A Map of Human Gene Expression

W. B. Langdon

Departments of Mathematical, Biological Sciences and Computing and Electronic Systems
University of Essex, Colchester, CO4 3SQ, UK

Technical Report CES-486
ISSN: 1744-8050

Abstract

We have calculated the correlation between most human genes, using thousands of public Affymetrix HG-U133 +2 high-density oligonucleotide array (HDONAs). The correspondences show highly structured interactions between EBI Ensembl exons across a wide range of tissues and disease states taken from NCBI GEO. Eigen values are used to find and display the principle components of the gene expression mRNA data. The PCA analysis suggests almost all genes interact in a connected graph. There are thousands of strongly interacting genes but the whole network is sparse, with many genes not correlating strongly. So far, few power laws typical of "small world" networks and anticipated in gene regulatory networks have been found.

The ≈300 million correlations are organised by gene/exon and are available via a web interface.

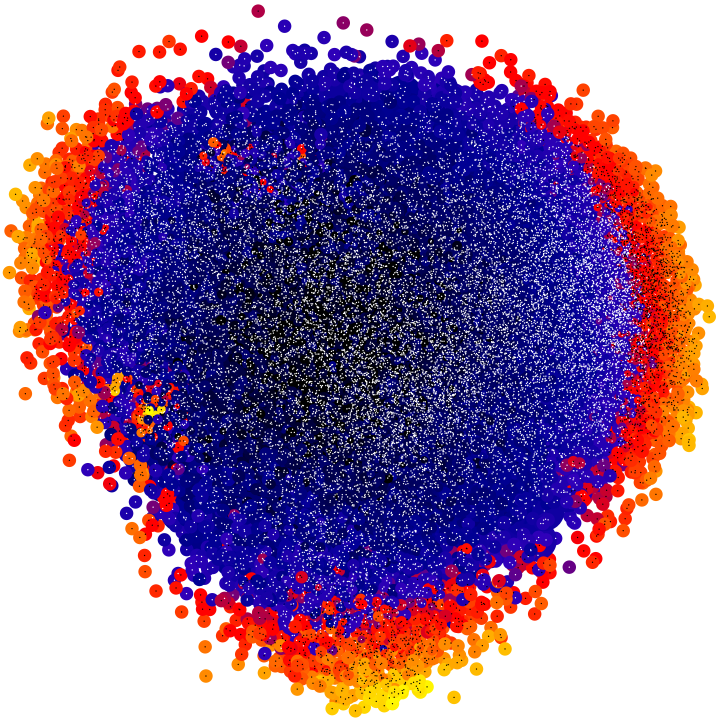


Figure 1: Principle component analysis is used to group 24 132 similarly behaved exons using data from 2757 GEO HG_U133 2+ GeneChips. Background colour indicates median correlation with ten nearest neighbours (black/blue low, yellow high). Each white dot shows an individual exon whose average correlation with its neighbours is less than 0.55. Exons with high average correlation shown by black dots.

1 Introduction

Much of post-genome Biology has concentrated upon either on DNA or on proteins and regarded RNA as a simple intermediate step. This is changing [RNA, 2007]. A menagerie of exotic species of RNA in addition to messenger RNA has been discovered. Much of its function is unclear. But its expression influences not only the production of proteins but also, directly or indirectly, the regulation of genes.

The availability of thousands of simultaneous measurements of RNA from a wide range of tissues and disease states gives a new data driven way to explore the complex network of gene product interactions. We place gene exons which interact in similar ways together in Figure 1.

2 Data Collection

The Affymetrix HG-U133 +2 GeneChip provides simultaneous 604 258 measurements of human gene expression (excluding controls). Some of these are duplicates and some cannot easily be assigned uniquely to exons. Using megablast, we have assigned 197 680 measurements to 29 637 Ensembl exons.

