BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–4

Gene Characterization From Patterns In Transcriptional Co-regulation

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: Microarray experiments are frequently designed to characterize classes of samples and genes. However, a challenge in the analysis of these data is the relatively small number of samples observed compared to the number of genes measured for each sample. Unless bolstered by additional sample and gene metadata, results produced under this condition are less strong, and the data less suitable for re-analysis when considering other biological questions.

Results: We present a gene-gene transcriptional correlation matrix of *Homo sapiens* prepared from a very large, quality-controlled set of Affymetrix HG-U133_Plus_2 microarrays. We demonstrate that gene function can be predicted and that linkage regions can be reduced to caual genes with high precision using this resource, even in the complete absence of sample metadata.

Availability: The authors have made gene expression and gene-gene correlation data available at `http://genome.ucla.edu/projects/XXX`. Software for manipulating these data is also provided here, as well as from the Bioconductor website `http://bioconductor.org`.

Contact: snelson@ucla.edu

1 INTRODUCTION

Text Text Text Text Text Text Text Text Text Text Text

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 [7, 2] and denote the S=

array $\times$ P =

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

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 [11], 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

*these authors contributed equally to this work

†to whom correspondence should be addressed

were observed more than 3 standard deviations more than 5% of the time (n=464) were removed from further consideration, leaving a matrix M'' with 11636 arrays and

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'' matrix to calculate C'', a

$\times$

matrix of Pearson correlation coefficients for every pair of probesets (Equation 1). C'' was used in all results presented in Section 4.

$$C'' = cor(M^T{}'') \quad (1)$$

3.2 Annotating Genes

For each probeset $p \in P$ on the HG-U133_Plus_2 arraydesign, we retrieved and sorted in descending order $r = C''_p$. We took r', the derivative of r, and used the R Bayesian Change Point *bcp* to identify δ, the index of the largest value of r' 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 [5] 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 [5] 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 $\rightarrow$ probeset mapping available from NetAffx [11]. 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'' (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 [11], genomic alignment [9], and human-reviewed Gene Ontology (GO) Biological Process (BP) [6, 5] metadata.

4.1 Data Processing

All HG-U133_Plus_2 arrays (n=

) were retrieved from Celsius [2]. 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

$\times$

column matrix, denoted M''. We calculated the Pearson correlation coefficient for every pair of probesets in M'', yielding a

$\times$

correlation matrix, denoted C''.

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 [16, 8]. 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) [9]. 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 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 [4]. 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.

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 results in our cartilage trial were reproducible on a completely different set of genes, as well as to see if the method was robust enough to identify the known gene given

a much smaller profile for comparison. We were able to correctly identify 75% (3/4) of the microcephaly genes.

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) [15, 12, 19, 1, 3] while regions JBTS1 and JBTS2 have so far only been linked to D9S158 [13] on chromosome 9 and a 17-megabase centromeric region of chromosome 11 [10, 18, 17], 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 [14] 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 Annotation

We observed that each probeset $p \in P$ on the HG-U133_Plus_2 arraydesign typically has only a few highly correlated neighbor probesets, while the majority of probesets are slightly correlated to slightly anti-correlated. An example of the rapid drop in correlation coefficients for a single probeset 1316_at is shown in Figure 5. We hypothesized that the nearest neighbors were likely to be involved in the the same biological process, and that the observed transcriptional co-regulation would allow us to suggest roles for unannotated genes, and to suggest previously unknown roles for characterized genes.

To this end, we first evaluated how well the co-regulated neighbors could be used to reconstruct human-curated annotation that were already assigned to the gene using the Biological Process (BP) aspect of the Gene Ontology (GO) [6]. To do so for a single BP-annotated probeset, we first needed to discriminate between nearby and distant neighbors and did so using the *bcp*, the Bayesian Change Point library of the R programming language. The subset of nearby neighbors for which BP was available was then tested for category enrichment using a hypergeometric-based test provided by the Bioconductor [5] package *GOstats*. We tested for enrichment of all GO BP categories associated with any of the nearby neighbors, and so adjusted all p-values output by the hypergeometric test using Bonferroni correction in order to compensate for multiple hypothesis testing.

We then used the BP annotations that were *already known* from existing literature and "recovered" solely based on nearby neighbor annotations to estimate the p-value below which novel predictions should be accurate.

We were able to recover the correct annotation for XXX% of known annotations. The relationship between fractional recovery and confidence of the prediction (p-value) is given in Figure XXX. The vast majority of recovered annotations (XXX%) had an adjusted p-value of less than XXX, and we used this as a cutoff below which predicted annotations should be accurate.

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.

5 CONCLUSION

Text Text Text Text Text Text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text

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 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.

[3]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.

[4]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.

[5]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.

[6]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.

[7]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.

[8]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.

[9]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.

