The Title

Allen Day Jun Dong Stanley F. Nelson

April 23, 2008

1 Abstract

...

2 Introduction

3 Methods

3.1 Data Processing

We retrieved RMA-processed gene expression data for the HG-U133_Plus_2 arraydesign from the Celsius microarray data warehouse (`http://genome.ucla.edu/projects/celsius`) [5, 1]. In total, there were 12826 arrays S, each of which reports measurements for 54675 probesets P. We denote this initial $S \times P$ matrix as M.

It was clear from a cursory examination of summary statistics prepared from M that there were abberant arrays present, and that these arrays would have a negative impact on any downstream analyses. At a coarse level, there appeared to be XXX types of aberrant arrays:

A. arrays with extremely high gene expression values across many probesets.

B. arrays with extremely low gene expression values across many probesets.

C. arrays with dissimilar expression values for two probesets reputedly measuring the same gene.

We sought to remove these abberant arrays from the dataset.

3.1.1 Removal of Dim and Bright Arrays

Class A & B arrays were easiest to identify. We calculated the mean expression value of all probesets for each array, then calculated the mean and standard deviation of a 10 % trimmed distribution of those means.

The trimmed means had a mean of 231.1081 and a standard deviation of 21.01693 . A histogram of the distribution is given in Figure 5.

There were 726 arrays with mean a expression value more than 3 standard deviations away from the mean of trimmed means. These were primarily dim arrays (n=711) but there were also bright arrays (n=15). These arrays were removed from further consideration, leaving a matrix $M\prime$ with 12100 arrays and 54675 probesets.

3.1.2 Removal of Inconsistent Arrays

Class C arrays were slightly more difficult to find. To identify them, we exploited the fact that, via NetAffx [8], Affymetrix publishes a probeset $\rightarrow$ gene symbol mapping for their array designs. We assumed that pairs of probesets designed to target the same gene were more likely to be linearly related than randomly selected pairs because they were targetting the same gene, and that these relationships could be used as a starting point to identify inconsistent arrays.

We took all 19632 unique gene symbols from the NetAffx HG-U133_Plus_2 gene annotation, and identified the subset G (n=10433) for which there were two or more probesets. We constructed G groups, each corresponding to a single gene symbol, i.e. $g_1 = p_{g_1}1, \ldots, p_{g_1}n, \ldots, g_G = p_{g_G}1, \ldots, p_{g_G}n$. Then, for each $g \in G$, we performed a linear regression of $log_10(signal)$ for all possible probeset pairs p_gA, p_gB (n=38682). A distribution of r^2 values from those linear regressions is given in Figure 5.

Examination of the probeset pairs with the largest value of r^2 revealed that the majority were control probesets that targeted spike-in sequences that are added as part of the microarray hybridization for quality control. We performed 62 multiple regressions, allowing each control probeset to be the response variable once. In the context of a single regression if an array's residual was, relative to all other arrays' residuals, more than 3 standard deviations away from the line, we incremented a counter for that array. Outlier frequencies per per array is shown in Figure 5.

After performing all 62 regressions, all arrays that were observed more than 3 standard deviations more than 5% of the time (n=464) were removed from further consideration, leaving a matrix $M\prime\prime$ with 11636 arrays and 54675 probesets.

3.1.3 Gene-Gene Correlation Matrix

After filtering out abberant arrays from out dataset (Section 3.1), we used the $M\prime\prime$ matrix to calculate $C\prime\prime$, a 54675 $\times$ 54675 matrix of Pearson correlation coefficients for every pair of probesets. $C\prime\prime$ was used in all results presented in Section 4.