We down loaded all the HG-U133 +2 data in GEO. After cleaning and removing duplicates we had 2757 CEL files covering a wide variety of human tissue types and disease states, from numerous individuals. After normalising to a log scale and excluding data near spatial flaws [Langdon *et al.*, 2007b], for each exon we calculate the correlation between all the HG-U133 +2 probes which we mapped uniquely to that exon. The resulting "heat maps" can be found via http://bioinformatics.essex.ac.uk/users/wlangdon/. Affymetrix data for the same exon is often far from perfectly correlated. In some case we were able to advance technological reasons why some probes might not be only measuring their target transcript [Langdon *et al.*, 2007a; Langdon, 2008; Upton *et al.*, ; Langdon and Harrison, 2008]. [Langdon and Harrison, 200 provides a fast a simple rule (does it match `GGGG|CGCC|G(G|C){4}|CCC`) whereby we excluded probes based on their DNA sequence as being potentially suspect. For each exon we selected the probe which correlated (above 0.8) with most other members of the exon. (Excluding overlapping probes and controls). Ties were broken by choosing the maximum correlation and the earliest along the transcript. Where less than 20% of the possible correlations were less than 0.8 (or where the probes overlapped), we selected the probe with the maximum mean normalised intensity. This yielded 24 132 exons, which cover 14 288 of the 32 561 human genes described in Ensembl.

We calculated the correlations across all of GEO of each Ensembl exon with all the others. This is more than 290 million correlation coefficients across almost three thousand tissue samples.

3 Correlation Coefficients

The correlation coefficients cover a wide range of values. A small number of exons are highly correlated and some are inversely correlated. However most do not covary strongly with each other. Figure 2 shows the distribution of correlation coefficients is nearly symmetric about zero and looks approximately Normal. However the data are very far from random. (Since we are looking for correlations across thousands of samples, independent data would have observed correlations very close to zero. More than 95% would be expected to lie within 0.05 of zero.) In contrast the standard deviation of the observed distribution is 0.21. Figure 3 shows the scatter plot for the most negatively correlated pair of exons.

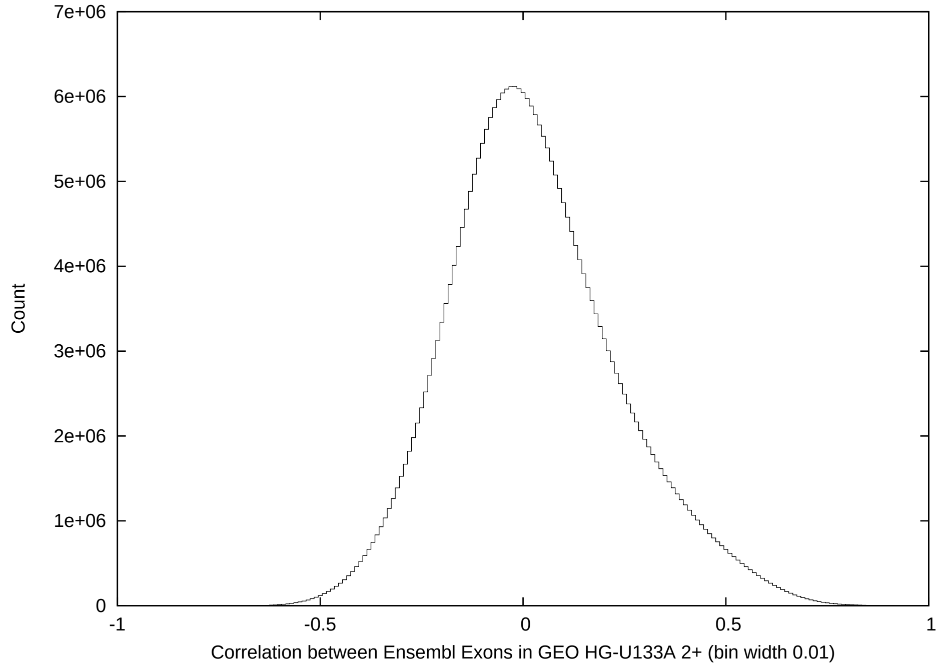


Figure 2: Distribution of correlation between human exons in all HG-U133 2+ GEO CEL files.

