BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–5

Gene Characterization Throught Large-Scale Co-expression Analysis

Allen Day * Jun Dong * and Stanley F. Nelson †

Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The major goal of human genetics is to identify normal variants and mutations that lead to specific human phenotypes. Linkage and association are powerful tools to identify limited regions of the genome for further analysis to identify DNA variants causative of disease. However, the identification of individual genes responsible for the phenotype is often elusive. One way to accelerate the identification of gene mutations and variations is to integrate gene expression patterns, which provide an additional dimension of information.

Results: We describe a broad-based tool that uses all publicly available gene expression data on a single arraydesign to construct gene-gene expression correlations, and can be used by individual scientists to explore the gene-gene co-expression relationships present in this massive dataset to infer biological roles for individual genes. These matrices include virtually all human genes and aggregate data from all laboratories depositing data on the web, as available from Celsius, a community resource of image files and pre-processed Affymetrix microarray data.

We show the power of this approach to prioritize sequencing of individual genes within typical linkage intervals, and suggest novel candidate genes within linkage regions for Joubert syndrome and Type 2 Limb girdle muscular dystrophy. We demonstrate a technique that can reveal associations between genes and biological processes, given a seed set of characterized genes, and present a set of novel genes involved in cartilage development.

Availability: All data used in these analyses, as well as web-based exploratory tools, are available from `http://genome.ucla.edu/projects/XXX`. Additionally, software for programmatic access to these data is provided at the Bioconductor website `http://bioconductor.org` in the XXX package.

Contact: snelson@ucla.edu

1 INTRODUCTION

Genetic linkage and association studies lead to the identification of links between causal genes or genomic regions and their resulting phenotypes. These studies contain unique data, such as pedigree structures, that are not obtainable through other methods of inquiry. Data collected using DNA microarrays also give a unique perspective on genes associated with a particular phenotype, as the regulation of gene expression is independent of genomic location.

As the genome era draws to a close, the number of uncharacterized diseases that are influenced by a single or small number of genes has become very small. This is because the statistical power required to identify causal links between genes and phenotypes increases with the number of genes involved in the biological processes that manifest in the phenotype. Thus, a major theme in the post-genome era is that advances will come through integrative approaches that combine biological data from many sources [6, 22, 23].

Previous efforts have demonstrated the utility of using gene-gene co-expression patterns to perform gene characterization [19, 3, 10, 15, 17]. Indeed, this type of retrospective analysis of large volumes of data is characteristic of microarray studies in general. However, microarray studies frequently suffer from lack of statistical power because of the relatively small numbers of samples observed relative to the number of genes measured [30]. We addressed this problem of dimension transposition simply by increasing the amount of data in our analyses. That is, we constructed Celsius, a publicly accessible data warehouse of Affymetrix microarray data [4]. Celsius contains more than 120,000 microarrays, and for the current-generation human arraydesigns more than 12,000 arrays each. Further, microarrrays added to Celsius are processed in such a way to allow for progressive growth of the data set without needing to reprocess data.

Celsius contains very little experimental metadata. This position is in stark contrast to the microarray community at large, which favors metadata collection and modeling in the extreme [31, 21, 2]. While we concede that metadata provides additional statistical power in data analysis, policies that require metadata to be deposited concommitantly with assay measurements may actually be harmful. By increasing barriers to public release of data via metadata requirements, the amount of publicly available data is diminished. This reduction in data volume then effectively reduces our ability to draw biological conclusions from assay data alone.

Here, we demonstrate that fears of having vast volumes of unlabeled microarray data are unfounded. Quantitative microarray data are useful in their own right, and can be used to address biological problems completely unrelated to the initial study for which they were produced.

*these authors contributed equally to this work

†to whom correspondence should be addressed

2 APPROACH

Text Text Text Text Text Text Text Text Text Text Text

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 [11, 4] and denote the S=12826 array $\times$ P =54675 probeset matrix as M.

A cursory examination of 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 at least 3 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. Details of the exclusion procedure are shown in Algorithm 1. Essentially, 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 themselves 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 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 [16], Affymetrix publishes a probeset → 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).

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. Thus, we concluded that using the built-in control probesets was a robust way to identify aberrant arrays. 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. 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. This procedure is expressed Algorithm 2. Outlier frequencies per array are shown in Figure 5.

3.1.3 Correlating Genes Subsequent to 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 (Equation 1). $C\prime\prime$ was used in all results presented in Section 4.

$$C\prime\prime = cor(M^T\prime\prime) \quad (1)$$

3.2 Annotating Genes

For each probeset $p \in P$ on the HG-U133_Plus_2 arraydesign, we retrieved and sorted in descending order $r = C\prime\prime_p$. We took $r\prime$, the derivative of r, and used the R Bayesian Change Point *bcp* to identify δ, the index of the largest value of $r\prime$ that preceded a mostly-linear portion of the curve. The subset of probesets where $r > \delta$ were defined as Q, and used as input to the *hyperGTest* function of the *GOstats* package of Bioconductor [8] 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 applied Bonferroni correction to the p-values associated with each prediction by multiplying each p-value by the total number of predictions made for the corresponding probeset. We used these corrected p-values from predicted annotations N_p that were known to be non-computationally assigned from the *hgu133plus* package of Bioconductor [8] to establish a conservative cutoff, below which predicted annotations should all be high-quality. This process is presented in Algorithm 4.

3.3 Analyzing Linkage Regions

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 → probeset mapping available from NetAffx [16]. 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 $Q \times G (G \ni g)$ slab was retrieved from $C\prime\prime$ (Section 3.1.3), and row-summarized to produce a Q-length vector $\vec{r}$ of mean correlation coefficients to $G \ni g$. The algorithm for this procedure is presented in Algorithm 3.

