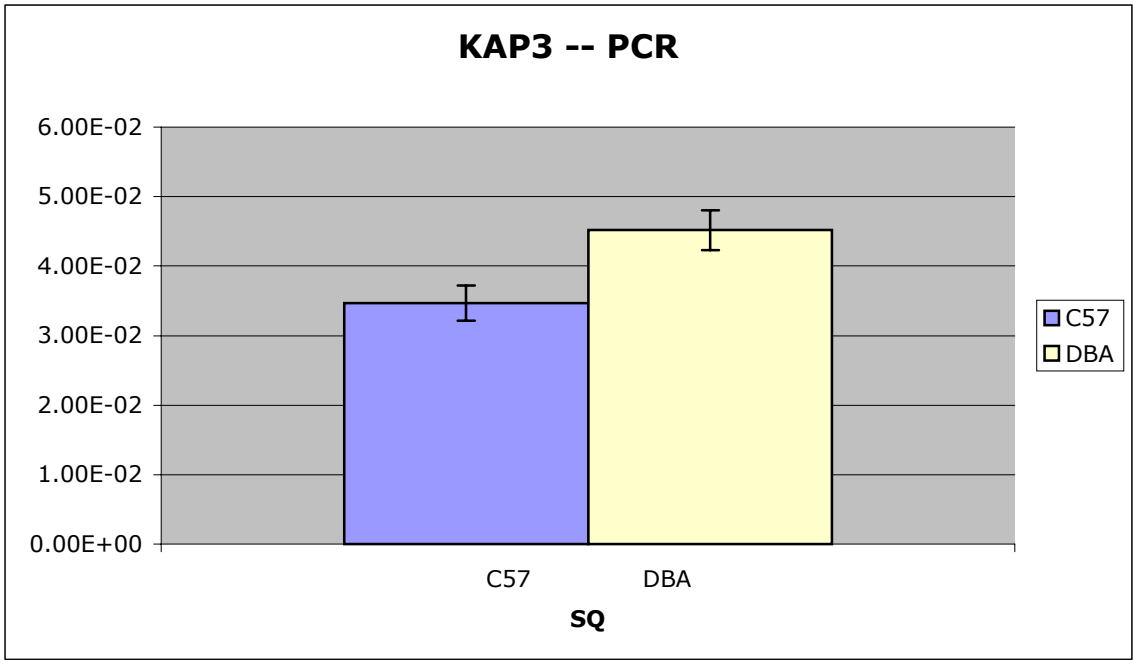


Figure 3

P< 015



P< 002

Figure 4

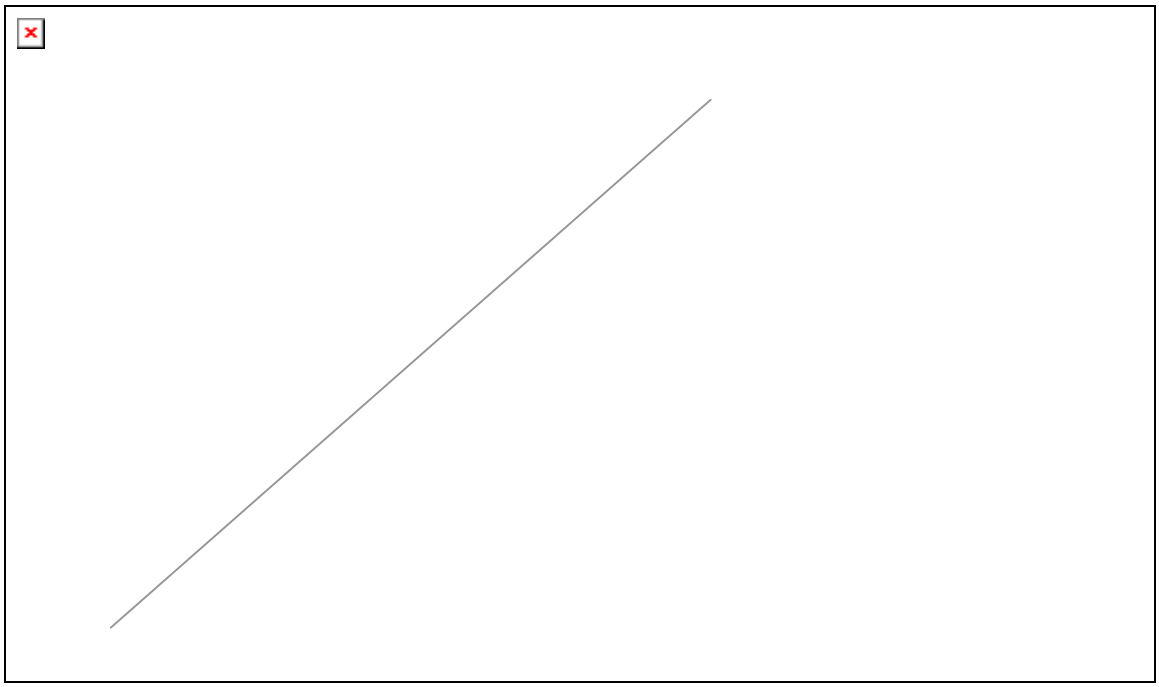


Figure 5

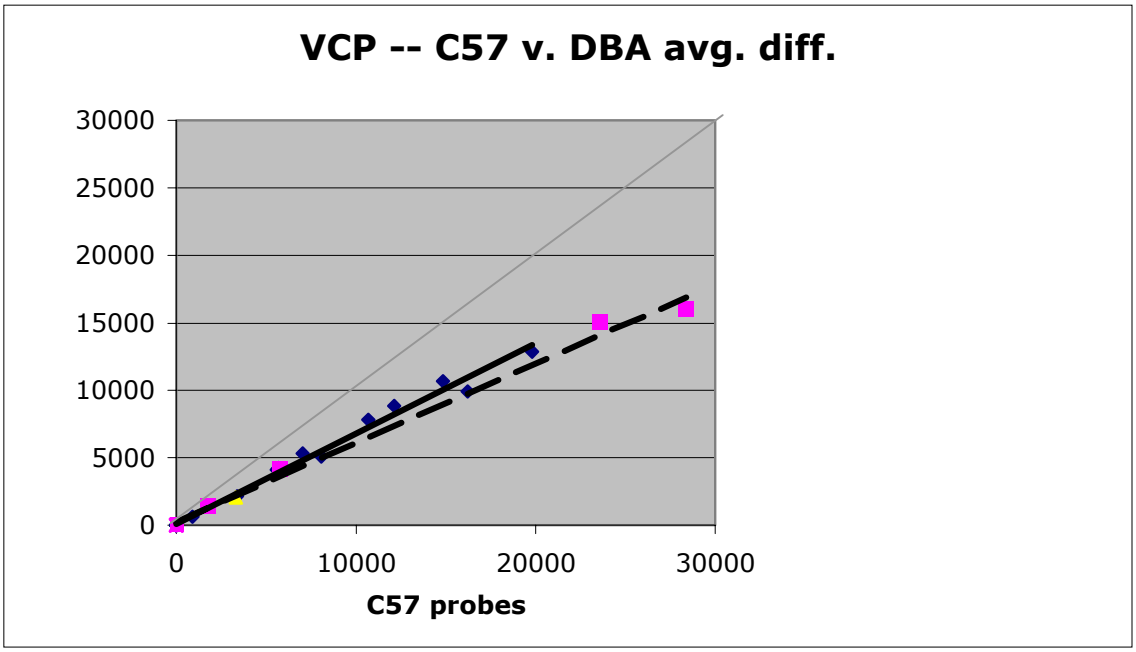
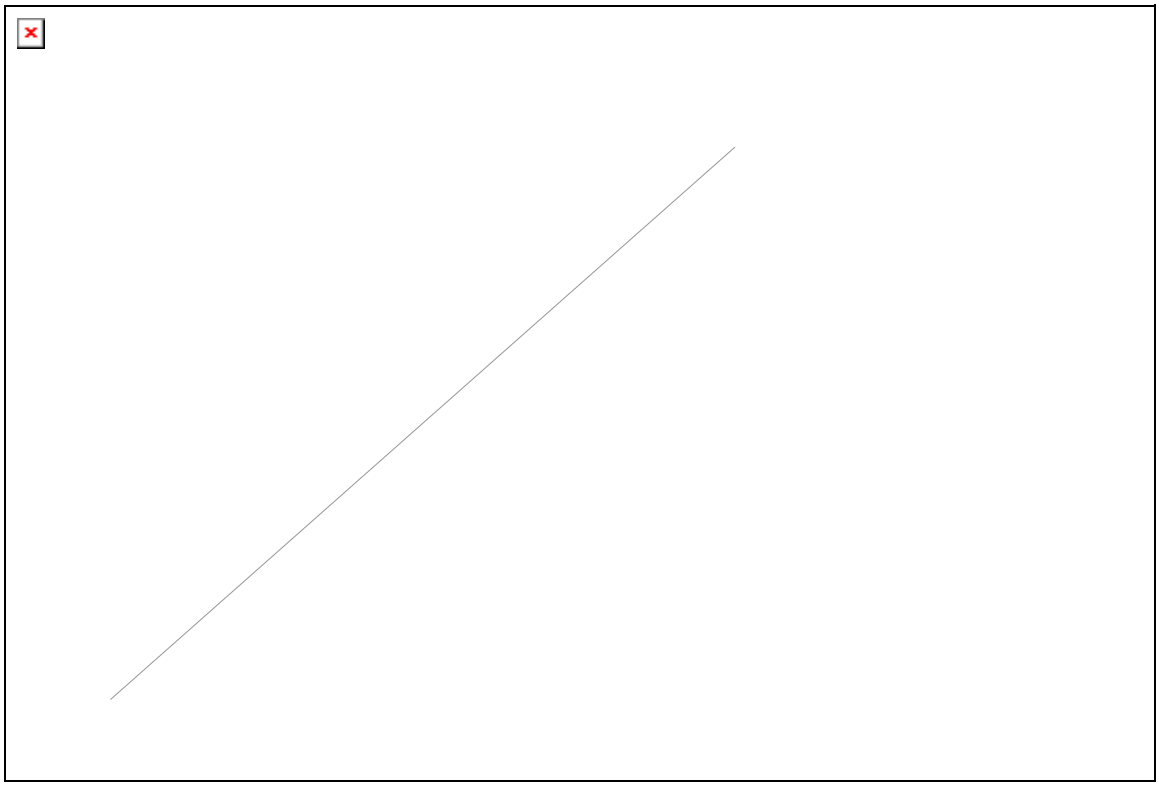

VCP -- C57 v. DBA avg. diff.
30000
25000
20000
15000
10000
5000
0
0
10000
20000
30000
C57 probes

Figure 6



References

1. Ideker T, Ozier O, Schwikowski B, Siegel A. Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks. *Bioinformatics* **18** S233-S240 (2002).

2. Robertson J, Ziegle J, Kronick M, Madden D, Budowle B. Genetic Typing Using Automated Electrophoresis and Fluorescence Detection. *EXS.* **58** 391-398 (1991).

3. Zhang L, Miles M, AldapeK. A Model of Molecular Interactions on Short Oligonucleotide Microarrays. *Nature Biotechnology* **21** 818-821 (2003).

4. Cytoscape: Analyzing and Visualizing Biological Network Data. http://www.cytoscape.org

5. NIH – DAVID (Database for Annotation, Visualization and Integrated Discovery). http://david.niaid.nih.gov/david/upload.asp/

6. The WebQTL Project. http://webqtl.roswellpark.org/