[10]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 identifies the CORS2 locus on chromosome 11p12-q13.3. *Am J Hum Genet*, 73(3):656–62, 2003.

[11]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.

[12]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.

[13]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 identifies a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. *Am J Hum Genet*, 65(6):1666–71, 1999.

[14]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.

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

[16]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.

[17]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.

[18]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.

[19]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.

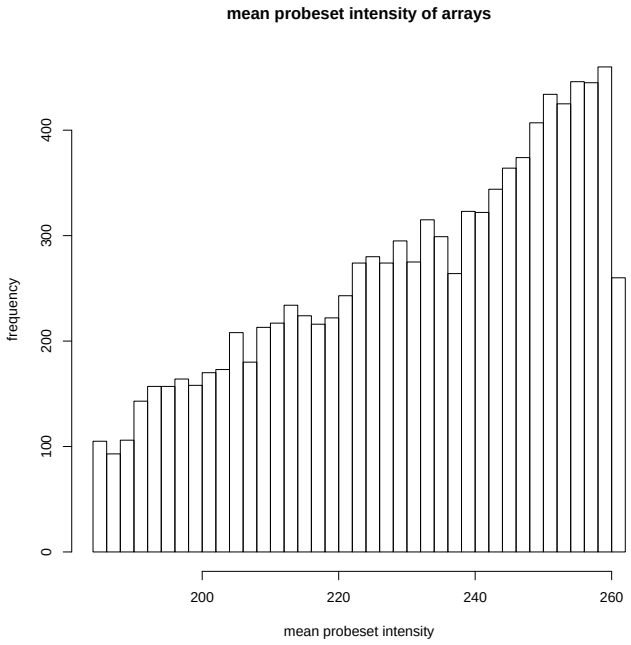

Fig. 1. Mean probeset signal intensity of arrays. The mean value for all probesets was calculated for each array (x-axis) and the frequency of any given mean is plotted by bin (y-axis). Several extremely bright arrays (n=15) are not shown.

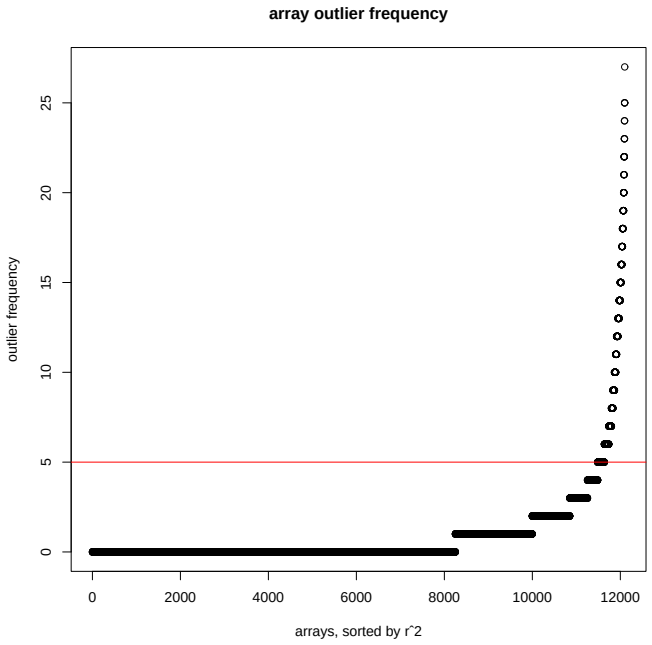

Fig. 2. Regressions of control probesets reveal aberrant arrays. Multiple regressions were performed for all 62 HG-U133_Plus_2 control probesets. Arrays (x-axis) are plotted versus the fraction of observations with regression residual $> 3\sigma$ (y-axis). A red horizontal line indicates a cutoff above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

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

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

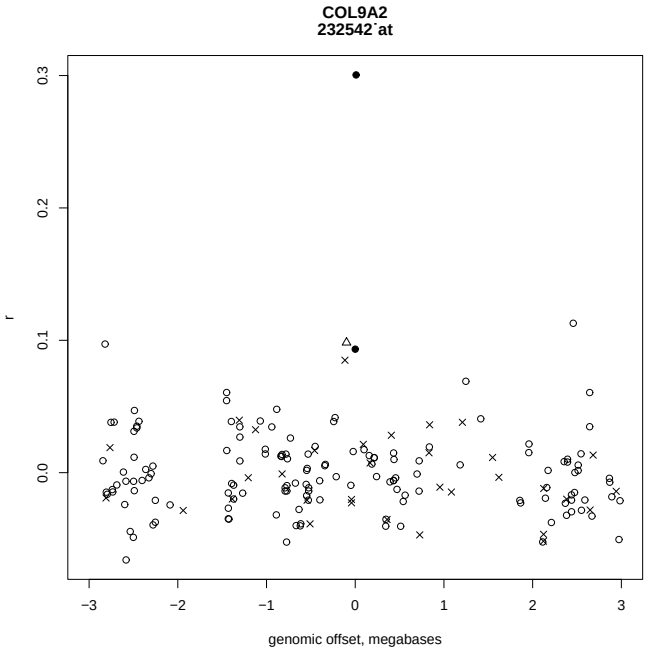

Fig. 3. Gene correlations to a cartilage-specific gene list within a 6-megabase region surrounding the location of an associated gene. The genomic position (x-axis) of probesets within a 6-megabase region centered at the location of COL9A2, a gene known to be associated with cartilage development, is 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.