4 DISCUSSION

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 [16], genomic alignment [13], and human-reviewed Gene Ontology (GO) Biological Process (BP) [9, 8] metadata.

4.1 Data Processing

All HG-U133_Plus_2 arrays (n=12826) were retrieved from Celsius [4]. We assessed the arrays using some simple quality control (QC) metrics, and excluded several hundred arrays as described in Section 3.1 yielding a 11636 array $\times$ 54675 column matrix, denoted $M\prime\prime$. We calculated the Pearson correlation coefficient for every pair of probesets in $M\prime\prime$, yielding a 54675 $\times$ 54675 correlation matrix, denoted $C\prime\prime$.

4.2 Disease Gene Recovery

Usually 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 [26, 12]. 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, human build hg18 (March 2006) [13]. 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 $L(q)$ 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 maximum value of $\bar{r}$ for named genes in the simulated linkage region. If the gene identified was present in G we evaluated the result as positive. In the event that genes not in G met the criterion of $max(\bar{r})$, we evaluated the result as negative.

We first applied our method to the limb-girdle musscular dystrophy, type 2 (LMGD2) phenotype. There are 11 genes known to be associated with LMGD2. For each gene, we considered all probesets targeting a named gene within a 6-megabase genomic region centered at the gene's locus. We calculated the mean correlation coefficient $\bar{r}$ to the 11-gene LMGD2 profile for each probeset within the region, but excluding any probesets targeting the gene used to select the region. In 55% (6/11) cases $\max \bar{r}$ corresponded to the causal gene for LMGD2.

Next, we applied this method to see if we could identify the four genes known to be associated with microcephaly. The purpose of this test was to confirm that the methods used in our LGMD2 trial would effectively identify genes for a completely different phenotype, as well as to see if the method was robust enough to identify the known gene given a much smaller profile for comparison. In 75% (3/4) of cases the most correlated gene with the other known microcephaly genes was correctly identified from a 6-megabase linkage region surrounding that gene.

Finally, we applied our scanning method to Joubert syndrome. Seven linkage regions for Joubert syndrome have been identified (JBTS1-JBTS7). Five of these have had the associated gene in the region identified (JBTS3=AHI1; JBTS4=NPHP1; JBTS5=CEP290; JBTS6=TMEM67; JBTS7=RPGRIP1L) [25, 18, 29, 1, 5] while regions JBTS1 and JBTS2 have so far only been linked to D9S158 [20] on chromosome 9 and a 17-megabase centromeric region of chromosome 11 [14, 28, 27], respectively.

The purpose of this third test was another instance of reproducibility of results, as well as to see if we could make a prediction as to the identity of the genes in remaining linked regions JBTS1 and JBTS2 for which a gene has not yet been identified. We were able to correctly identify 80% (4/5) of the five genes known to be associated with Joubert syndrome. A plot of r-values surrounding NPHP1 is given in Figure 5. We also show data from the Gene Expression Atlas [24] for the same region in Figure 5 to demonstrate that NPHP1 could not be identified merely by scanning this region for brain-specific or even brain-expressed genes.

Based solely on the correlation data, we are able to suggest the uncharacterized gene C9orf116 as the most likely candidate for the 6-megabase region surrounding D9S158 that is synonymous with JBTS1 (Figure 5).

We also examined JBTS2 to see if we could suggest any genes that might be associated with Joubert syndrome in this region as well. JBTS2 is a centromere-spanning 17-megabase region on chromosome 11 between markers D11S1915 and D11S4191 (Figure 5). We included an additional 3-megabases upstream and downstream of the outer markers. The best candidate for JBTS2 is AGBL2. However, this is a large region and there are highly-correlated genes on both sides of the centromere. For these reasons we suggest that the 17-megabase centromere-spanning linkage region JBTS2 is actually two separate linkage regions which we designate as JBTS2p and JBTS2q located on the p- and q-side of the centromere, respectively. The best candidate in JBTS2p is AGBL2. Additional candidates in JBTS2p are C11orf49 and probeset 229687_s_at. The best candidate in region JBTS2q is M4A8B. Additionally candidates in JBTS2q are SCGB1A1 and probeset 229688_at.

4.3 Gene Functional Assessment

Next, we considered whether these correlation data are more generally useful. In the case of identifying disease genes, the search space was constrained to a linkage region. However, some analytical methods are position-independent and seek to characterize all genes. One such task is the assignment of functional annotation.

To evaluate the performance of these correlation data for the assignment of gene annotation into biological processes, we selected 5 genes that were non-computationally annotated for the Gene Ontology term for "muscle contraction" (GO:0006936). The genes selected were UTRN, KCNQ1, MYOM1, SGCA, and ASPH. These are represented by 13 probesets on the HG-U133_Plus_2 arraydesign. Next, we calculated $\bar{r}$, the mean correlation coefficient to the 13-probeset profile, as described in Section 4.2, but did so genome-wide rather than restricting to a single linkage region. Then, we rank-ordered the set of all probesets and the set of the initially selected 13 probesets by $\bar{r}$. We observed that the muscle probesets were generally more highly-correlated to the profile, presumably because of the internal consistency of the set. However a large number of genes not used present in the profile also had high values of $\bar{r}$, and we sought to evaluate how well the 5-gene profile was able to recover other genes known to be involved in muscle contraction. To do so, we tested for enrichment of muscle contraction annotations for the probesets in the top decile of $\bar{r}$ values, partitioned into one-tenth percentile sized quantiles. We observed that enrichment for muscle contraction annotation is strongly correlated to an increased value of $\bar{r}$ relative to our 5-gene profile. The value of $\bar{r}$ versus the cumulative

distribution of probesets is given in Figure 5, along with the significance of the enrichment of muscle contraction annotation within each quantile.

