

HIV Infection of Human T cells

Steen Knudsen

February 16, 2004

Contents

1	Introduction	2
2	Materials and Methods	2
2.1	Experimental Details	2
2.2	Statistical Analysis	3
2.3	Array Normalization	3
2.4	Expression index calculation	3
2.5	Clustering and PCA on chips	4
2.6	Classification	4
2.7	Statistical Significance	4
2.8	Analysis of Variance	4
2.9	Log fold change calculation	5
2.10	Gene Clustering	5
2.11	Correspondence Analysis	5
2.12	Gene Annotation	5
2.13	Protein Function Prediction	6
2.14	Promoter analysis	6
3	Results	10
3.1	Normalization	10
3.2	PCA and clustering of chips	10
3.3	Classification of chips	11
3.4	Statistical Analysis	12
3.5	Functional categories	19
3.6	Prediction of orphan function	22
3.7	Signal transduction pathway analysis	22
3.8	Metabolic pathway analysis	28
3.9	Clustering of Genes	28
3.10	Promoter analysis	34
3.11	Correspondence Analysis	37
4	Appendix A: parameters used in this report	39

Abstract

A DNA microarray experiment was performed using a chip of type HU6800. Principal Component Analysis and clustering was performed to reveal groupings in the samples. A statistical analysis was performed to reveal genes differentially expressed between the categories. A correspondence analysis was performed to identify genes associated with the individual categories

and experiments. Significantly regulated genes with unknown function were analyzed for properties of the encoded proteins and their function predicted using the ProtFun software. The TRANSPATH and KEGG databases were searched for differentially expressed genes annotated on known signal transduction or metabolic pathways. The promoter regions of differentially regulated genes were searched for regulatory elements.

1 Introduction

This report was generated automatically by the GenePublisher automatic DNA microarray analysis system¹.

Guide to interpretation of results: first look at the MVA plots before and after normalization to see if there are any obvious outlying chips (high variance and steep slope). Outlying chips may also be identified in the chip clustering, the PCA or the KNN classifier. Then look at the table of genes with significant changes in expression. Help in interpreting the biology of these genes may come from the LocusLink (if available), and from the TRANSPATH and KEGG analysis. Typically, one or more genes on this list need to be verified as differentially regulated by another method before publication, for example a quantitative PCR against the messenger RNA or an immunoassay against the protein. The gene cluster analysis is usually only of interest if there are more than two conditions compared in the experiment. Whether there are two or more conditions, you may look at the promoter analysis. The list of potential promoter elements may be overwhelming, but you can try to look for elements that are found by more than one method, or elements that show up in genes with a related role or function. For more information on the analysis methods used in this report, see Knudsen, S. (2002) *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley, New York.

2 Materials and Methods

This section describes the analysis in general terms. Details of the parameters and methods used can be found in the appendix section of this report.

2.1 Experimental Details

The purpose of this study is to measure the effect of HIV-1 on the transcription of genes in the infected host cell. The human cell line MT4 was infected *in vitro*

¹Knudsen, S., Workman, C., Sicheritz-Ponten, T., and Friis, C. (2003) GenePublisher: Automated Analysis of DNA Microarray Data. *Nucleic Acids Research*. Vol. 31, No. 13 3471-3476

with HIV-1. Control cultures were grown without HIV-1 infection. After 7 days of growth of the control cultures cells were harvested, RNA extracted and run on Affymetrix chips. These chips were compared to chips run on HIV-1 infected cultures harvested 7 days after infection. Replicates were performed to assure reproducibility and allow measurement of experimental variation.

2.2 Statistical Analysis

The statistical analysis was performed using the R statistics programming environment available from www.r-project.org. False positive predictions were assessed by multiplying P -values with the number of genes and by performing a permutation of the data.

2.3 Array Normalization

The individual chips were made comparable to each other by applying the *qspline*² method. Qspline is a robust non-linear method for normalization using array signal distribution analysis and cubic splines. Qspline fits cubic splines to the quantiles of the array signal distribution, and uses those splines to normalize signals dependent on their intensity.

2.4 Expression index calculation

For each gene, the expression index was calculated based on the probes by using the Li-Wong Model-Based Expression Index³. This model takes into account that probe pairs respond differently to changes in expression of a gene and that the variation between replicates is also probe-pair dependent.

The model-based expression index for each gene is calculated as:

$$\tilde{\theta} = \frac{\sum_N PM_n \phi_n}{N}$$

where ϕ_n is a scaling factor that is specific to probe PM_n and is obtained by fitting a statistical model to a series of experiments.

The model is run without the mismatch (MM) probes, using only perfect match (PM) probe information, by specifying "Background correction" as "bg.adjust".

²Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Saxild, H.H., Gautier, L., Nielsen, C., Nielsen, H.B., Brunak, S., and Knudsen, S. (2002) A new non-linear method for reducing variance between DNA microarray experiments. *Genome Biology* 3(9):0048.

³Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98:31–36.

This uses a model-based background subtraction from PM probes⁴. This latter PM-bg method is preferred over PM-MM methods because the resulting noise level is lower and because negative expression values are avoided.

2.5 Clustering and PCA on chips

Before any statistical analysis was performed, all genes on the chip were used for a hierarchical cluster analysis and principal component analysis to discover any grouping in the data (chips).

2.6 Classification

Three chip classifiers were automatically built on the input data, and cross-validated using the leave-one-out cross-validation principle as follows. K Nearest Neighbor (KNN) classification was performed for each chip by comparing it to the three nearest neighbors (K=3) among the remaining chips. The predicted class of the chips was the majority class among the three neighbors. For very small datasets, a K=1 classifier may be more accurate, so classification was performed with only one neighbor as well.

For the Nearest Centroid classifier (NC), each chip was compared to the centroids of the classes for the remaining chips. The predicted class of the chip was the class of the nearest centroid using Euclidean distance.

No feature selection was performed for the classifiers.

2.7 Statistical Significance

Differentially expressed genes between two categories of replicated experiments were identified by applying the *t*-test. The P-values calculated for each gene were used to calculate a False Discovery Rate⁵. It is possible to specify use of a paired *t*-test in the parameter file.

2.8 Analysis of Variance

Differentially expressed genes between more than two categories of replicated experiments were identified by applying an Analysis of Variance (ANOVA). The

⁴ Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2002) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Accepted for publication in Biostatistics., Available at <http://biosun01.biostat.jhsph.edu/ririzar/>

⁵Benjamini, Y., and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* 57:289-300

P-values calculated for each gene were used to calculate a False Discovery Rate⁶.

2.9 Log fold change calculation

The logarithm of the fold change of gene expression was calculated in order to obtain a symmetric distribution of regulation around zero (upregulated genes have positive logfold values, downregulated genes have negative logfold values). Expression values less than 1 were set to 1 before calculating the log fold change in order to avoid negative expression values that can occur if mismatch probe values are subtracted.

2.10 Gene Clustering

Hierarchical clustering was performed using the ClusterExpress software developed by Christopher Workman. Distances were calculated as the angle between vectors, and the expression values visualized as the logarithm of fold change relative to the average of category A.

2.11 Correspondence Analysis

Associations between categories and genes significant in the statistical test were visualized with correspondence analysis. Expression values were first converted to positive numbers by setting all negative numbers to zero. After correspondence analysis, genes and experiments were plotted in the same plot using the first two principal components⁷

2.12 Gene Annotation

Genes were annotated with Gene Ontologies (www.geneontology.org), which provides a unique identifier for each gene known to be responsible for a cellular process or function. Genes were grouped according to high-level function categories in the Gene Ontology database. Genes grouped under more than one functional category were only counted once. Genes were matched to the KEGG⁸ (Kyoto Encyclopedia of Genes and Genomes) description of known cellular pathways

⁶Benjamini, Y., and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* 57:289-300

⁷Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., and Vingron, M. (2001), Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA* 98:10781–10786.

⁸Kanehisa M, Goto S, Kawashima S, Nakaya A. "The KEGG databases at GenomeNet." *Nucleic Acids Res.* 2002 Jan 1;30(1):42-6.

(<http://www.genome.ad.jp>). For genes matching more than one pathway, only one pathway is shown. Genes were matched to the TRANSPATH⁹ database of signal transduction (www.gene-regulation.com). If genes match more than one pathway, only one pathway is shown.

2.13 Protein Function Prediction

For those genes where a gene ontology number has not been assigned and the function has not been inferred by homology to another protein, an attempt was made at predicting the function using the ProtFun¹⁰ method. The ProtFun methods predicts the function not based on homology, but based on properties of the protein sequence as well as predicted features such as post-translational modification.

2.14 Promoter analysis

Upstream regions (5000 bp for human, 300 bp for yeast) were extracted from the genes of each cluster using Ensembl (www.ensembl.org) or GenBank. The software program `saco_patterns`¹¹ was run on each cluster to identify significantly overrepresented patterns in the upstream regions. `saco_patterns` looks for conserved (identical) patterns in sequences, it does not allow for degeneration of the pattern.