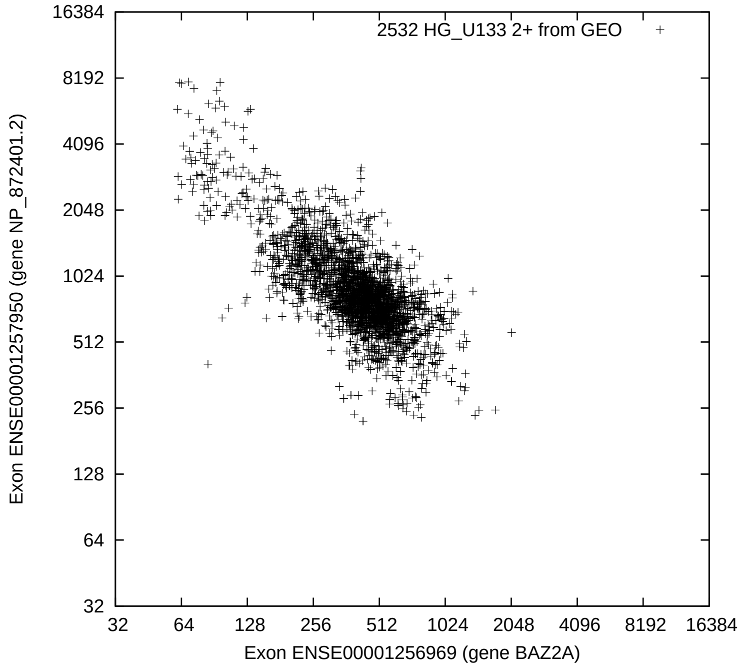


Figure 3: Example scatter plot of normalised HG_U133 +2 gene expression intensity measurements from GEO. (Pair of Ensembl exons with lowest correlation chosen. 225 reading close to a spatial flaw omitted).

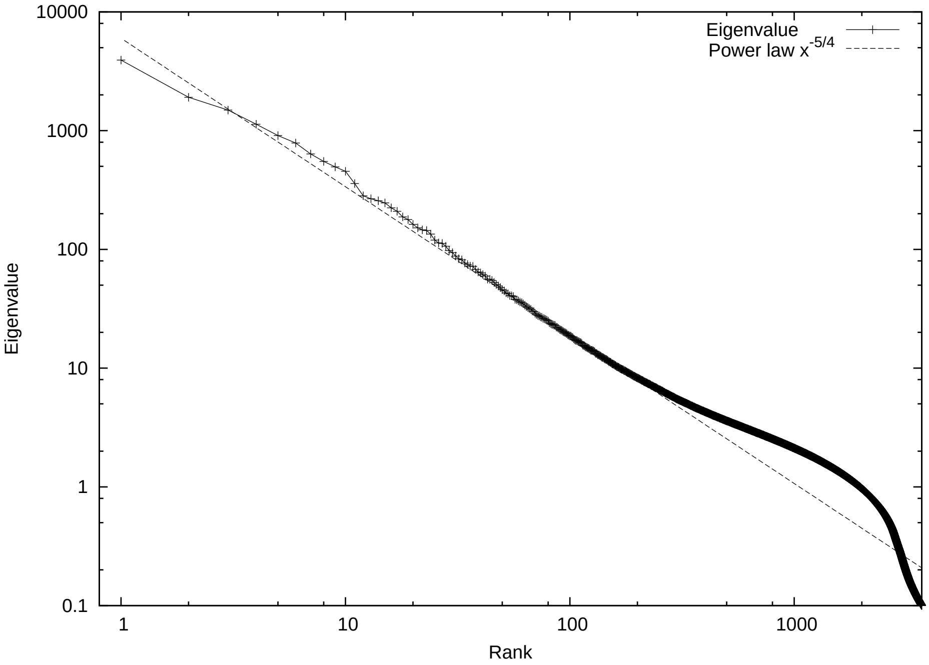


Figure 4: Eigenvalues (note log scales) of the correlation matrix between Human exons. The rapid decline indicates correlations can be approximated with only a few principle eigenvectors. The above average (i.e. > 1) eigenvalues follow an approximate power law. There are 24 132 eigenvalues. They are all positive [Everitt and Dunn, 1998].

4 Eigen Analysis

How should we display 290 million numbers?

They naturally fit into a $24\,132 \times 24\,132$ square matrix. Each element i, j of the matrix is the correlation between exon$_i$ and exon$_j$. Since this is the same as the correlation between exon$_j$ and exon$_i$, the matrix is square and all its diagonal elements are 1.

The eigenvalues and eigenvectors of such a matrix can be calculated. They can be thought of as rotations of the matrix. The eigenvalues can always be sorted and the eigenvectors corresponding to the largest eigenvalues can be thought of as its principle components. Given sufficient eigenvectors the original matrix can be reconstructed. However often the original matrix has some structure and only the subset of the eigenvectors corresponding to the largest eigenvalues are needed to approximate the matrix, cf. Figure 4. These eigenvectors capture the important aspects of the matrix. The first eigenvectors capture the essential nature of the matrix as successive eigenvectors are included they refine (either increasing or decreasing) elements within the matrix.