We then applied this method to the results of a previous study by Funari, *et al*. that identified several previously unannotated genes that are expressed in cartilage tissue [7]. This study identified 114 genes represented by 133 probesets from the HG-U133A arraydesign that were highly expressed only in cartilage (Table 1). We performed the scan described, and were able to positively identify the correct gene 66% (75/114) times. A graphical representation of the 6-megabase region surrounding one cartilage gene, COL9A2, is given in Figure 5.

5 CONCLUSION

These correlation data are generally useful, and may be used to address a wide variety of biological questions. We have demonstrated that we are able to correctly identify causal genes within their linkage regions for several unrelated disease phenotypes. We also demonstrated the ability use these data to extrapolate what is known about a disease process to prioritize candidates within linkage regions for which the causal gene has not yet been identified.

Further, these correlation data are useful for characterizing non-disease phenotypes and can be used to characterize genes at global-scale as opposed to within specific genomic regions. We demonstrated this using a small set of muscle contraction genes to identify genes that are known to also be involved in muscle contraction according to the Gene Ontology Consortium. We reproduced this result and demonstrated that this type of enrichment analysis is not process- or annotation-specific by using a set of cartilage development genes identified from a previous microarray study to find more genes likely to be involved in cartilage development.

The versatility of this resource initially surprised us. The methods used to assemble the correlation matrix is completely metadata independent – only the genomic alignment of probe sequences and the quantitative measurements made by the microarray were used. The data set is also very heterogenous. It is composed from microarray data generated from thousands of individual experiments, from hundreds of experimenters, with each experiment using different biological materials and variations of the Affymetrix protocols.

What conclusions can we draw? XXX ... XXX

These results are in are somewhat surprising, but speak to the power...

FUNDING

Text Text Text Text Text Text

ACKNOWLEDGEMENT

Text Text Text Text Text Text Text Text.

REFERENCES

[1]L Baala, S Romano, R Khaddour, S Saunier, UM Smith, S Audollent, C Ozilou, L Faivre, N Laurent, B Foliguet, A Munnich, S Lyonnet, R Salomon, F Encha-Razavi, MC Gubler, N Boddaert, P de Lonlay, CA Johnson, M Vekemans, C Antignac, and T Attie-Bitach. The Meckel-Gruber syndrome gene, MKS3, is mutated in Joubert syndrome. *Am J Hum Genet*, 80(1):186–94, 2007.

[2]A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, CA Ball, HC Causton, T Gaasterland, P Glenisson, FC Holstege, IF Kim, V Markowitz, JC Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo, and M Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001.

[3]MR Carlson, B Zhang, Z Fang, PS Mischel, S Horvath, and SF Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7:40, 2006.

[4]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.

[5]M Delous, L Baala, R Salomon, C Laclef, J Vierkotten, K Tory, C Golzio, T Lacoste, L Besse, C Ozilou, I Moutkine, NE Hellman, I Anselme, F Silbermann, C Vesque, C Gerhardt, E Rattenberry, MT Wolf, MC Gubler, J Martinovic, F Encha-Razavi, N Boddaert, M Gonzales, MA Macher, H Nivet, G Champion, JP Berthlm, P Niaudet, F McDonald, F Hildebrandt, CA Johnson, M Vekemans, C Antignac, U Rther, S Schneider-Maunoury, T Atti-Bitach, and S Saunier. The ciliary gene RPGRIP1L is mutated in cerebello-oculo-renal syndrome (Joubert syndrome type B) and Meckel syndrome. *Nat Genet*, 39(7):875–81, 2007.

[6]CT Ferrara, P Wang, EC Neto, RD Stevens, JR Bain, BR Wenner, OR Ilkayeva, MP Keller, DA Blasiole, C Kendziorski, BS Yandell, CB Newgard, and AD Attie. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profi ling. *PLoS Genet*, 4(3):e1000034, 2008.

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

[8]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.

[9]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.

[10]S Horvath, B Zhang, M Carlson, KV Lu, S Zhu, RM Felciano, MF Laurance, W Zhao, S Qi, Z Chen, Y Lee, AC Scheck, LM Liau, H Wu, DH Geschwind, PG Febbo, HI Kornblum, TF Cloughesy, SF Nelson, and PS Mischel. Analysis of oncogenic signaling networks in glioblastoma identifi es ASPM as a molecular target. *Proc Natl Acad Sci U S A*, 103(46):17402–7, 2006.

[11]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.

[12]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.

[13]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.

[14]LC Keeler, SE Marsh, EP Leeflang, CG Woods, L Sztriha, L Al-Gazali, A Gururaj, and JG Gleeson. Linkage analysis in families with Joubert syndrome plus oculo-renal involvement identifi es the CORS2 locus on chromosome 11p12-q13.3. *Am J Hum Genet*, 73(3):656–62, 2003.

[15]I Lee, B Lehner, C Crombie, W Wong, AG Fraser, and EM Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet*, 40(2):181–8, 2008.

[16]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.

[17]R Nagarajan, N Le, H Mahoney, T Araki, and J Milbrandt. Deciphering peripheral nerve myelination by using Schwann cell expression profi ling. *Proc Natl Acad Sci U S A*, 99(13):8998–9003, 2002.