The Gibbs sampler¹² was run on the same upstream regions. The Gibbs sampler looks for degenerate patterns which it tries to capture with a weight matrix description. In all sequences, the best match to this weight matrix is shown in the output. The Gibbs sampler starts with a new random matrix every time and is non-deterministic, meaning that it may give different results every time it is run.

The transcription factor binding sites in the TRANSFAC¹³ database were matched

⁹Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E. "TRANSPATH: an integrated database on signal transduction and a tool for array analysis." *Nucleic Acids Res.* 2003 Jan 1;31(1):97-100.

¹⁰Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A. F., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. (2002) Ab initio prediction of human orphan protein function from post-translational modifications and localization features. *Journal of Molecular Biology* 319:1257-1265

¹¹Jensen, L.J. and S. Knudsen, (2000) Automatic Discovery of Regulatory Patterns in Promoter Regions Based on Whole Cell Expression Data and Functional Annotation. *Bioinformatics* 16:326-333.

¹²Lawrence, Altschul, Boguski, Liu, Neuwald & Wootton (1993) "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment", *Science* 262:208-214.

¹³Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. "TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 2003 Jan 1;31(1):374-8.

against the same upstream regions. Factor matrices with hits more than 95% of the maximal score of the matrix were recorded.

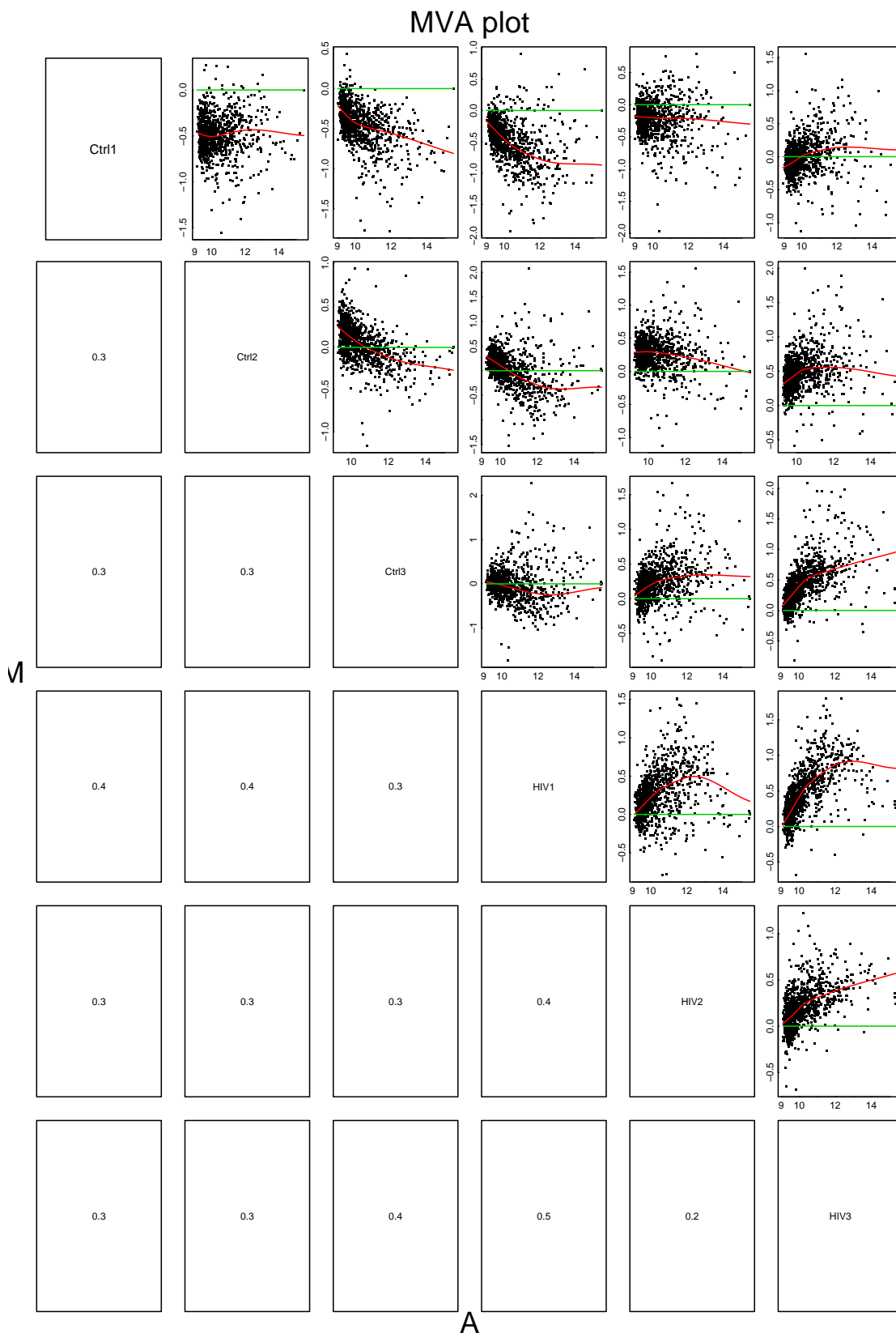


Figure 1: M versus A for all chip-to-chip comparisons before normalization. The diagonal shows the names of the chips being compared. The lower triangle shows the variance of the ratios between the two chips being compared. Two identical chips should have a variance of zero. Look for bad chips in this plot. They are revealed by a higher variance in comparisons to the other chips and by a consistent curvature when compared to other chips (indicating low amount of hybridization). The comparison is limited to 10 chips versus 10 chips.

MVA plot

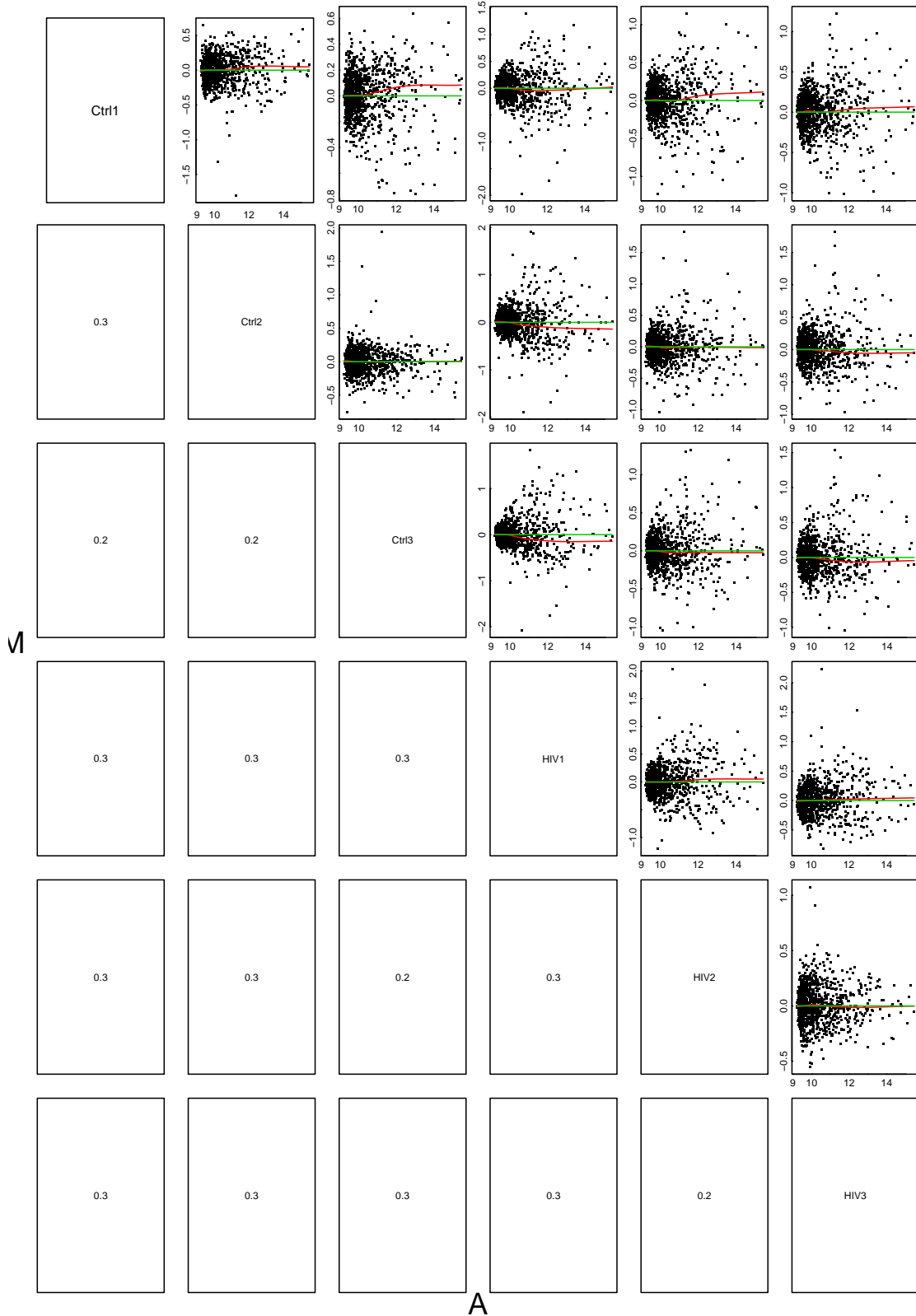


Figure 2: M versus A for all chip-to-chip comparisons after normalization. The diagonal shows the names of the chips being compared. The lower triangle shows the variance of the ratios between the two chips being compared. Two identical chips should have a variance of zero.