Where the matrix represents a sparsely connected graph, such as a small world network, the eigenvalues decrease rapidly and the components within the more important eigenvectors can be readily assigned a meaning in terms of the graph [Langdon *et al.*, 2008]. Usually eigenvectors are normalised so that they have unit length. In some cases, some of the components of an eigenvector can be very close to zero. In our case this means the corresponding exon plays very little part in the eigenvector. Therfore further compression of the 300 million data values can be had by only using the larger components of the principle eigenvectors. Figure 5 shows the number important components in eigenvectors corresponding to the fifty largest eigenvalues. Using only these, the almost 300 million or so human gene expression correlation coefficients can be approximated with only 375 013 numbers.

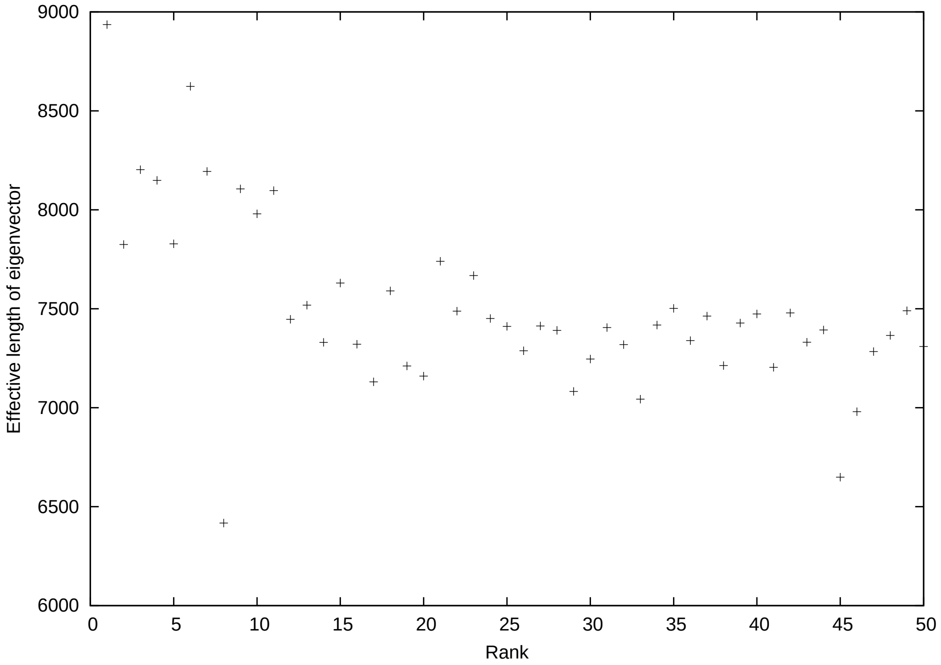


Figure 5: Each eigenvector has 24 132 components. The plot shows the smallest number of exons which need to be included in each of the 50 principle eigenvectors so to include $\geq 90\%$ of their length.

The eigenvectors corresponding to first and second largest eigenvalues are at right angles to each other and together form a plane onto which the original 24 312 axis of the correlation matrix can be projected. This is done in Figure 1. These is known as the (first two) principle components of the data and the plane captures, to the maximum extent possible by a linear transformation, the variation in the original GEO data [Everitt and Dunn, 1998].

Since the plane preserves as much as possible of the variation in the data, exons close to each other in it should be similarly behaved. This leads to the centre of the plane being dominated by exons with near zero correlation. Highly correlated exons are placed on the edge of the plane. See Figure 1.

5 Networks of Interactions

Figure 6 counts, for each exon, the number of other exons for which it has a strong positive correlation (i.e. > 0.578) and the number it has a strong negative correlation (i.e. < -0.410). (These thresholds were chosen to correspond to the lowest 1% and highest 1% of all the correlations. Cf. tails of Figure 2.) Most exons are strongly correlated (or anti-correlated) with very few other exons. The median number of negatively correlated exons is 48 (positive 31). In graph theory terms this corresponds to the degree of the node. I.e. the number of links it has. In other words the graph is very sparse in that most potential links are missing. Notice node degree does not follow a power law relationship which is typically found in small world or scale free networks.

Only a very few exons strongly vary in GEO with many others. Two distinct lobes can be seen in Figure 7. These correspond to the ≈ 350 exons which vary most strongly with 2000 or more others. Those exons with many negative correlations (perhaps caused by mutual inhibition or alternative splicing) tend not to have many strong positive correlations. And vice versa. That is, large numbers of positively correlated exons is matched by above average numbers of negatively correlated exons but the actual number of negative connections is much lower than the number of positive links.