[18]MA Parisi, CL Bennett, ML Eckert, WB Dobyns, JG Gleeson, DW Shaw, R McDonald, A Eddy, PF Chance, and IA Glass. The NPHP1 gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome. *Am J Hum Genet*, 75(1):82–91, 2004.

[19]S Rossi, D Masotti, C Nardini, E Bonora, G Romeo, E Macii, L Benini, and S Volinia. TOM: a web-based integrated approach for identifi cation of candidate disease genes. *Nucleic Acids Res*, 34(Web Server issue):W285–92, 2006.

[20]K Saar, L Al-Gazali, L Sztriha, F Rueschendorf, M Nur-E-Kamal, A Reis, and R Bayoumi. Homozygosity mapping in families with Joubert syndrome identifi es a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. *Am J Hum Genet*, 65(6):1666–71, 1999.

[21]PT Spellman, M Miller, J Stewart, C Troup, U Sarkans, S Chervitz, D Bernhart, G Sherlock, C Ball, M Lepage, M Swiatek, WL Marks, J Goncalves, S Markel, D Iordan, M Shojatalab, A Pizarro, J White, R Hubley, E Deutsch, M Senger, BJ Aronow, A Robinson, D Bassett, CJ Stoeckert, and A Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3(9):RESEARCH0046, 2002.

[22]D Steinhauser, BH Junker, A Luedemann, J Selbig, and J Kopka. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, 20(12):1928–39, 2004.

[23]D Steinhauser, B Usadel, A Luedemann, O Thimm, and J Kopka. CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, 20(18):3647–51, 2004.

[24]AI Su, T Wiltshire, S Batalov, H Lapp, KA Ching, D Block, J Zhang, R Soden, M Hayakawa, G Kreiman, MP Cooke, JR Walker, and JB Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–7, 2004.

[25]B Utsch, JA Sayer, M Attanasio, RR Pereira, M Eccles, HC Hennies, EA Otto, and F Hildebrandt. Identifi cation of the fi rst AHI1 gene mutations in nephronophthisis-associated Joubert syndrome. *Pediatr Nephrol*, 21(1):32–5, 2006.

[26]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.

[27]EM Valente, SE Marsh, M Castori, T Dixon-Salazar, E Bertini, L Al-Gazali, J Messer, C Barbot, CG Woods, E Boltshauser, AA Al-Tawari, CD Salpietro, H Kayserili, L Sztriha, M Gribaa, M Koenig, B Dallapiccola, and JG Gleeson. Distinguishing the four genetic causes of Jouberts syndrome-related disorders. *Ann Neurol*, 57(4):513–9, 2005.

[28]EM Valente, DC Salpietro, F Brancati, E Bertini, T Galluccio, G Tortorella, S Briuglia, and B Dallapiccola. Description, nomenclature, and mapping of a novel cerebello-renal syndrome with the molar tooth malformation. *Am J Hum Genet*, 73(3):663–70, 2003.

[29]EM Valente, JL Silhavy, F Brancati, G Barrano, SR Krishnaswami, M Castori, MA Lancaster, E Boltshauser, L Boccone, L Al-Gazali, E Fazzi, S Signorini, CM Louie, E Bellacchio, E Bertini, B Dallapiccola, and JG Gleeson. Mutations in CEP290, which encodes a centrosomal protein, cause pleiotropic forms of Joubert syndrome. *Nat Genet*, 38(6):623–5, 2006.

[30]Y Wang, DJ Miller, and R Clarke. Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br J Cancer*, 98(6):1023–8, 2008.

[31]PL Whetzel, H Parkinson, HC Causton, L Fan, J Fostel, G Fragoso, L Game, M Heiskanen, N Morrison, P Rocca-Serra, SA Sansone, C Taylor, J White, and CJ Stoeckert. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–73, 2006.

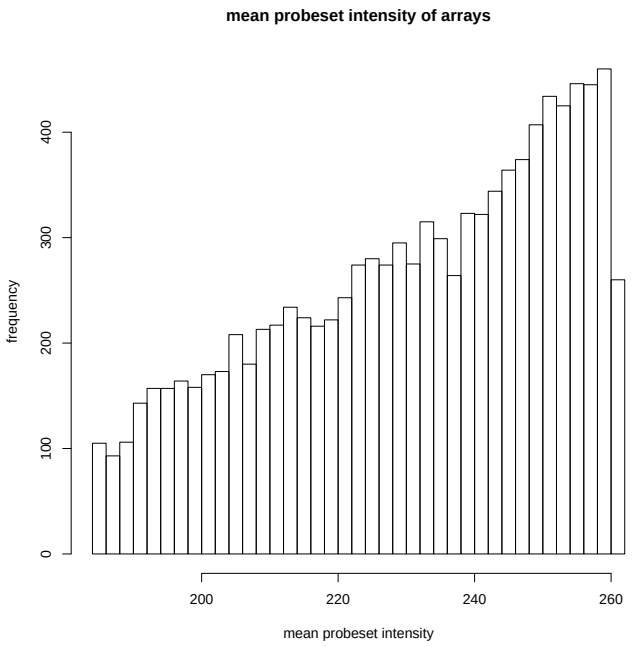

Fig. 1. Mean probeset signal intensity of arrays. The mean value for all probesets was calculated for each array (x-axis) and the frequency of any given mean is plotted by bin (y-axis). Several extremely bright arrays (n=15) are not shown.