The comparison is limited to 10 chips versus 10 chips.

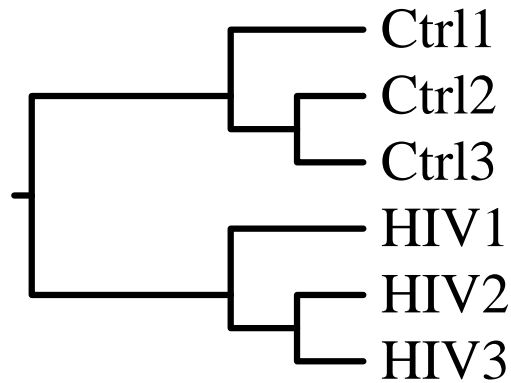


Figure 3: Hierarchical clustering of categories using Euclidean distance between vectors of all genes and complete linkage.

3 Results

3.1 Normalization

Figure 1 shows a comparison of all chips before normalization. This is a so-called M versus A plot; instead of plotting each probe on one chip against each probe on another, the scales are changed so it plots, for each probe, the logarithm of the ratio of expression between the two chips as a function of the logarithm of the mean of the expression of the two chips. Two identical chips would yield a straight, flat line through zero. Two comparable chips ideally have a straight, flat line through zero and a few probes off the line indicating differential expression. Deviation of the line from zero reveals a need for normalization before the two chips can be compared, and deviation from a straight line reveals a need for non-linear normalization (different normalization factors for highly and weakly expressed genes).

Figure 2 shows the comparison of all the chips after normalization.

3.2 PCA and clustering of chips

All chips were clustered based on the Euclidean distance of all genes (Figure 3). Such a clustering shows the relationship between individual chips, in particular

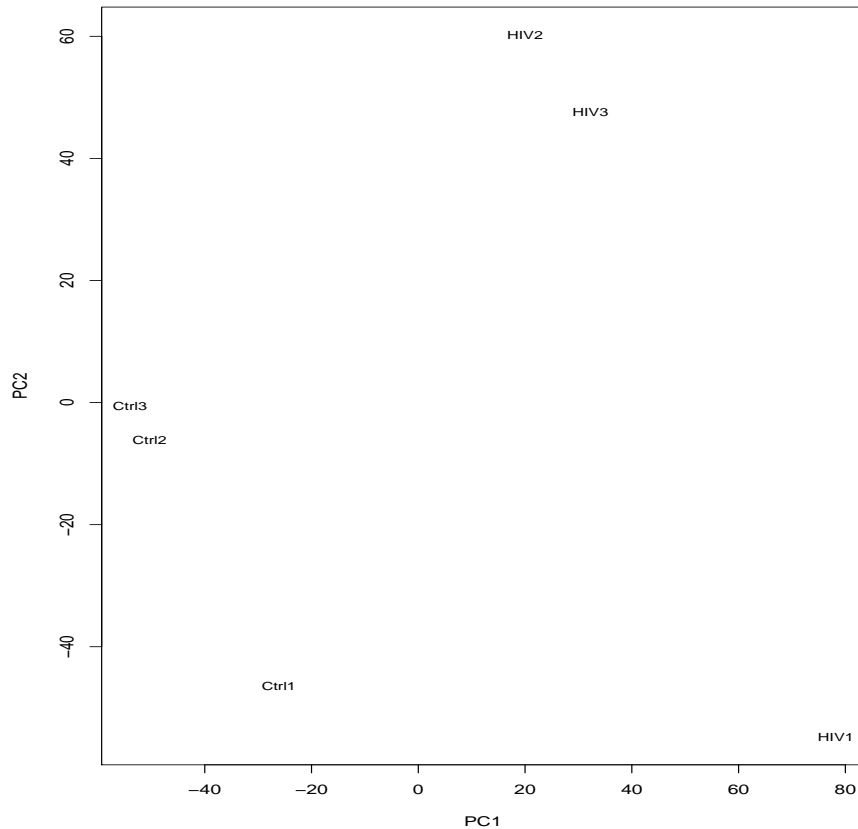


Figure 4: Principal Component Analysis showing all chips plotted according to their first two principal components.

if the cluster together in the categories they have been assigned. If they do not cluster together in the categories assigned, or if one chip clusters separately, this may be indicative of a problem, for example an outlier (bad quality) chip. In that case the analysis should be repeated without that chip to see if the results from the statistical analysis increase in significance.

Another way to look at the same information is to look at the first two principal components. Figure 4 shows a principal component analysis of the individual chips in order to determine any structure in the relationship between chips. The PCA is based on all genes.

3.3 Classification of chips

A K nearest neighbor (KNN) classifier was built to classify chips based on the expression of all genes. Each chip was compared to all other chips and the category assignment of the three closest chips ($k=3$) in Euclidean gene expression space was used to predict its category. Table 1 shows the prediction for each chip. The

total accuracy of class prediction reached was 100 and 100 percent for a k=1 and a k=3 classifier, respectively. It may be possible to improve on this accuracy by selecting predictive genes and by optimizing the number of nearest neighbors K. Doing this, however, will necessitate an evaluation on an independent test set that was not used for optimizing the classifier.

A Nearest Centroid (NC) Classifier was built as well. Instead of the closest chip in Euclidean space, the closest class centroid was used to predict the class of each chip. The total accuracy of class prediction reached was 100 percent. Also here, the performance may be improved by using a selection of genes.

Table 1: Predictions of the KNN and NC Classifiers

Chip	Category assigned in input	Prediction K=1	Prediction K=3	Prediction NC
Ctrl1	A	A	A	A
Ctrl2	A	A	A	A
Ctrl3	A	A	A	A
HIV1	B	B	B	B
HIV2	B	B	B	B
HIV3	B	B	B	B

3.4 Statistical Analysis

The cutoff in P -values used was 0.000946. 100 genes had P -values below that cutoff and are presented in Table 2 and Table 3 below. At that cutoff, we expect 7 false positive genes (0.000946×7129 genes on the chip). That means that we have a false discovery rate of 0.07 in Table 2 and Table 3 (7/100). We have, however, no way of knowing which genes are false positive unless we verify the findings with an independent method.

The genes are divided into upregulated genes (Table 2) and downregulated genes (Table 3) and ranked according to P -value. The most significant gene (rank=1) is ranked at the top, the least significant gene is ranked at the bottom. For each gene there is a list of gene ontology annotations (GO), if available. Information on the P -values and expression levels of *all* genes on the array is available in the file all.annotated.genes in the same directory as this report.

In the Adobe Acrobat (PDF) version of this report, the probe ID is hyperlinked to the LocusLink database (if available). Clicking on the probe ID will take you to a detailed description of the gene in that database.

Table 2: The top ranking upregulated genes in statistical analysis. Numbers in parenthesis help evaluate the significance and relevance of the result: expression level of gene on the first chip, P value from the statistical analysis, and the average fold change between the last and the first category. Example: A fold change of 2.5 means 2.5-fold upregulated in the last category relative to the first category.

Rank	Gene	Annotations (expressionlevel Pvalue foldchange)
8	HG3344-H	ubiquitin-conjugating enzyme E2D 1 (UBC4/5 homolog, yeast). GO: ubiquitin-protein ligase ; ubiquitin conjugating enzyme ; ubiquitin-dependent protein degradation ; (1308 5.1e-05 1.2)
9	Z29074_a	keratin 9 (epidermolytic palmoplantar keratoderma). GO: regulation of cell shape ; intermediate filament ; epidermal differentiation ; structural constituent of cytoskeleton ; (1942 5.9e-05 2.2)
13	U62317_r	arylsulfatase A. GO: lysosome ; arylsulfatase ; (2491 7.4e-05 1.4)
20	M34079_a	proteasome (prosome, macropain) 26S subunit, ATPase, 3. GO: nucleus ; 26S proteasome ; adenosinetriphosphatase ; transcription co-activator ; transcription co-repressor ; (1886 1.2e-04 1.7)
21	HG64-HT6	human immunodeficiency virus type I enhancer binding protein 3. (777 1.3e-04 1.1)
22	M14218_a	argininosuccinate lyase. GO: cytoplasm ; urea cycle ; arginine catabolism ; argininosuccinate lyase ; (2575 1.3e-04 1.6)
23	U66048_a	NA. (2082 1.3e-04 1.1)
27	X71428_a	fusion, derived from t(12. GO: nucleus ; RNA binding ; (3858 2.0e-04 2.4)
28	M19684_a	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 2. GO: serine protease inhibitor ; (3900 2.0e-04 1.5)
32	Z80783_a	H2B histone family, member L. (844 2.5e-04 1.5)
33	M16967_a	coagulation factor V (proaccelerin, labile factor). GO: blood coagulation ; blood coagulation factor ; (1249 2.6e-04 1.9)
37	M21142_c	GNAS complex locus. GO: olfaction ; plasma membrane ; Golgi trans cisterna ; adenylate cyclase activation ; Golgi to secretory vesicle transport ; heterotrimeric G-protein GTPase, alpha-subunit ; (18675 3.0e-04 1.2)
38	L33799_a	procollagen C-endopeptidase enhancer. GO: collagen binding ; development ; cell growth and/or maintenance ; (2744 3.0e-04 1.4)
42	D38073_a	MCM3 minichromosome maintenance deficient 3 (S. cerevisiae). GO: DNA binding ; adenosinetriphosphatase ; DNA replication initiation ; alpha DNA polymerase:primase complex ; (2881 3.4e-04 1.7)
44	X98507_a	myosin IC. GO: motor ; myosin ATPase ; actin cytoskeleton ; (772 3.5e-04 1.5)
45	M94630_a	heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kD). GO: nucleus ; RNA binding ; RNA catabolism ; RNA processing ; (4589 3.5e-04 1.9)