3.2 Gene Annotation

For each probeset $p \in P$ on the HG-U133_Plus_2 arraydesign, we retrieved the correlation coefficient vector r to all other probesets $P \ni p$ from $C\prime\prime$ (Section 3.1.3). We then selected the nearest neighbors. This was done by placing r into descending order, and calculating the derivative $r\prime$ as $r_i\prime = r_i - r_{i+1}$. We used the Bayesian Change Point *bcp* package XXX to produce $\hat{p}$, an estimate the probability of a changepoint for each position in $r\prime$. A single best change point δ was selected as the index of the largest

value of r that preceded 10 consecutive values $\hat{p} < 0.5$. Probesets Q where $r > \delta$ were used as input to the ***hyperGTest*** function of the ***GOstats*** package of Bioconductor [3] to test for enrichment of Gene Ontology (GO) Biological Process (BP) annotations in a gene set. ***hyperGTest*** produced a set of predicted gene annotations N_p for each $p \in P$ based on the annotation of neighbors Q. We used the p-values from predicted annotations N_p that were known to be non-computationally assigned from the ***hgu133plus*** package of Bioconductor [3] to establish a conservative cutoff, below which predicted annotations should all be high-quality.

3.3 Linkage Region Candidate Selection

For a given phenotype, a group of known genes G known to be associated with that phenotype were retrieved from previous publications and online databases. The list of genes was transformed to a list of probesets P present on the HG-U133_Plus_2 arraydesign using the gene symbol $\rightarrow$ probeset mapping available from NetAffx [8]. In the event that a gene in G mapped to multiple probesets, only the first of the probesets in alphanumerical order was selected to prevent multiple counting. Probesets P were then mapped to 6-megabase genomic regions A by finding the center point of each probeset's alignment to UCSC's March 2006 (hg18) version of the human genome and expanding by 3 megabases in each direction. Each region in A was then mapped to a list of all HG-U133_Plus_2 probesets Q aligned to that region. Then, for each $p \in P$, a $Q \times P - 1$ slab was retrieved from C'' (Section 3.1.3), and row-summarized to produce a Q-length vector $\vec{r}$ of mean correlation coefficients to $P \ni p$.

4 Results

Our aim was to mine the matrix of correlation coefficients for all probesets on the Affymetrix HG-U133_Plus_2 arraydesign for new information.

We wanted to let the data speak for themselves, and so included only a minimum of metadata. Metadata for samples hybridized to the arrays were excluded entirely from analyses. For probesets, we only included gene-symbol [8], genomic alignment [7], and human-reviewed Gene Ontology (GO) Biological Process (BP) [4, 3] metadata.

4.1 Data Processing

All HG-U133_Plus_2 arrays (n=12826) were retrieved from Celsius [1]. We assessed the arrays using some simple quality control (QC) metrics, and excluded several hundred based on unusual array intensities and unusual behavior of control probesets (Section 3.1), yielding a 11636 array $\times$ 54675 column matrix, denoted M''. We calculated the Pearson correlation coefficient for every pair of probesets in M'', yielding a 54675 $\times$ 54675 correlation matrix, denoted C''.

4.2 Disease Gene Recovery

Commonly, the first published evidence of association between a hereditary disease and one or more genes does not explicitly refer to the associated genes but rather describe the association to multiple associated genetic loci that should be examined more closely [9, 6]. These so-called linkage regions are commonly up to 10 megabases in size, and thus typically contain 60-100 genes, assuming an average gene size of 50 kilobases.

When the associated genes are eventually identified, it is frequently the case that they are all involved in the same biological process, and that this process is disrupted when one of its components is dysfunctional. Given that the genes are involved in the same biological process, it is reasonable to assume that they will be coexpressed in cells where the process occurs and thus be positively correlated.

Extending the idea that genes involved in the same biological process will generally be positively correlated, we sought to use C'' (Section 4.1) to simulate the identification of a disease gene.

Our method was to assemble a list of genes G known to be associated with a disease. Each gene identifier $g \in G$ was mapped to the corresponding list of probesets on the HG-U133_Plus_2 arraydesign. The list is denoted P_g, and is derived from the mapping function denoted $J(g)$. For each probeset in $p_g \in P_g$, the genomic position was retrieved using the UCSC Genome Browser [7]. We then retrieved a list of probesets which aligned to a 6 megabase genomic region surrounding the initial probeset. The list is denoted Q_{p_g}, and is derived form the mapping function denoted as $K(p)$. Next, the vector of mean correlaton coefficient $\bar{r}_{p_g}$ of probeset $q_{p_g} \in Q_{p_g}$ to $P \ni J(g)$ was calculated using function $L()$ from C''. Finally, the best gene in the region was identified as the one matching $J^{-1}(K^{-1}(L^{-1}(max(\bar{r}_{p_g}))))$, the largest value of $\bar{r}$ in the region.