Input: List of arrays
Output: List of arrays of typical brightness
$m = []$;
$good = []$;
foreach *Array* s **do**
  $m = [m, mean(M_s)]$;
end
$m_{trim} = sort(m)_{m*0.1} \ldots sort(m)_{m*0.9}$;
$\mu = mean(m_{trim})$;
$\sigma = standard_deviation(m_{trim})$;
foreach *Array* s **do**
  if $|mean(M_s) - \mu|/\sigma < 3$ **then**
    $good = [good, s]$;
  end
end
return *good*;

Algorithm 1: Identification and removal of dim and bright arrays

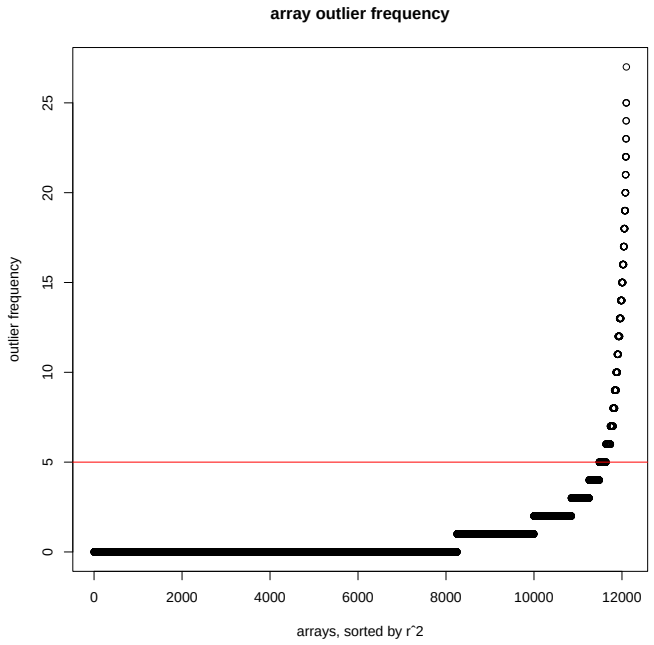

Fig. 2. Regressions of control probesets reveal aberrant arrays. Multiple regressions were performed for all 62 HG-U133_Plus_2 control probesets. Arrays (x-axis) are plotted versus the fraction of observations with regression residual $> 3\sigma$ (y-axis). A red horizontal line indicates a cutoff above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

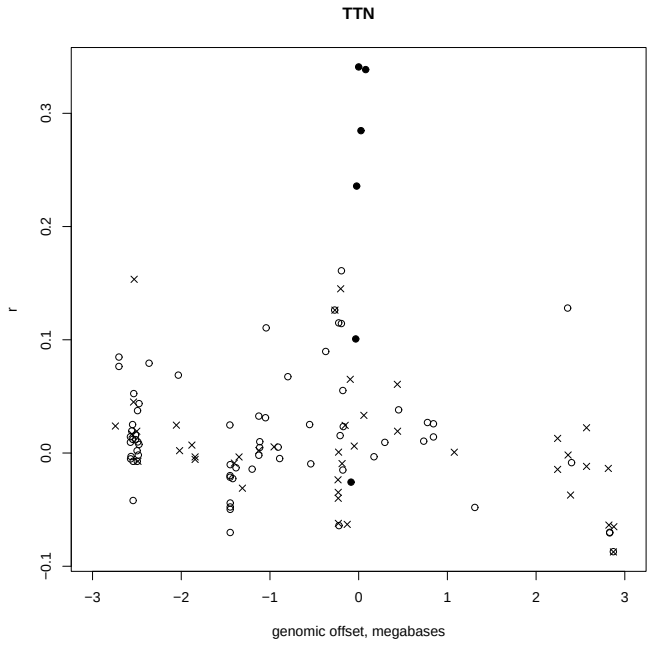

Fig. 3. Gene correlations to a list of LMGD2-associated genes within a 6-megabase region surrounding the location of an associated gene. The genomic position (x-axis) of probesets within a 6-megabase region centered at the location of a TTN, a gene known to be associated with LMGD2, is plotted versus the Pearson correlation coefficient r (y-axis) to a list of probesets targetting other genes known to be associated with Joubert syndrome across 11636 HG-U133_Plus_2 microarrays. Solid circles: probesets targeting TTN, $\times$s: probesets not designed to target a known gene, open circles: other probesets.

Input: List of arrays, List of control probesets
Output: List of arrays with consistent control probesets
$mark = []$;
$good = []$;
foreach *Control probeset* $c \in C$ **do**
  $lm = regression(M_{C \ni c}, response = M_c)$;
  $r = residuals(lm)$;
  $\mu = mean(r)$;
  $\sigma = standard_deviation(r)$;
  foreach *Array z-score* $z = (r - \mu)/\sigma$ **do**
    if $|z| > 3$ **then**
      $mark[z] + +$;
    end
  end
end
foreach *Array* a **do**
  if $mark[a]/length(C) < 0.05$ **then**
    $good = [good, a]$;
  end
end
return *good*;

Algorithm 2: Identification and removal of arrays with deviant control probeset signals