48 X89398_c uracil-DNA glycosylase. GO: DNA repair ; base-excision repair ; uracil DNA N-glycosylase ; (1690 3.6e-04 1.7)

49 L07594_a transforming growth factor, beta receptor III (betaglycan, 300kD). GO: receptor ; signal transduction ; development ; integral membrane protein ; glycosaminoglycan binding ; TGFbeta receptor signaling pathway ; (520 3.7e-04 1.4)

54 L11708_a hydroxysteroid (17-beta) dehydrogenase 2. GO: estrogen biosynthesis ; endoplasmic reticulum membrane ; (772 4.1e-04 1.2)

56 AB002382 catenin (cadherin-associated protein), delta 1. GO: cell adhesion ; (1807 4.2e-04 1.5)

59 M64497_a nuclear receptor subfamily 2, group F, member 2. GO: nucleus ; lipid metabolism ; signal transduction ; transcription co-repressor ; ligand-dependent nuclear receptor ; ligand-regulated transcription factor ; regulation of transcription from Pol II promoter ; (1152 4.4e-04 1.4)

60 D32002_s nuclear cap binding protein subunit 1, 80kD. GO: nucleoplasm ; RNA binding ; mRNA splicing ; binding to mRNA cap ; mRNA-nucleus export ; (956 4.5e-04 1.1)

67 L15388_a G protein-coupled receptor kinase 5. GO: cytoplasm ; soluble fraction ; phospholipid binding ; protein kinase C binding ; G-protein-coupled receptor phosphorylating protein kinase ; G-protein signaling, coupled to cAMP nucleotide second messenger ; regulation of G-protein coupled receptor protein signaling pathway ; (3078 5.1e-04 1.2)

69 U88898_a unnamed HERV-H protein. (249 5.3e-04 1.6)

73 K03192_f cytochrome P450, subfamily IIA (phenobarbital-inducible), polypeptide 6. GO: microsome ; monooxygenase ; cytochrome P450 ; coumarin 7-hydroxylase ; (1574 5.6e-04 2.3)

74 Z80776_a H2A histone family, member G. (466 5.7e-04 1.4)

80 M15205_a thymidine kinase 1, soluble. GO: cytoplasm ; thymidine kinase ; nucleobase, nucleoside, nucleotide and nucleic acid metabolism ; (3763 6.6e-04 1.6)

87 L07540_a replication factor C (activator 1) 5 (36.5kD). (1064 7.5e-04 1.4)

89 X13293_a v-myb myeloblastosis viral oncogene homolog (avian)-like 2. GO: chromatin ; anti-apoptosis ; regulation of cell cycle ; transcription factor ; development ; transcription from Pol II promoter ; (1871 8.3e-04 2.0)

90 U21128_a lumican. GO: vision ; proteoglycan ; extracellular matrix ; cartilage condensation ; extracellular matrix glycoprotein ; (122 8.4e-04 1.6)

92 S94421_a NA. (2793 8.6e-04 1.2)

94 M15465_s pyruvate kinase, liver and RBC. GO: pyruvate kinase ; (2934 8.7e-04 1.5)

95 M63589_a T-cell acute lymphocytic leukemia 1. GO: oncogenesis ; DNA binding ; cell proliferation ; (1189 8.8e-04 1.3)

96 HG3921-H homeo box A6. (2298 9.0e-04 3.2)

97 HG1102-H ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1). GO: GTPase ; cell adhesion ; cell motility ; response to wounding ; inflammatory response ; embryogenesis and morphogenesis ; intracellular signaling cascade ; (1815 9.0e-04 1.4)

98	L05187_a	small proline-rich protein 1A. GO: structural molecule ; epidermal differentiation ; (761 9.2e-04 1.3)
100	U70867_a	solute carrier family 21 (prostaglandin transporter), member 2. GO: lipid transport ; membrane fraction ; lipid transporter ; integral plasma membrane protein ; (1425 9.5e-04 1.2)

Table 3: The top ranking downregulated genes in statistical analysis. Numbers in parenthesis help evaluate the significance and relevance of the result: expression level of gene on the first chip, P value from the statistical analysis, and the average fold change between the last and the first category. Example: A fold change of -2.5 means 2.5-fold downregulated in the last category relative to the first category.

Rank	Gene	Annotations (expressionlevel Pvalue foldchange)
1	HG1872-H	CD74 antigen (invariant polypeptide of major histocompatibility complex, class II antigen-associated). GO: integral membrane protein ; class II major histocompatibility complex antigen ; (23965 3.4e-06 -3.5)
2	X63717_a	tumor necrosis factor receptor superfamily, member 6. GO: receptor ; apoptosis ; anti-apoptosis ; soluble fraction ; signal transduction ; signal transducer ; induction of apoptosis ; transmembrane receptor ; protein complex assembly ; integral plasma membrane protein ; integral plasma membrane proteoglycan ; (2325 6.8e-06 -1.8)
3	HG3576-H	major histocompatibility complex, class II, DR beta 5. GO: integral plasma membrane protein ; perception of pest/pathogen/parasite ; class II major histocompatibility complex antigen ; (20466 1.7e-05 -2.2)
4	D50925_a	PAS domain containing serine/threonine kinase. (1445 2.6e-05 -1.3)
5	D16227_a	hippocalcin-like 1. GO: calcium ion binding ; (3333 3.8e-05 -1.7)
6	L13744_a	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila). GO: nucleus ; oncogenesis ; (204 4.7e-05 -1.4)
7	U89336_c	chromosome 6 open reading frame 9. (4443 4.7e-05 -2.5)
10	L06797_s	chemokine (C-X-C motif), receptor 4 (fusin). GO: apoptosis ; virulence ; cytoplasm ; chemotaxis ; coreceptor ; neurogenesis ; pathogenesis ; immune response ; invasive growth ; plasma membrane ; activation of MAPK ; chemokine receptor ; response to viruses ; inflammatory response ; G-protein coupled receptor ; histogenesis and organogenesis ; integral plasma membrane protein ; cytosolic calcium ion concentration elevation ; G-protein coupled receptor protein signaling pathway ; (3684 6.5e-05 -1.8)
11	U03105_a	proline rich 2. GO: nucleus ; protein binding ; (6322 6.9e-05 -4.0)
12	M92843_s	zinc finger protein 36, C3H type, homolog (mouse). GO: cytoplasm ; mRNA catabolism ; single-stranded RNA binding ; (5341 7.0e-05 -2.6)