In the process of identifying direct linkage between a gene and a

We evaluated the possibility that C'' (Section 4.1) could be used to narrow the scope of disease gene candidates

as a tool for prioritizing genes present in regions known to be associated with a disease.

We started with a set of 56 probesets representing 48 genes that were identified in a microarray study that searched for genes associated with skeletal abnormality [2]. We used the UCSC Genome Browser [7] to align these probesets to the March 2006 build of the human genome.

4.3 Gene Annotation

We evaluated the possibility of assigning new Gene Ontology annotations to a gene using only the identities and annotations attached to that gene's nearest neighbors. We measured distance between genes using the Pearson correlation coefficient r^2, and considered genes as neighbors where $r^2 > 0.5$. This value was selected from empirical observation of the minima and maxima of all correlation coefficients, as it identified at least one neighbor for more than 90% of all probesets (Figure 5).

4.3.1 Calibration

We were creating putative gene annotations using an automated process, so we wanted to observe the properties of pre-existing, non-computation annotations. This allowed us to choose parameters for the process that would yield only high-confidence annotations.

A subset of 1000 probesets were randomly selected, 536 of which had associated a non-computer-assigned annotation (Figure 5). There were a total of 1158 of these annotations, or approximately 2.16 annotations per probeset. For this same subset, we used the Bioconductor *GOstats* package to perform a hypergeometric test (hyperGTest) on the annotation of neighboring genes to measure which annotations were significantly enriched. We made an adjustment to our neighbor criterion and only used the 50 most correlated neighbors as input to hyperGTest in the event that there were more than 50 neighbors. This allowed us to compute the significance of neighbors' annotations in a reasonable amount of time.

hyperGTest produced a total of 30,280 possible annotations where $p - value < 0.05$. For each of the previously known, non-computer-assigned annotations, we looked up the p-value of the same annotation, if any, that came out of hyperGTest. We were able to recover $588/1158, 51\%$ of the annotations. The distribution of p-values signifying the test's confidence in each of those recoveries is given in Figure 5, and indicates that the majority of recalled annotations have $p - value < 0.001$, and nearly all have $p - value < 0.00001$. We chose $p - value < 0.00001$ as the threshold for high-quality, novel annotations suggested by hyperGTest. There were 966 annotations (approximately one per probeset) with p-value below this threshold and the distribution of their p-values is given in Figure 5.

Manual examination of these suggested annotations revealed that... XXX FAKE XXX many of the annotations were "near misses" to the existing annotation, as they were neighbors in the structure of the Gene Ontology graph. 90% of annotation were within one edge traversal of the Gene Ontology graph structure. XXX END FAKE XXX.

4.3.2 Prediction

XXX FILL THESE NUMBERS IN XXX

While the majority of the arraydesign probes for characterized, known genes,

More than XXX about 10,000 XXX probesets on the HG-U133_Plus_2 arraydesign are not assigned to any existing gene symbol. This typically means they were designed to measure a transcript that is computationally predicted only, and not supported by any *in vivo* observation. An additional XXX about 5,000-10,000 XXX probesets are assigned to a gene symbol that are supported by little or no *in vivo* data. This group contains predicted genes, as well as transcripts that have been observed in EST libraries.

Typical probesets in both of these groups have no annotation whatsoever. We were able to assign XXX annotations to XXX probesets (XXX %) in these groups. The data provided here represent an initial, and thus significant, step forward in the characterization of the roles of these hypothetical and rarely observed genes.

Our assignment of an additional XXX annotations to genes which are already characterized is also significant, as it suggests direct linkage between biological processes

previously known to be only indirectly related.

———-

References

[1] A Day, MR Carlson, J Dong, BD O'Connor, and SF Nelson. Celsius: a community resource for Affymetrix microarray data. *Genome Biol*, 8(6):R112, 2007.