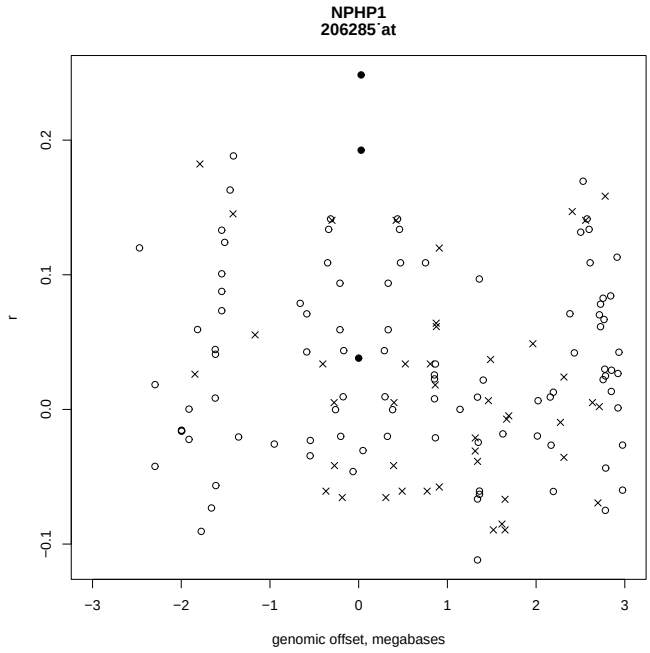

Fig. 4. Gene correlations to a list of Joubert syndrome-associated genes within a 6-megabase region surrounding the location of an associated gene. The genomic position (x-axis) of probesets within a 6-megabase region centered at the location of a NPHP1, a gene known to be associated with Joubert syndrome, 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.

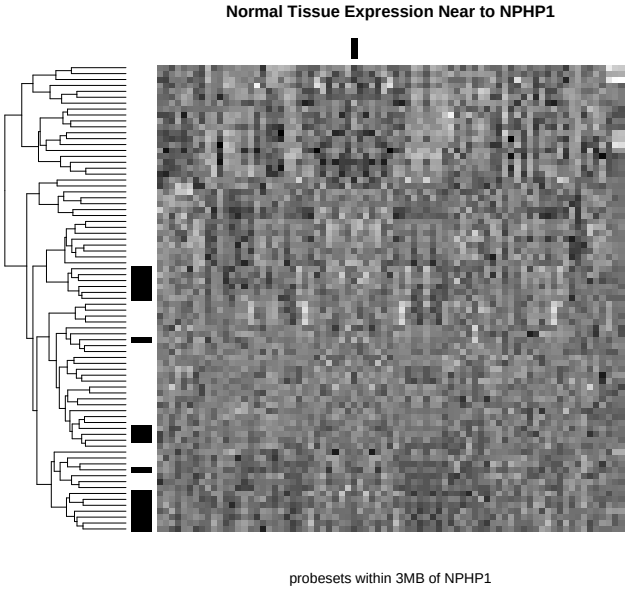

Fig. 5. Normal tissue expression surrounding NPHP1. Probeset position in ascending genomic order (x-axis) versus tissue (y-axis) from the GNF Expression Atlas 2 are presented as a tissue-clustered, column-scaled heatmap. Black=low expression, white=high expression. Black bars in the margin indicate brain tissue rows, and the column representing Joubert syndrome-associated gene NPHP1.