- 14 M63904_a guanine nucleotide binding protein (G protein), alpha 15 (Gq class). GO: plasma membrane ; phospholipase C activation ; heterotrimeric G-protein GTPase, alpha-subunit ; muscarinic acetyl choline receptor, phospholipase C activating pathway ; (7958 8.0e-05 -2.8)
- 15 M14219_a decorin. GO: extracellular matrix ; histogenesis and organogenesis ; chondroitin sulfate/dermatan sulfate proteoglycan ; (1649 8.1e-05 -1.7)
- 16 U50527_s hypothetical gene CG018. (404 8.9e-05 -3.1)
- 17 M32011_a neutrophil cytosolic factor 2 (65kD, chronic granulomatous disease, autosomal 2). GO: cytosol ; soluble fraction ; electron transporter ; superoxide metabolism ; cellular defense response ; (3500 9.3e-05 -3.6)
- 18 M80563_a S100 calcium binding protein A4 (calcium protein, calvasculin, metastasin, murine placental homolog). GO: calcium ion binding ; invasive growth ; (18426 9.6e-05 -2.7)
- 19 X79067_a zinc finger protein 36, C3H type-like 1. GO: nucleus ; transcription factor ; (4318 1.2e-04 -2.1)
- 24 L35249_s ATPase, H⁺ transporting, lysosomal 56/58kD, V1 subunit B, isoform 2. GO: proton transport ; hydrogen ion transporter ; vacuolar hydrogen-transporting ATPase ; (1069 1.4e-04 -2.1)
- 25 L46720_s ectonucleotide pyrophosphatase/phosphodiesterase 2 (autotaxin). GO: chemotaxis ; cell motility ; plasma membrane ; phosphodiesterase I ; phosphate metabolism ; nucleotide pyrophosphatase ; transcription factor binding ; integral plasma membrane protein ; G-protein coupled receptor protein signaling pathway ; (2481 1.8e-04 -1.8)
- 26 Y00062_a protein tyrosine phosphatase, receptor type, C. GO: protein tyrosine phosphatase ; integral plasma membrane protein ; cell surface receptor linked signal transduction ; transmembrane receptor protein tyrosine phosphatase ; (1945 1.8e-04 -2.6)
- 29 X55666_a upstream transcription factor 1. GO: nucleus ; transcription from Pol II promoter ; specific RNA polymerase II transcription factor ; (4783 2.0e-04 -1.2)
- 30 U44754_a small nuclear RNA activating complex, polypeptide 1, 43kD. GO: transcription from Pol II promoter ; transcription from Pol III promoter ; (402 2.1e-04 -2.0)
- 31 M59465_a tumor necrosis factor, alpha-induced protein 3. (6581 2.4e-04 -2.5)
- 34 U37518_a tumor necrosis factor (ligand) superfamily, member 10. GO: receptor binding ; soluble fraction ; signal transduction ; cell-cell signaling ; induction of apoptosis ; integral plasma membrane protein ; (749 2.7e-04 -2.8)
- 35 U79256_a hypothetical protein MGC14258. (2573 2.8e-04 -2.0)
- 36 M33600_f major histocompatibility complex, class II, DR beta 1. GO: pathogenesis ; class II major histocompatibility complex antigen ; (28308 2.9e-04 -2.2)
- 39 U15085_a major histocompatibility complex, class II, DM beta. GO: chaperone ; immune response ; MHC-interacting protein ; perception of pest/pathogen/parasite ; (7320 3.2e-04 -2.2)
- 40 U68494_a NA. (1587 3.3e-04 -1.6)

- 41 AC002477 NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1 (7.5kD, MWFE). GO: energy pathways ; membrane fraction ; NADH dehydrogenase complex (ubiquinone) (sensu Eukarya) ; NADH dehydrogenase (ubiquinone) ; (1412 3.4e-04 -1.3)
- 43 U66838_a cyclin A1. GO: cytosol ; male meiosis I ; spermatogenesis ; regulation of cell cycle ; regulation of CDK activity ; (1055 3.4e-04 -1.3)
- 46 U77735_a pim-2 oncogene. GO: male meiosis ; cell proliferation ; protein amino acid phosphorylation ; protein serine/threonine kinase ; (6745 3.6e-04 -1.7)
- 47 S73591_a thioredoxin interacting protein. (3054 3.6e-04 -1.9)
- 50 M57466_s major histocompatibility complex, class II, DP beta 1. GO: pathogenesis ; perception of pest/pathogen/parasite ; class II major histocompatibility complex antigen ; (14496 3.8e-04 -2.2)
- 51 M27533_s CD80 antigen (CD28 antigen ligand 1, B7-1 antigen). GO: receptor binding ; immune response ; plasma membrane ; signal transduction ; (2640 4.0e-04 -1.9)
- 52 HG3484-H CDC-like kinase 1. GO: regulation of cell cycle ; cell proliferation ; protein serine/threonine kinase ; non-membrane spanning protein tyrosine kinase ; (1583 4.0e-04 -3.1)
- 53 U60975_a sortilin-related receptor, L(DLR class) A repeats-containing. GO: transmembrane receptor ; internalization receptor ; receptor mediated endocytosis ; integral plasma membrane protein ; (4633 4.0e-04 -2.3)
- 55 X61123_a B-cell translocation gene 1, anti-proliferative. GO: cell proliferation ; cell cycle regulator ; negative regulation of cell proliferation ; (3864 4.1e-04 -2.1)
- 57 U20734_s jun B proto-oncogene. GO: chromatin ; DNA binding ; transcription co-activator ; transcription co-repressor ; RNA polymerase II transcription factor ; regulation of transcription from Pol II promoter ; (5585 4.2e-04 -2.3)
- 58 K01383_a metallothionein 1A (functional). GO: heavy metal binding ; response to heavy metal ; heavy metal sensitivity/resistance ; heavy metal ion transport ; (6085 4.4e-04 -2.0)
- 61 U21551_a branched chain aminotransferase 1, cytosolic. GO: cytosol ; cell proliferation ; G1/S transition of mitotic cell cycle ; branched-chain amino acid aminotransferase ; branched chain family amino acid biosynthesis ; (588 4.5e-04 -2.0)
- 62 M25322_a selectin P (granule membrane protein 140kD, antigen CD62). GO: selectin ; cell adhesion molecule ; plasma membrane ; soluble fraction ; secretory vesicle ; cell adhesion receptor ; integral plasma membrane protein ; integral plasma membrane proteoglycan ; (560 4.6e-04 -1.0)
- 63 L07956_a glucan (1,4-alpha-), branching enzyme 1 (glycogen branching enzyme, Andersen disease, glycogen storage disease type IV). GO: energy pathways ; glycogen metabolism ; 1,4-alpha-glucan branching enzyme ; (1947 4.6e-04 -3.1)
- 64 HG688-HT major histocompatibility complex, class II, DR beta 1. GO: pathogenesis ; class II major histocompatibility complex antigen ; (18910 4.7e-04 -2.2)
- 65 K02765_a complement component 3. GO: receptor binding ; immune response ; signal transduction ; G-protein coupled receptor protein signaling pathway ; (2852 5.0e-04 -1.6)

- 66 Z29066_s NIMA (never in mitosis gene a)-related kinase 2. GO: mitosis ; centrosome ; regulation of cell cycle ; regulation of mitosis ; protein serine/threonine kinase ; (1564 5.0e-04 -3.7)
- 68 U58091_a cullin 4B. (308 5.2e-04 -2.2)
- 70 X71661_a lectin, mannose-binding, 1. GO: chaperone ; Golgi membrane ; protein folding ; blood coagulation ; ER to Golgi transport ; mannose binding lectin ; integral membrane protein ; endoplasmic reticulum membrane ; (522 5.3e-04 -1.5)
- 71 D31767_a DAZ associated protein 2. (7814 5.3e-04 -1.7)
- 72 U38545_a phospholipase D1, phosphatidylcholine-specific. GO: membrane ; chemotaxis ; phospholipase D ; phospholipid metabolism ; RAS protein signal transduction ; small GTPase mediated signal transduction ; (1678 5.5e-04 -1.9)
- 75 X53587_a integrin, beta 4. GO: integrin ; oncogenesis ; cell adhesion ; invasive growth ; cell adhesion receptor ; (4144 5.7e-04 -3.0)
- 76 L38487_a estrogen-related receptor alpha. GO: nucleus ; DNA binding ; ligand-dependent nuclear receptor ; (6398 5.8e-04 -1.2)
- 77 U90551_a H2A histone family, member L. (482 6.3e-04 -2.7)
- 78 M58459_a ribosomal protein S4, Y-linked. GO: RNA binding ; protein biosynthesis ; structural constituent of ribosome ; cytosolic small ribosomal subunit (sensu Eukarya) ; (12961 6.4e-04 -2.6)
- 79 M65217_a heat shock transcription factor 2. GO: response to heat shock ; transcription factor ; transcription co-activator ; transcription from Pol II promoter ; (1096 6.5e-04 -2.0)
- 81 U30521_a P311 protein. (1476 6.8e-04 -2.7)
- 82 X77366_a nuclear factor (erythroid-derived 2)-like 1. GO: nucleus ; heme biosynthesis ; transcription factor ; inflammatory response ; transcription cofactor ; embryogenesis and morphogenesis ; transcription from Pol II promoter ; (3437 6.9e-04 -1.8)
- 83 L40379_a thyroid hormone receptor interactor 10. GO: protein binding ; signal transduction ; actin cytoskeleton reorganization ; (6021 7.0e-04 -3.8)
- 84 M37721_a peptidylglycine alpha-amidating monooxygenase. GO: soluble fraction ; protein modification ; electron transporter ; peptidyl-glycine monooxygenase ; integral plasma membrane protein ; (1640 7.0e-04 -1.9)
- 85 X03100_c major histocompatibility complex, class II, DP alpha 1. (15529 7.1e-04 -2.0)
- 86 X69111_a inhibitor of DNA binding 3, dominant negative helix-loop-helix protein. GO: development ; transcription co-repressor ; (2854 7.2e-04 -1.3)
- 88 M20681_a solute carrier family 2 (facilitated glucose transporter), member 3. GO: glucose transport ; membrane fraction ; glucose transporter ; carbohydrate metabolism ; integral membrane protein ; (12317 8.2e-04 -2.2)
- 91 U39317_a ubiquitin-conjugating enzyme E2D 2 (UBC4/5 homolog, yeast). GO: oncogenesis ; invasive growth ; protein modification ; ubiquitin-protein ligase ; ubiquitin conjugating enzyme ; ubiquitin-dependent protein degradation ; (2572 8.5e-04 -1.7)
- 93 X51408_a chimerin (chimaerin) 1. GO: GTPase activator ; SH3/SH2 adaptor protein ; (7540 8.6e-04 -2.3)