[2] VA Funari, A Day, D Krakow, ZA Cohn, Z Chen, SF Nelson, and DH Cohn. Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression. *BMC Genomics*, 8:165, 2007.

[3] RC Gentleman, VJ Carey, DM Bates, B Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, AJ Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, JY Yang, and J Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[4] MA Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, GM Rubin, JA Blake, C Bult, M Dolan, H Drabkin, JT Eppig, DP Hill, L Ni, M Ringwald, R Balakrishnan, JM Cherry, KR Christie, MC Costanzo, SS Dwight, S Engel, DG Fisk, JE Hirschman, EL Hong, RS Nash, A Sethuraman, CL Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, SY Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, EM Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, and R White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, 2004.

[5] RA Irizarry, BM Bolstad, F Collin, LM Cope, B Hobbs, and TP Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31(4):e15, 2003.

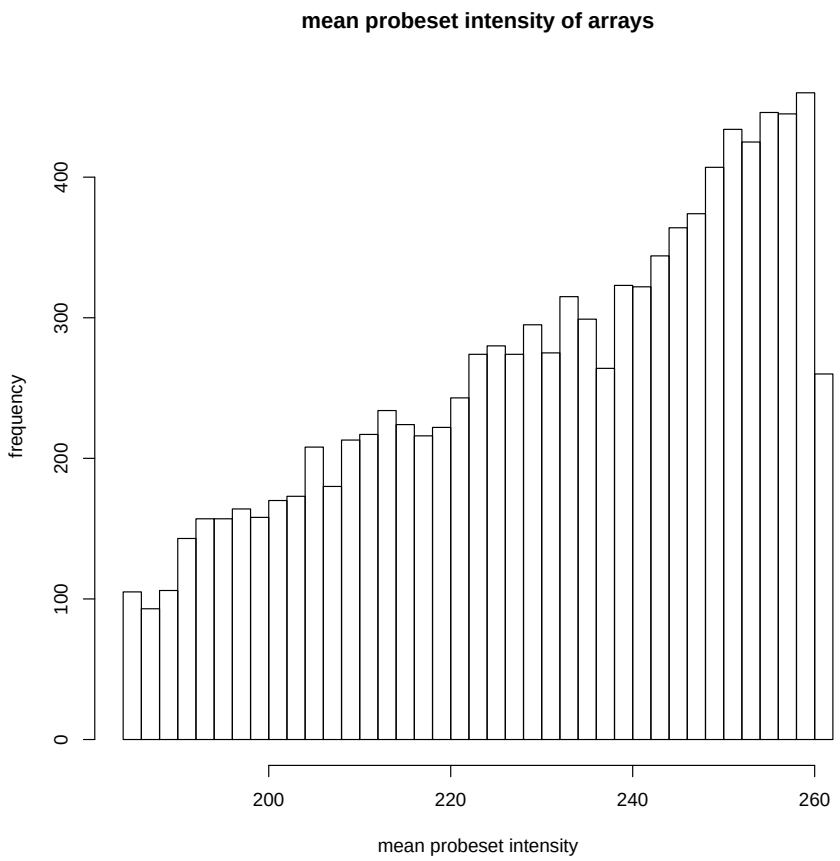
[6] AP Jackson, DP McHale, DA Campbell, H Jafri, Y Rashid, J Mannan, G Karbani, P Corry, MI Levene, RF Mueller, AF Markham, NJ Lench, and CG Woods. Primary autosomal recessive microcephaly (MCPH1) maps to chromosome 8p22-pter. *Am J Hum Genet*, 63(2):541–6, 1998.

[7] D Karolchik, RM Kuhn, R Baertsch, GP Barber, H Clawson, M Diekhans, B Giardine, RA Harte, AS Hinrichs, F Hsu, KM Kober, W Miller, JS Pedersen, A Pohl, BJ Raney, B Rhead, KR Rosenbloom, KE Smith, M Stanke, A Thakkapallayil, H Trumbower, T Wang, AS Zweig, D Haussler, and WJ Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–9, 2008.