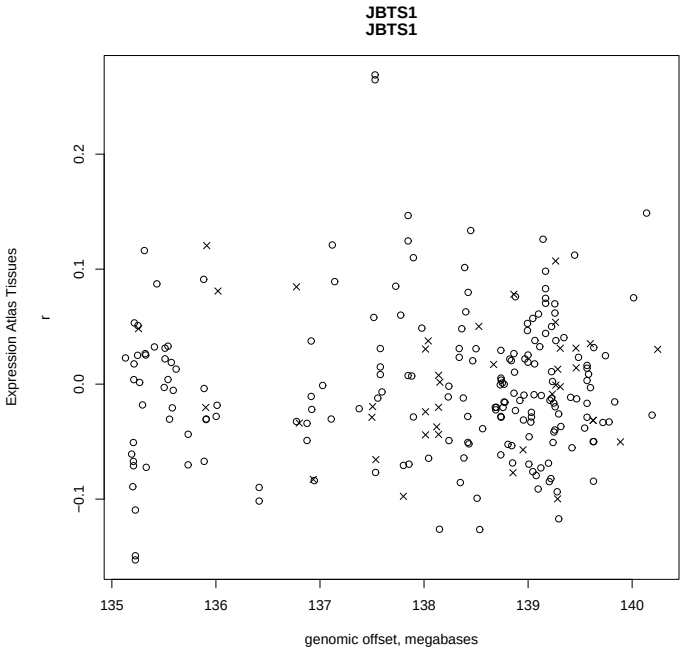

Fig. 6. Gene correlations to a list of Joubert syndrome-associated genes within 3 megabases of JBTS1 marker D9S158. The genomic position (x-axis) of probesets within a 6-megabase region on chromosome 9 centered at the location of marker D9S158, the peak of Joubert syndrome associated region JBTS1, is 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. $\times$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**
    $T = probesets(E) \ni probesets(g)$;
    $r = \frac{\sum_t^T C\prime\prime_t}{length(T)}$;
    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

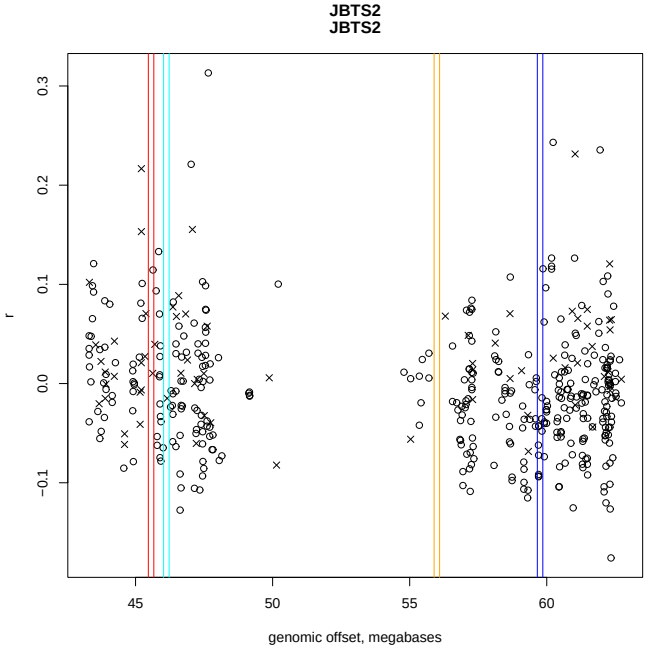

Fig. 7. Gene correlations to a list of Joubert syndrome-associated genes within 3 megabases of the 17-megabase region JBTS2. The genomic position (x-axis) of probesets within a 26-megabase region centered at JBTS2, a centromere-spanning 17-megabase region of chromosome 11 demarcated by D11S1915 and D11S4191 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.

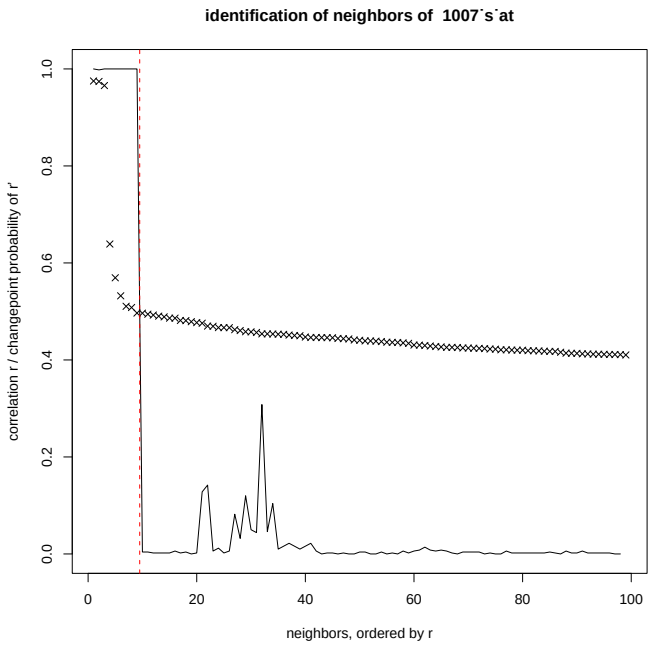

Fig. 8. Identification of most correlated neighbors for probeset 1007_s_at. Pearson correlation coefficient r (y-axis) is plotted for the 100 most highly correlated probesets to 1007_s_at (x-axis) using $\times$s. Solid lines indicates the probability of a change in slope of $r\prime$. A vertical dotted line indicates the position above which slope changes of $r\prime$ are common.