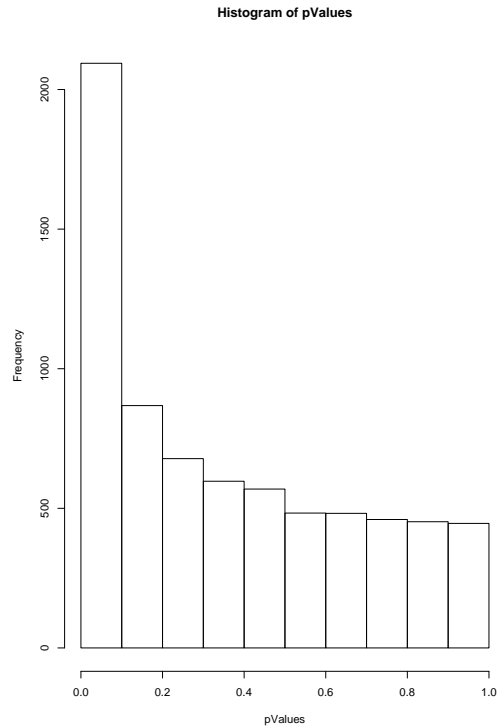


Figure 5: A histogram of all P-values. A uniform distribution of P-values over the interval 0 to 1 is indicative of few or none differentially expressed genes. A peak at the low end of the distribution is indicative of differential expression of many genes.

3.5 Functional categories

The top ranking genes that have a function annotated by Gene Ontology terms have been placed into functional and process categories as defined by the Gene Ontology Consortium. Figure 7 shows the distribution of the upregulated and downregulated genes by function. Upregulation and downregulation is determined based on the last category compared to the first category. Figure 8 compares upregulated and downregulated genes directly by category.

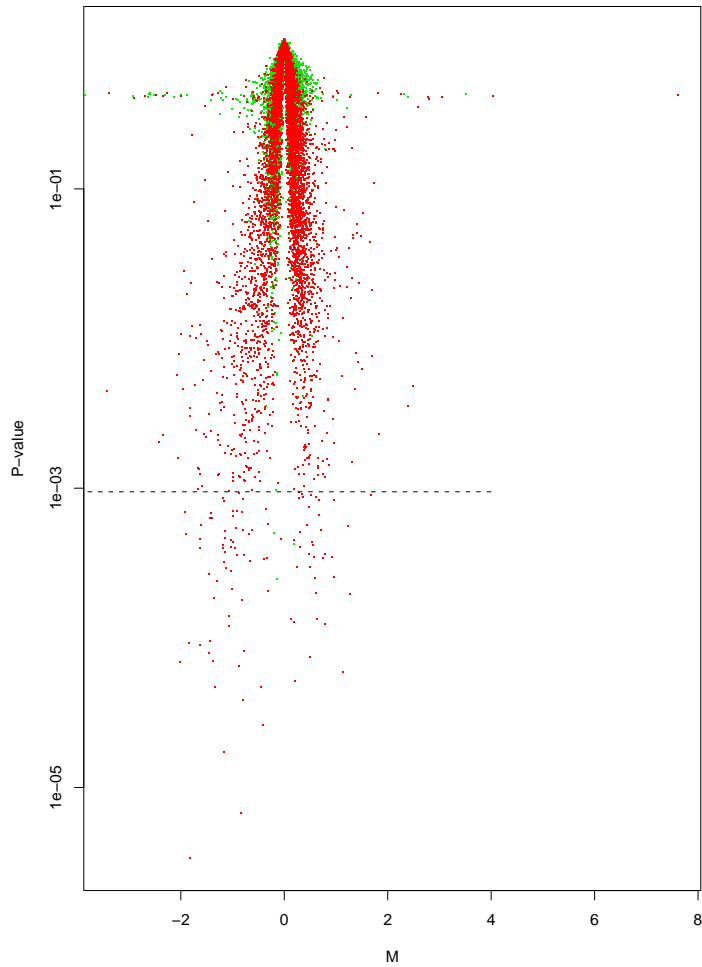
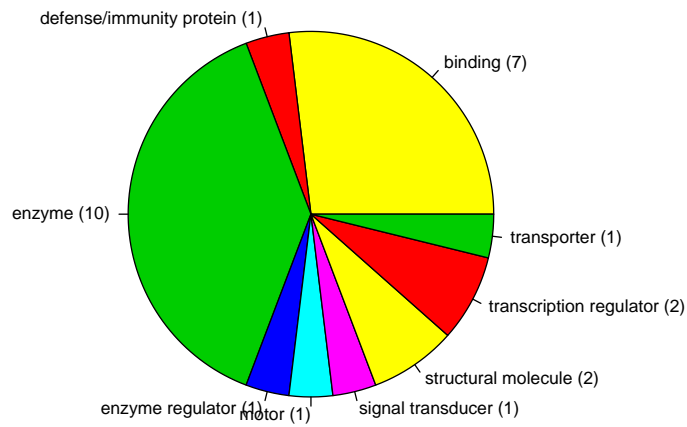


Figure 6: A "volcano" plot (Wolfinger, R.D. et. al. (2001) J. Comp. Biol. 8:625-638) showing the relationship between P -value and \log_2 fold change (M). The relationship is shown both for the original data (red) and for a permutation of the columns (green). The permutation (shuffling of the data) should remove the signal and leave only the noise, allowing an estimate of the P -values and fold changes that can occur by chance alone. The chosen P -value cutoff of 0.000946 is shown by a dotted line. Note that to save time only one permutation is performed. Ideally all possible permutations should be tried.

Functional Categories of upregulated genes



Functional Categories of downregulated genes

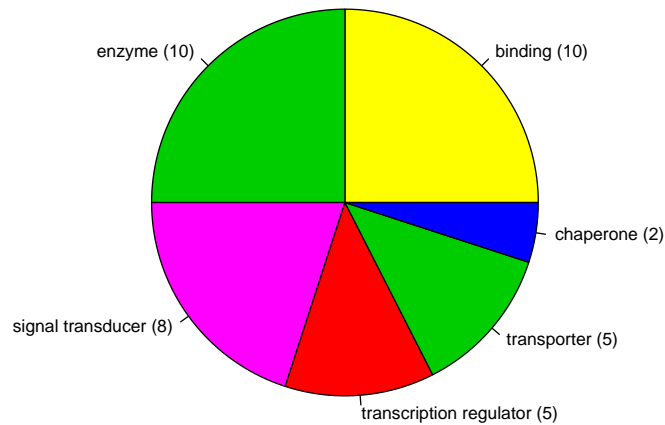


Figure 7: Gene ontology function categories of those top ranking genes that have been annotated. The number of genes in each category is shown in parenthesis. Note that only a fraction of the top ranking genes have been categorized with a gene ontology function.

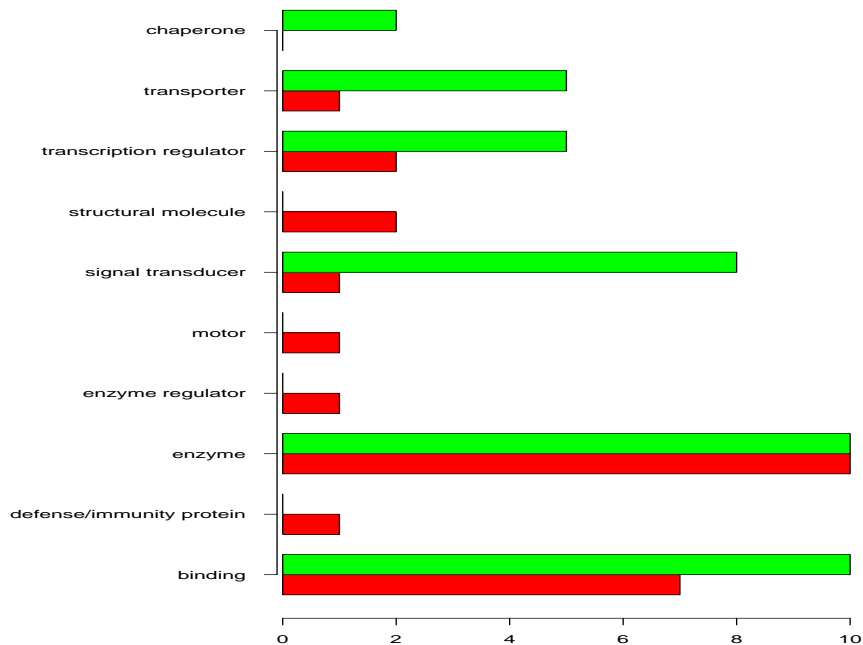


Figure 8: Gene ontology function categories of those top ranking genes that have been annotated. Upregulated genes are shown in red, downregulated genes are shown in green.