[8] G Liu, AE Loraine, R Shigeta, M Cline, J Cheng, V Valmeekam, S Sun, D Kulp, and MA Siani-Rose. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*, 31(1):82–6, 2003.

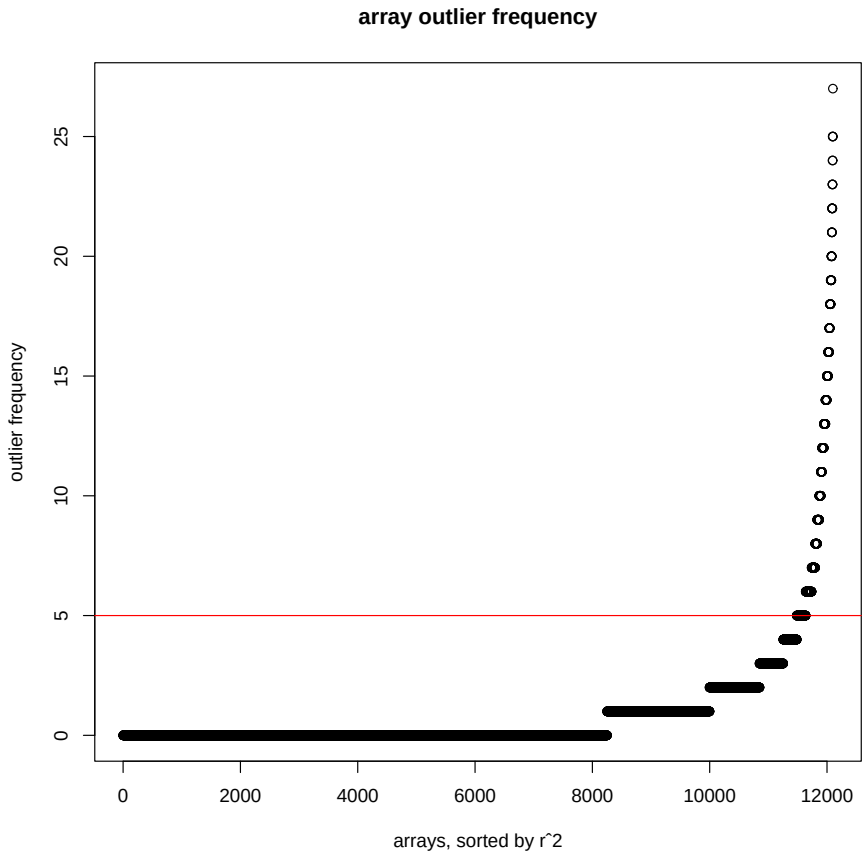
[9] EM Valente, F Brancati, and B Dallapiccola. Genotypes and phenotypes of Joubert syndrome and related disorders. *Eur J Med Genet*, 51(1):1–23, 0.

5 Figures



Mean probeset signal intensity of arrays. The mean value for all probesets was calculated for each array (x-axis) and the frequency of any given mean is plotted by bin (y-axis). Severa extremely bright arrays (n=15) are not shown.

Figure 1:
Mean probeset signal intensity of arrays. The mean value for all probesets was calculated for each array (x-axis) and the frequency of any given mean is plotted by bin (y-axis). Severa extremely bright arrays (n=15) are not shown.



Regressions of control probesets reveal aberrant arrays. Multiple regressions were performed for all 62 HG-U133_Plus_2 control probesets. Arrays (x-axis) are plotted versus the fraction of observations with regression residual $> 3\sigma$ (y-axis). A red horizontal line indicates a cutoff above which arrays are omitted from analysis.

Figure 2:
Regressions of control probesets reveal aberrant arrays. Multiple regressions were performed for all 62 HG-U133_Plus_2 control probesets. Arrays (x-axis) are plotted versus the fraction of observations with regression residual $> 3\sigma$ (y-axis). A red horizontal line indicates a cutoff above which arrays are omitted from analysis.

Figure 3: Correlation r^2 Minima vs. Maxima, All Probesets.

Figure 4: Fraction of Annotated Probesets.

Figure 5: Recall of Known Annotation.

Figure 6: P-values of High-Quality, Novel Annotation.