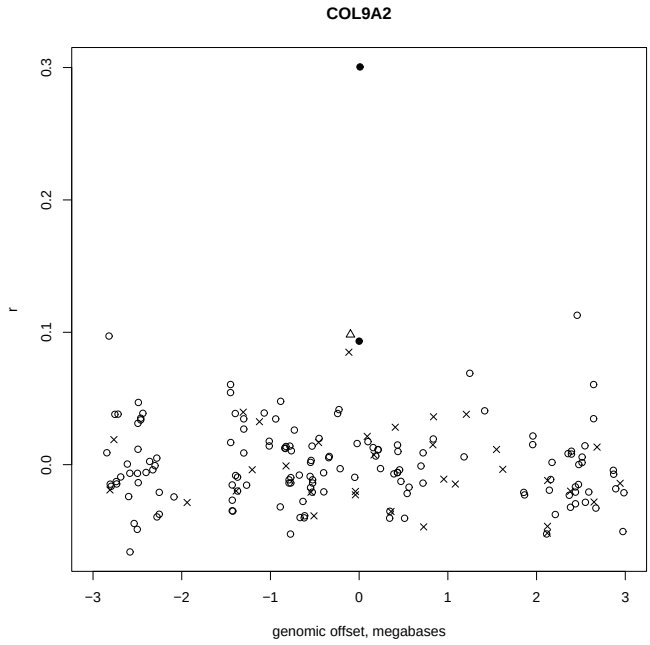

Fig. 4. Gene correlations to a cartilage-specific gene list within a 6-megabase region surrounding the location of an associated gene. The genomic position (x-axis) of probesets within a 6-megabase region centered at the location of COL9A2, a gene known to be associated with cartilage development, is plotted versus the Pearson correlation coefficient r (y-axis) to a list of cartilage-expressed probesets targetting other genes across 11636 HG-U133_Plus_2 microarrays. Solid circles: probesets targeting COL9A2, open triangles: other probesets in the cartilage-expressed list, $\times$s: probesets not designed to target a known gene, open circles: other probesets.

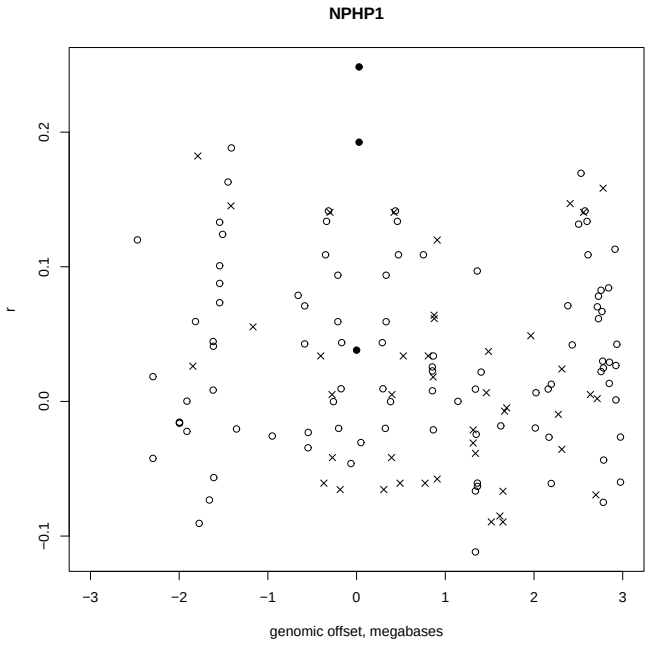

Fig. 5. Gene correlations to a list of Joubert syndrome-associated genes within a 6-megabase region surrounding the location of an associated gene. The genomic position (x-axis) of probesets within a 6-megabase region centered at the location of a NPHP1, a gene known to be associated with Joubert syndrome, is plotted versus the Pearson correlation coefficient r (y-axis) to a list of probesets targetting other genes known to be associated with Joubert syndrome across 11636 HG-U133_Plus_2 microarrays. Solid circles: probesets targeting NPHP1, open triangles: other probesets in the Joubert syndrome gene list, $\times$s: probesets not designed to target a known gene, open circles: other probesets.

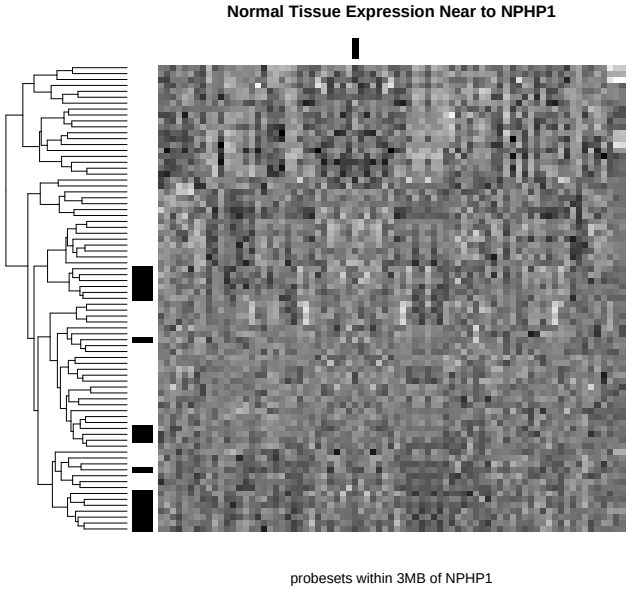

Fig. 6. Normal tissue expression surrounding NPHP1. Probeset position in ascending genomic order (x-axis) versus tissue (y-axis) from the GNF Expression Atlas 2 are presented as a tissue-clustered, column-scaled heatmap. Black=low expression, white=high expression. Black bars in the margin indicate brain tissue rows, and the column representing Joubert syndrome-associated gene NPHP1.