3.6 Prediction of orphan function

Among the top ranking genes are genes with unknown function. For those genes where the complete amino acid sequence is known or predicted, the ProtFun software was used to predict the function in general categories (Table 4).

Table 4: ProtFun prediction of orphan gene function, if any.

Gene	ProtFun Predicted Categories
U89336_cds1_at	Cell_envelope; Nonenzyme; Growth_factor;
AB002382_at	Cell_envelope; Enzyme; Ligase; Ion_channel;
D31767_at	Cell_envelope; Enzyme; Cation_channel;

3.7 Signal transduction pathway analysis

The top genes were searched against the TRANSPATH¹⁴ signal transduction database (www.transpath.de or www.gene-regulation.com). Table 5 shows the results.

Table 5: Table of top ranking genes found in TRANSPATH. Expression refers to absolute expression of the gene on the first chip, P-value of differential expression and logfold change in expression. Pathway refers to the name of the pathway in TRANSPATH in which the gene was found and the gene name refers to the name used for the gene in that pathway. If you click on a gene identifier, your browser will take you to a database description of it.

Gene	Expression	Gene name in pathway	Pathway	Figure
X63717_a	(2325 6.8e-06 -1.8)	Fas	cancernet	13
M63904_a	(7958 8.0e-05 -2.8)	G-alpha-16	IL-8	10
M59465_a	(6581 2.4e-04 -2.5)	A20	TNF_alpha	12
L07594_a	(520 3.7e-04 1.4)	TGFR-III	TGFbetamap	11
M27533_s	(2640 4.0e-04 -1.9)	CD80	CD28	9

The figures shown on the following pages give a schematic overview of the signal transduction pathways in which differentially expressed genes were found. Remember that the signal is usually transmitted by protein-to-protein contact. Such protein-to-protein contact is not detected in a DNA microarray experiment. What is detected instead is if any genes encoding the proteins in the pathway are regulated or if any target genes of the pathways are regulated.

The signal transduction pathway analysis was extended beyond the top ranking genes to look for all genes in the experiment which could be mapped to a TRANSPATH annotated pathway. The purpose of this is to discover pathways with a number of differentially regulated genes, even though they on an individual gene basis do not pass a statistical significance test.

Figure 14 shows all the TRANSPATH pathways in which genes were found and summarizes their rank in the statistical analysis.

¹⁴Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E. "TRANSPATH: an integrated database on signal transduction and a tool for array analysis." Nucleic Acids Res. 2003 Jan 1;31(1):97-100.

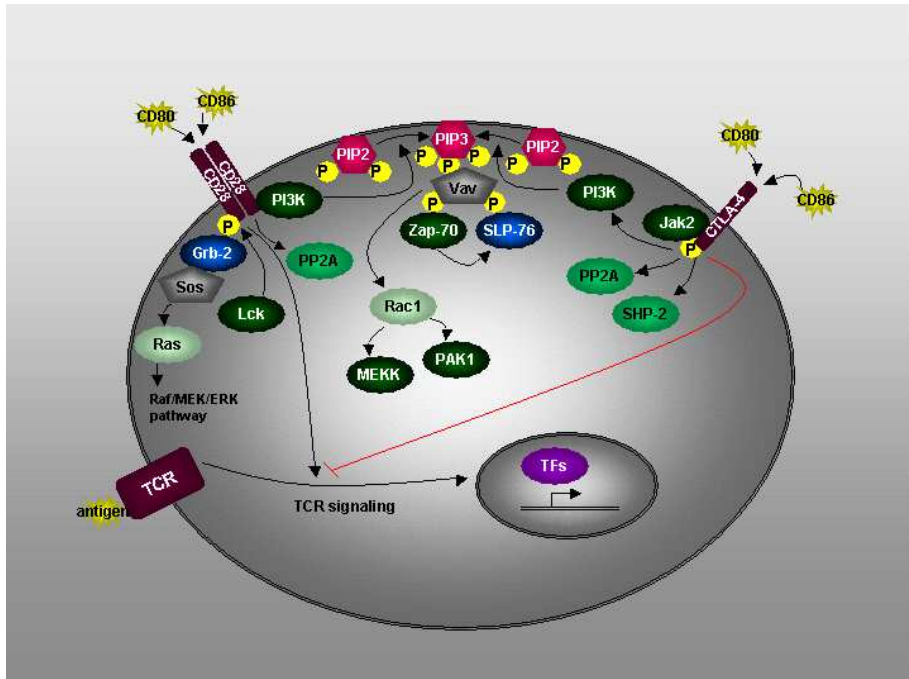


Figure 9: The CD28 signal transduction pathway.

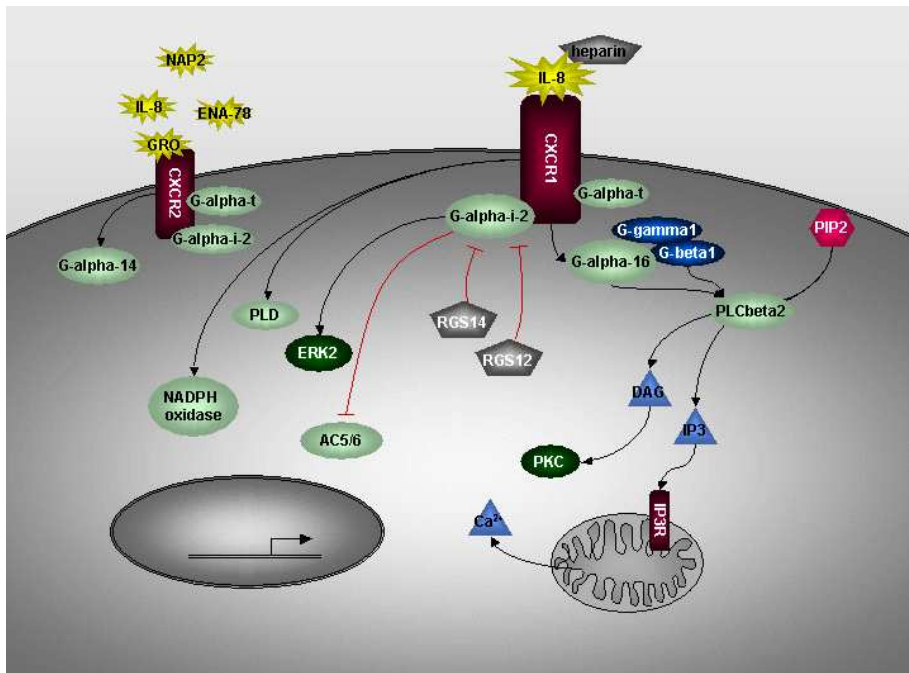


Figure 10: The IL-8 signal transduction pathway.

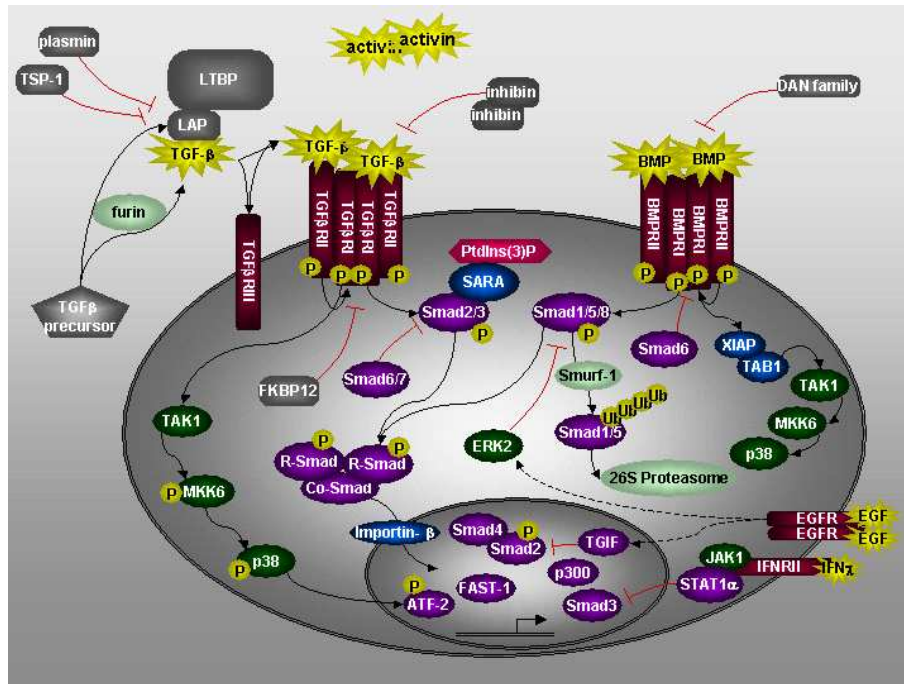


Figure 11: The TGFbetamap signal transduction pathway.