If we create a graph where each node is an exon and strongly correlated (positive or negative) exons are connected by a link then we find almost all exons are in one connected central component. The central component is sparse with only 5 816 501 links between its 20 862 exons (about 2% of the fully connected graph). There are two isolated networks with 5 exons, ten with four, 38 with three, 93 pairs and 2920 exons which are not connected. Figure 8 shows the 1816 links that refer to exons which are strongly correlated with an example exon. (Figure 9 shows the top of the corrseponding web page.

6 Conclusions

290 million correlations between 24000 exons which together cover almost half the human genes, including the non protein coding genes, have been calculated. The strong interaction network is huge but sparse and contains thousands of genes which interact strongly with thousands of others. Conversely it contains tens of thousands of genes which interact strongly with less than a hundred others.

References

[Everitt and Dunn, 1998] Brian S. Everitt and Graham Dunn. *Applied Multivariate Data Analysis*. Arnold, 1998.

[Langdon and Harrison, 2008] W. B. Langdon and A. P. Harrison. Evolving regular expressions for genechip probe performance prediction. In *PPSN X*, LNCS, Dortmund, 13-17 September 2008. Springer.

[Langdon *et al.*, 2007a] W. B. Langdon, R. da Silva Camargo, and A. P. Harrison. Spatial defects in 5896 HG-U133A GeneChips. In Joaquin Dopazo, Ana Conesa, Fatima Al Shahrour, and David Montener, editors, *Critical Assesment of Microarray Data*, Valencia, 13-14 December 2007. Presented at EMERALD Workshop.

[Langdon *et al.*, 2007b] W. B. Langdon, G. J. G. Upton, R. da Silva Camargo, and A. P. Harrison. A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007. Submitted.

[Langdon *et al.*, 2008] W. B. Langdon, R. Poli, and W. Banzhaf. An eigen analysis of the GP community. *Genetic Programming and Evolvable Machines*, 2008. Online first.

[Langdon, 2008] W. B. Langdon. Evolving GeneChip correlation predictors on parallel graphics hardware. In Jun Wang, editor, *Proceedings of the IEEE World Congress on Computational Intelligence*, pages 4152–4157, Hong Kong, 1-6 June 2008. IEEE.

[RNA, 2007] Biology's big bang. *Economist*, 14 June 2007. Leader.

[Upton *et al.*,] Graham JG Upton, William B Langdon, and Andrew P Harrison. Incorrect measurement of gene expression by microarrays. *BMC Genomics*. Submitted.