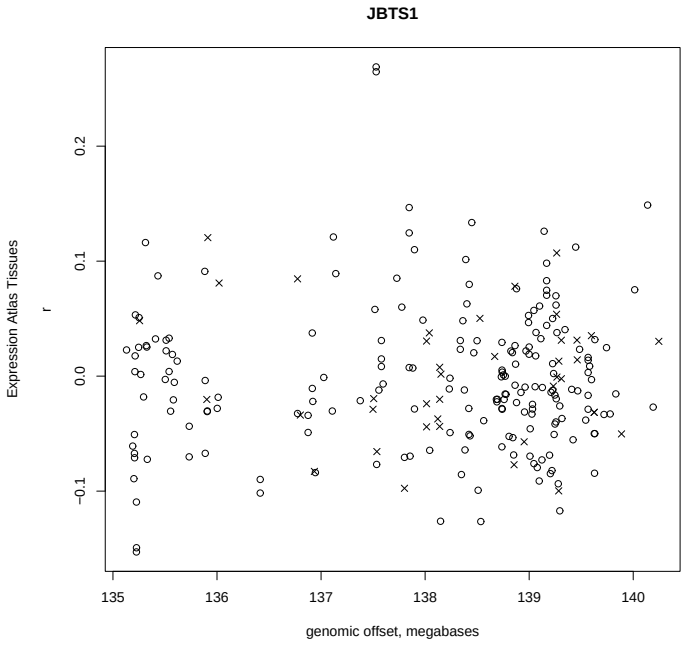

Fig. 7. Gene correlations to a list of Joubert syndrome-associated genes within 3 megabases of JBTS1 marker D9S158. The genomic position (x-axis) of probesets within a 6-megabase region on chromosome 9 centered at the location of marker D9S158, the peak of Joubert syndrome associated region JBTS1, is plotted versus the Pearson correlation coefficient *r* (y-axis) to a list of probesets targetting other genes known to be associated with Joubert syndrome across 11636 HG-U133_Plus_2 microarrays. ×s: probesets not designed to target a known gene, open circles: other probesets. The probesets located at 137.5 megabases for C9orf116 are the best candidate for a Joubert syndrome gene within this region.

Input: E, a list of "profile" gene symbols
Output: F, a list of "candidate" probesets / gene symbols
$P = []$;
$best = []$;
$hit = []$;
foreach *Gene symbol* $g \in G$ **do**
  $P = [P, probesets(g)]$;
end
foreach *Gene symbol* $g \in G$ **do**
  $b = genomic\_position(g)$;
  $b_{min} = b - 3{\times}10^6$;
  $b_{max} = b + 3{\times}10^6$;
  $Q = probesets\_in\_region(b_{min}, b_{max})$;
  $best[g] = -1$;
  foreach *Probeset* $q \in Q$ **do**
    $T = probesets(E) \ni probesets(g)$;
    $r = \frac{\sum_i^T C\prime\prime_t}{length(T)}$;
    if $r > best[g]$ **then**
      $bet[g] = r$;
      $hit[g] = gene\_symbol(q)$;
    end
  end
end
return hit;

Algorithm 3: A method for identifying the highest correlated gene to a gene list within a genomic region

Input: All HG-U133_Plus_2 probesets P
Output:
foreach *Probeset* $p \in P$ **do**
  $R = sort(C\prime\prime_p \ni p)$;
  $B = Bayesian\_change\_point(R\prime)$;
  $block = 0$;
  $offset = 1$;
  foreach $i \in 1...length(B)$ **do**
    if $B_i < 0.5$ **then**
      $block++$;
    else
      $block = 0$;
    end
    if $block >= 10$ **then**
      $offset = i$;
      Break;
    end
  end
end
return $hyperGTest(R_1, \ldots, R_{offset})$

Algorithm 4: A method for selecting probeset neighbors from correlation data

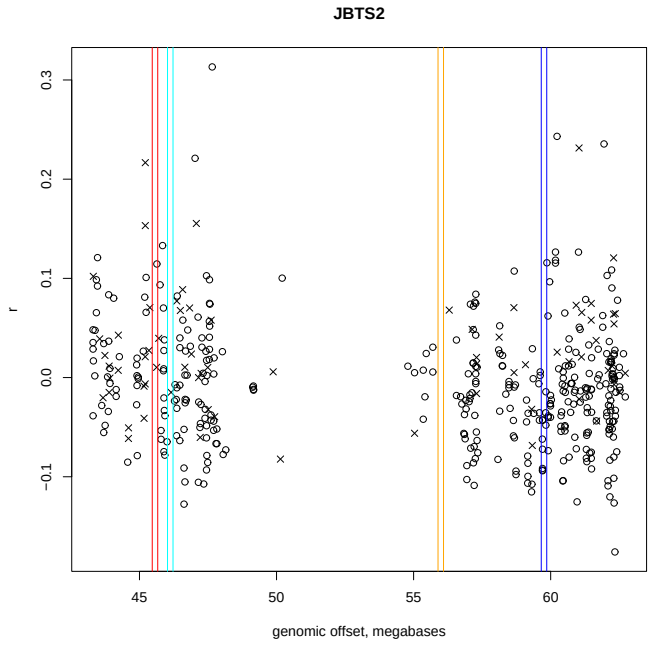

Fig. 8. Gene correlations to a list of Joubert syndrome-associated genes within 3 megabases of the 17-megabase region JBTS2. The genomic position (x-axis) of probesets within a 26-megabase region centered at JBTS2, a centromere-spanning 17-megabase region of chromosome 11 demarcated by D11S1915 and D11S4191 known to be associated with Joubert syndrome, is plotted versus the Pearson correlation coefficient r (y-axis) to a list of probesets targetting other genes known to be associated with Joubert syndrome across 11636 HG-U133_Plus_2 microarrays. $\times$s: probesets not designed to target a known gene, open circles: other probesets. Paired vertical lines indicate the posiiton of markers used in the mapping of this region. From left to right: D11S1915 (red), D11S1344 (cyan), D11S1313 (orange), D11S4191 (blue). The probeset located at 47 megabases for AGBL2 is the best candidate for a Joubert syndrome gene within this region.