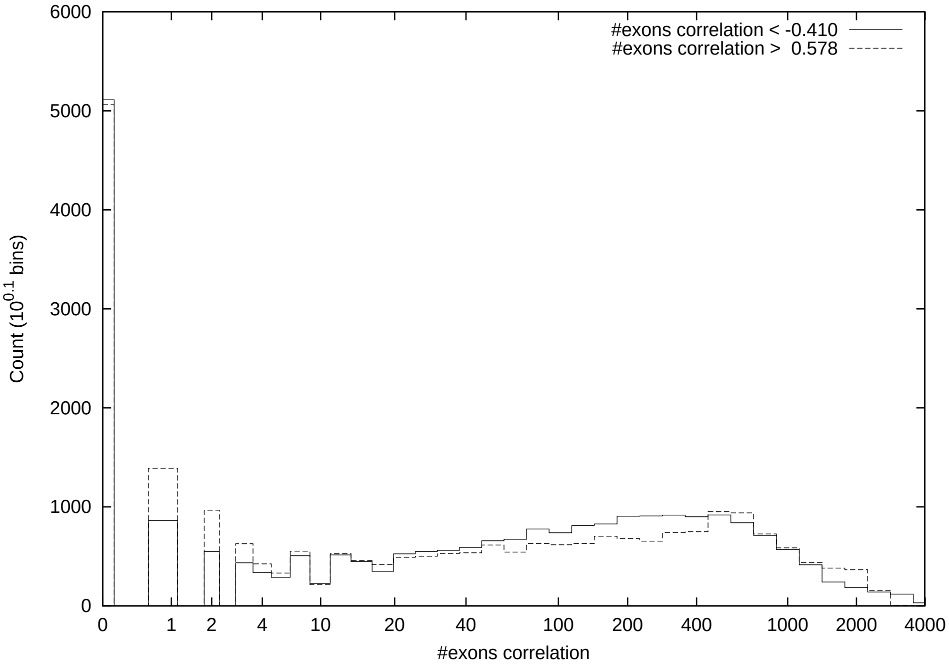


Figure 6: Number of exons which vary strongly together in 2757 HG-U133A +2 GEO CEL files. Most exons are strongly correlated (i.e. > 0.578) with more than 31 others. (For negatively correlated the threshold is < -0.410 and the median number of exons is 48.) However the value at zero shows, 6587 exons are not strongly correlated with any other and 5112 are not not strongly negatively correlated with any other. Together these include a subset of 2920 which are neither strongly correlated nor strongly anti-correlated with any other exon. (cf. bottom sqaure in Figure 7) Note non-linear horizontal scale and bin width.

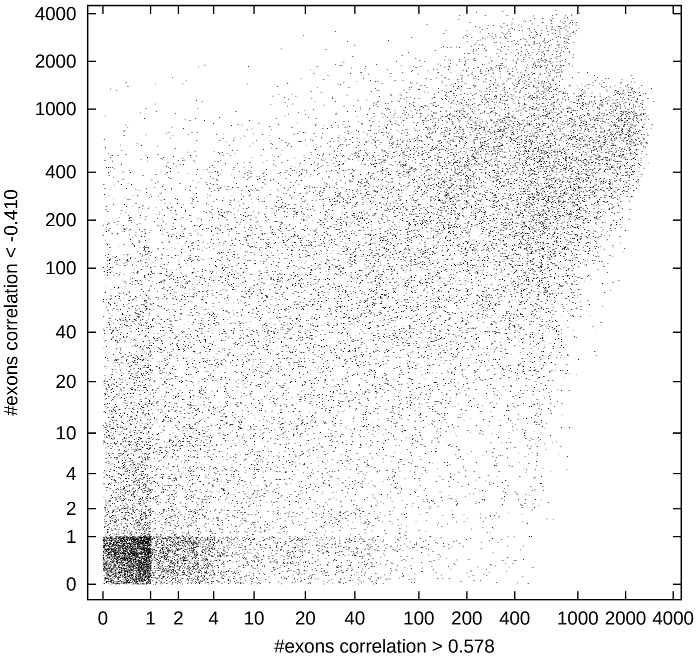


Figure 7: Same data as Figure 6. Note non-linear scales. Noise used to spread data.

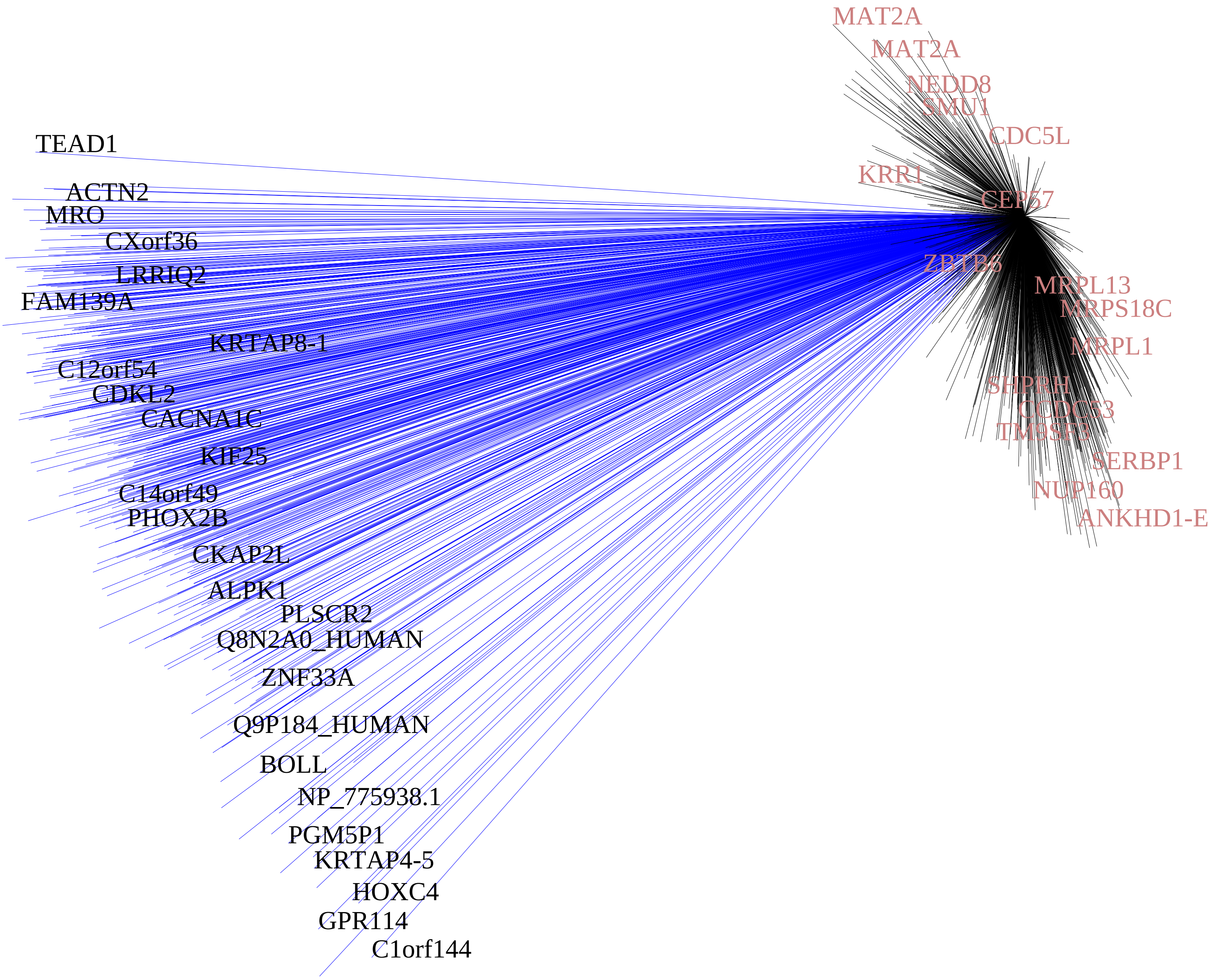
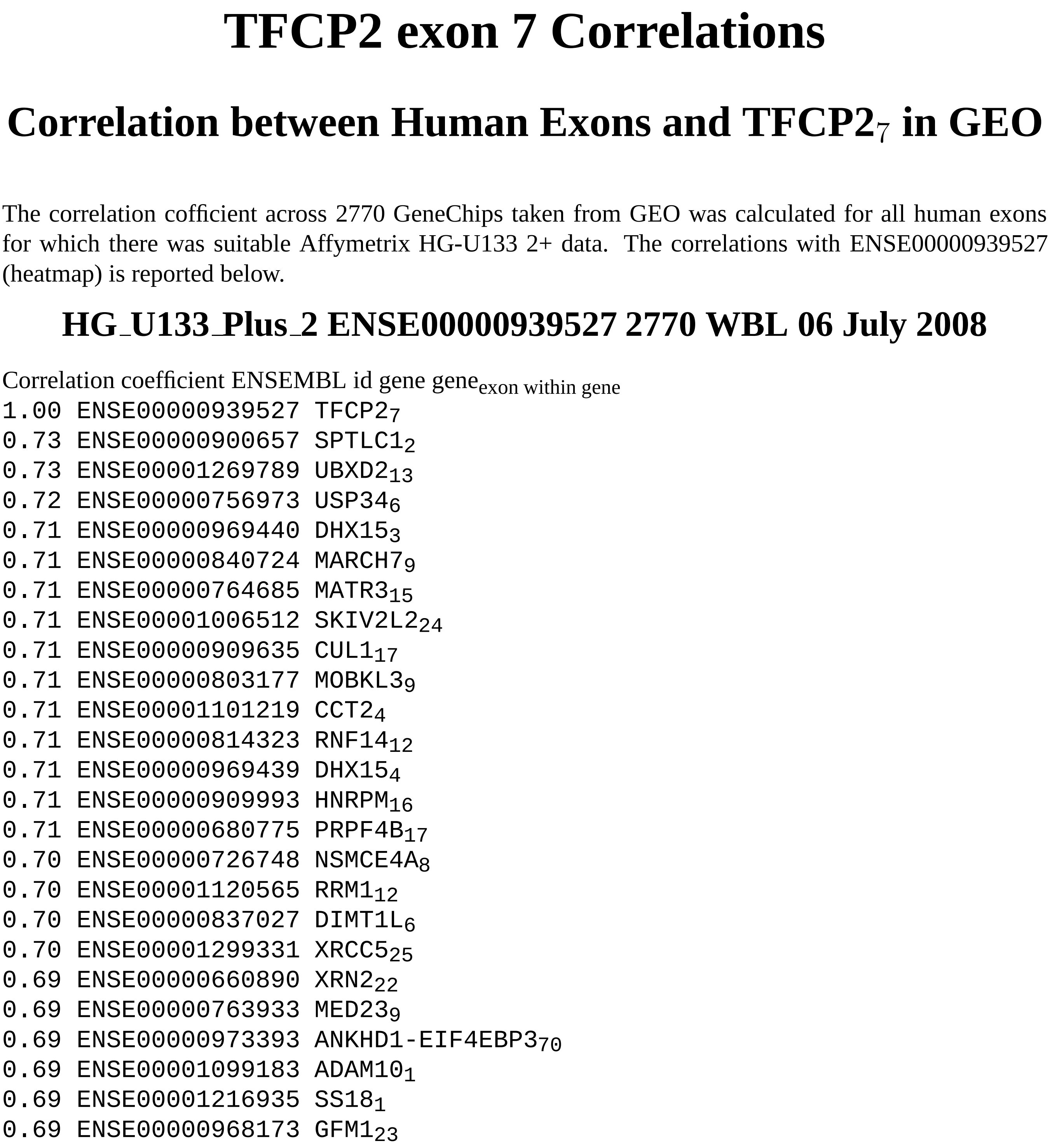