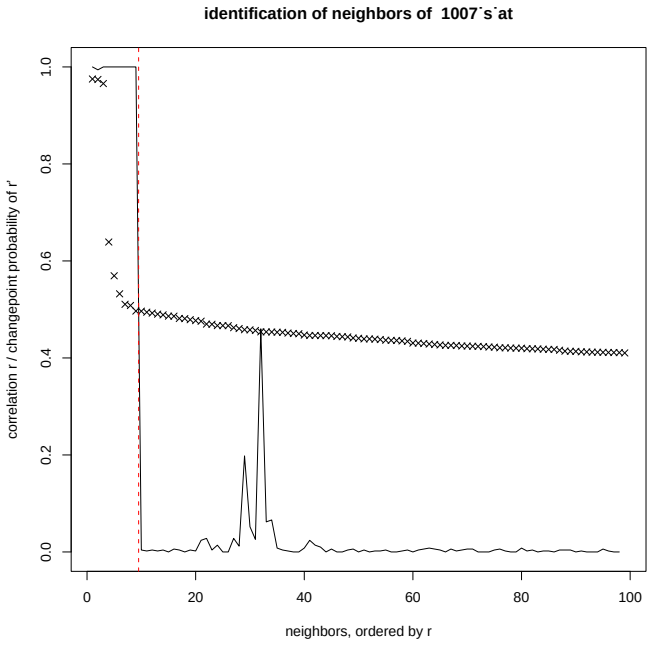

Fig. 9. Identification of most correlated neighbors for probeset 1007_s_at. Pearson correlation coefficient r (y-axis) is plotted for the 100 most highly correlated probesets to 1007_s_at (x-axis) using $\times$s. Solid lines indicates the probability of a change in slope of $r\prime$. A vertical dotted line indicates the position above which slope changes of $r\prime$ are common.

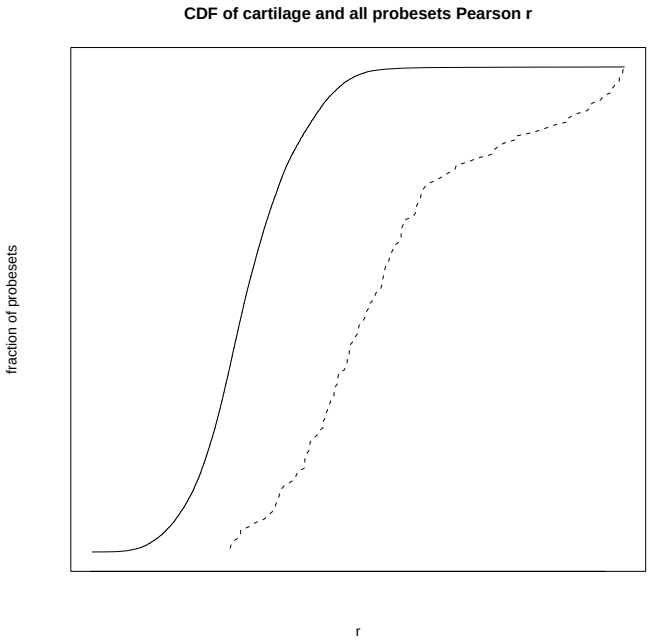

Fig. 10. What is plotted is the mean correlation coefficient to a cartilage profile of each probeset known to be cartilage-expressed (red, x-axis) vs. the total fraction of cartilage-expressed probesets ¡= a given value of r. A CDF, effectively. The same is plotted for all probesets (black). The purpose is to show that there are some probesets in the black curve that are not already known to be cartilage-expressed, i.e. they are not in the red curve. But these black points have higher correlation coefficient to the profile than some fraction of red points, so they are likely to be cartilage expressed. So is it clear? Is there a better way to do it?

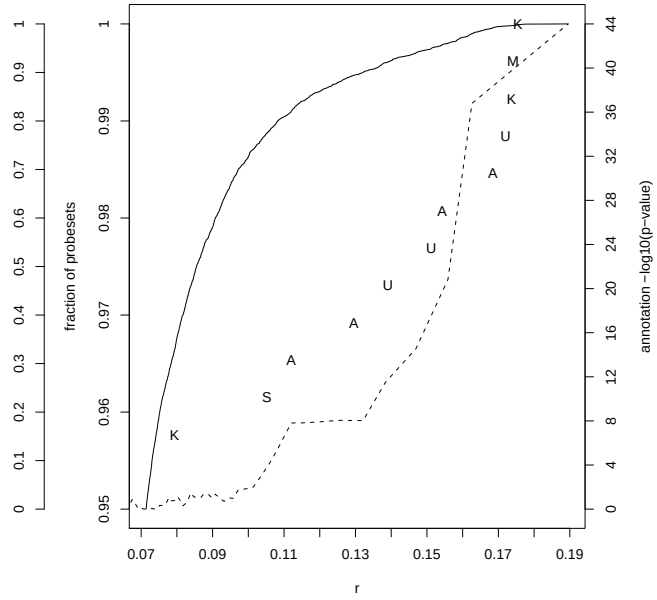

Fig. 11. Plot of $\bar{r}$ to a muscle profile of 5 genes (UTRN, KCNQ1, MYOM1, SGCA, ASPH; 13 probesets) annotated with "muscle contraction". The 13 probesets present in the profile (red, x-axis) and all probesets (black, x-axis) are plotted versus the cumulative fraction of cartilage-expressed probesets ¡= a given value of $\bar{r}$. Hypergeometric test enrichment of term "muscle contraction" $-log_10(p-values)$ within each 0.1%-sized quantiles for the top decile are also shown (cyan, right x-axis).