Figure 8: Exons which are strongly correlated with Ensembl exon ENSE00000939531 (blue indicates negative correlation). Labels are gene to which selected exons belong. Exons are spread out using two principle eigen values. I.e. using the positions given in Figure 1.



TFCP2 exon 7 Correlations

Correlation between Human Exons and $\mathrm{TFCP2}_{7}$ in GEO

The correlation cofficient across 2770 GeneChips taken from GEO was calculated for all human exons for which there was suitable Affymetrix HG-U133 2+ data. The correlations with ENSE00000939527 (heatmap) is reported below.

HG_U133_Plus_2 ENSE00000939527 2770 WBL 06 July 2008

Correlation coefficient	ENSEMBL id	gene $\mathrm{gene}_{\text{exon within gene}}$
1.00	ENSE00000939527	$\mathrm{TFCP2}_{7}$
0.73	ENSE00000900657	$\mathrm{SPTLC1}_{2}$
0.73	ENSE00001269789	$\mathrm{UBXD2}_{13}$
0.72	ENSE00000756973	$\mathrm{USP34}_{6}$
0.71	ENSE00000969440	$\mathrm{DHX15}_{3}$
0.71	ENSE00000840724	$\mathrm{MARCH7}_{9}$
0.71	ENSE00000764685	$\mathrm{MATR3}_{15}$
0.71	ENSE00001006512	$\mathrm{SKIV2L2}_{24}$
0.71	ENSE00000909635	$\mathrm{CUL1}_{17}$
0.71	ENSE00000803177	$\mathrm{MOBKL3}_{9}$
0.71	ENSE00001101219	$\mathrm{CCT2}_{4}$
0.71	ENSE00000814323	$\mathrm{RNF14}_{12}$
0.71	ENSE00000969439	$\mathrm{DHX15}_{4}$
0.71	ENSE00000909993	HNRPM_{16}
0.71	ENSE00000680775	$\mathrm{PRPF4B}_{17}$
0.70	ENSE00000726748	$\mathrm{NSMCE4A}_{8}$
0.70	ENSE00001120565	$\mathrm{RRM1}_{12}$
0.70	ENSE00000837027	$\mathrm{DIMT1L}_{6}$
0.70	ENSE00001299331	$\mathrm{XRCC5}_{25}$
0.69	ENSE00000660890	$\mathrm{XRN2}_{22}$
0.69	ENSE00000763933	$\mathrm{MED23}_{9}$
0.69	ENSE00000973393	$\mathrm{ANKHD1\text{-}EIF4EBP3}_{70}$
0.69	ENSE00001099183	$\mathrm{ADAM10}_{1}$
0.69	ENSE00001216935	$\mathrm{SS18}_{1}$
0.69	ENSE00000968173	$\mathrm{GFM1}_{23}$

Figure 9: First few lines of an example web page. Same exon as pictured in Figure 8.