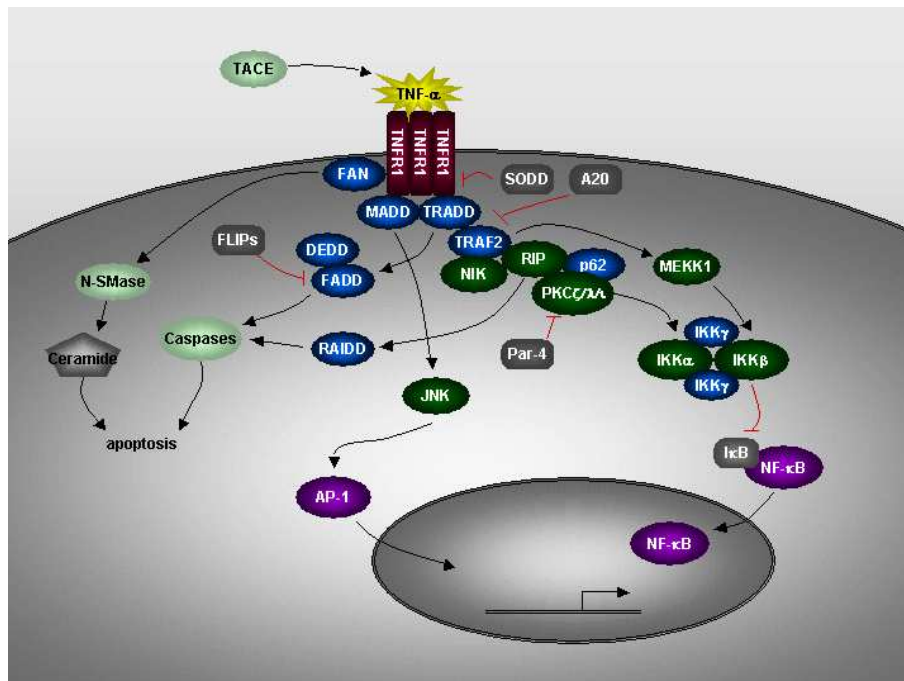


Figure 12: The TNF_alpha signal transduction pathway.

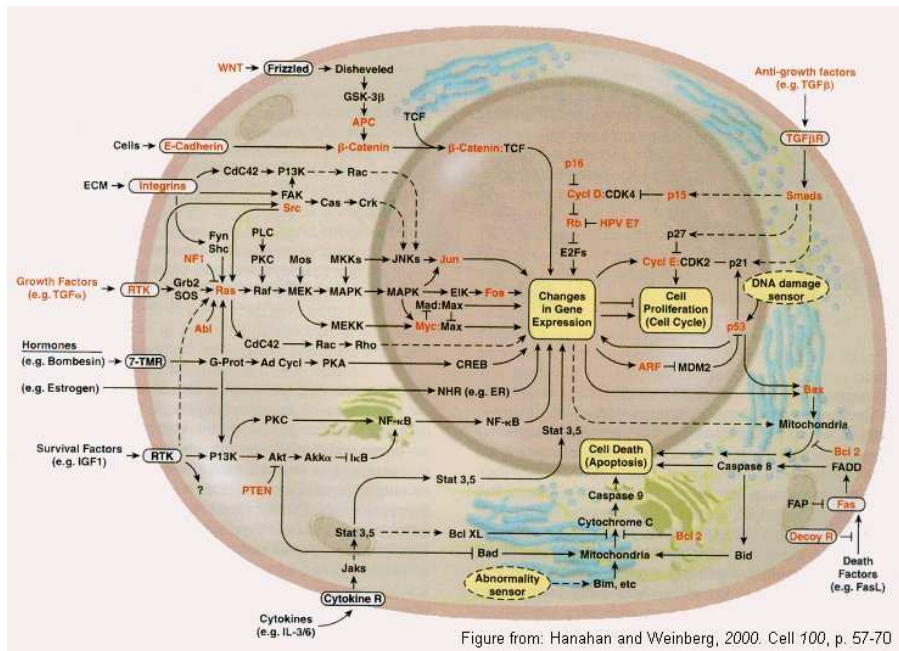


Figure 13: The cancer-related signal transduction pathway.

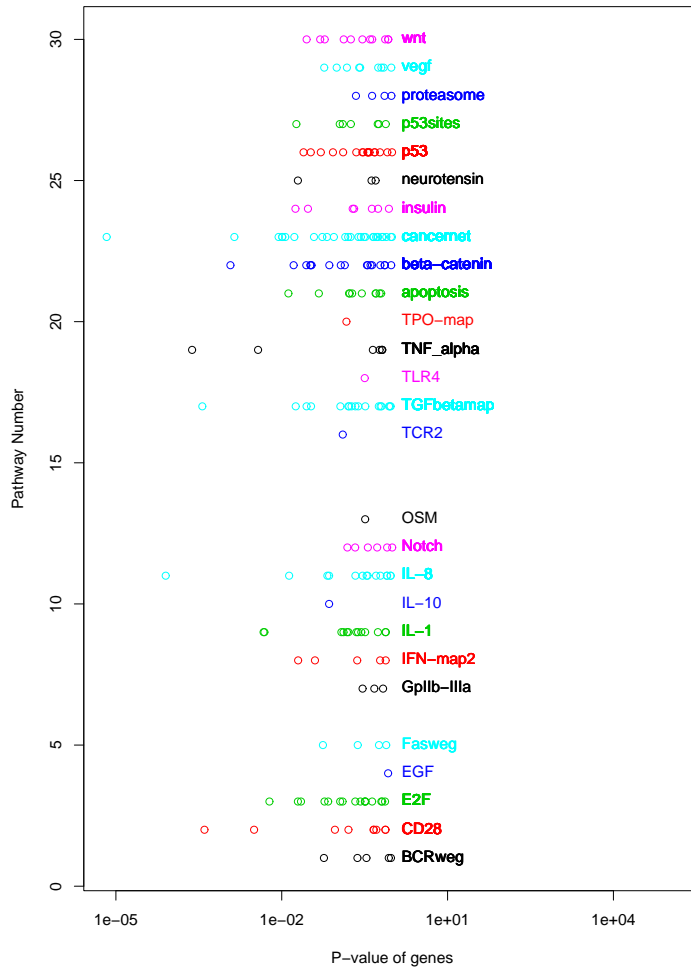


Figure 14: A list of all signal transduction pathways in which genes were found. The x-axis shows the P-value of each gene assigned to each pathway. A P-value close to 1 means the gene is almost certain to be unchanged in the experiment. The smaller the P-value, the greater the probability of differential regulation. Pathways with differential expression should stand out from the background level.

3.8 Metabolic pathway analysis

A pathway analysis was performed on the top ranking genes by running them against the KEGG database of cellular pathways. Table 6 shows the results.

Table 6: Table of top ranking genes found in KEGG. The pathway of the top gene can be seen in Figure 15 and the E.C. number refers to the step in that pathway. If you click on a pathway name, your browser will take you to a figure of the pathway. You can locate the E.C. numbers on the figures. If you click on a gene identifier, your browser will take you to a database description of it.

Gene	Description	Pathway
U62317_r	aslA; arylsulfatase [EC:3.1.6.1] (2491 7.4e-05 1.4)	Sphingoglycolipid metabolism
M14218_a	argininosuccinate lyase [EC:4.3.2.1] (2575 1.3e-04 1.6)	Arginine and proline metabolism
L46720_s	nucleotide pyrophosphatase [EC:3.6.1.9] (2481 1.8e-04 -1.8)	Pantothenate and CoA biosynthesis
AC002477	NADH dehydrogenase [EC:1.6.5.3] (1412 3.4e-04 -1.3)	Oxidative phosphorylation
U21551_a	branched-chain amino acid aminotransferase [EC:2.6.1.42] (588 4.5e-04 -2.0)	Pantothenate and CoA biosynthesis
L07956_a	1,4-alpha-glucan branching enzyme [EC:2.4.1.18] (1947 4.6e-04 -3.1)	Starch and sucrose metabolism
U38545_a	phospholipase D [EC:3.1.4.4] (1678 5.5e-04 -1.9)	Phospholipid degradation
M15205_a	thymidine kinase [EC:2.7.1.21] (3763 6.6e-04 1.6)	Pyrimidine metabolism
M15465_s	pyk; pyruvate kinase [EC:2.7.1.40] (2934 8.7e-04 1.5)	Carbon fixation

The KEGG pathway analysis was extended beyond the top ranking genes to look for all genes in the experiment which could be mapped to a KEGG pathway. The purpose of this is to discover pathways with a number of differentially regulated genes, even though they on an individual gene basis do not pass a statistical significance test.

Figure 16 shows all the KEGG pathways in which genes were found and summarizes their rank in the statistical analysis.

3.9 Clustering of Genes

A visualization of the expression of the top ranking genes in each of the experiments is performed by clustering with the ClusterExpress software (Figure 17).

A number of K-means clusterings were performed as well. First the number of clusters, K, was optimized by measuring how the number of clusters affects the quality of the clustering (Figure 18). Then a K-means clustering using the optimal number of clusters, 2, was performed (Figure 19).

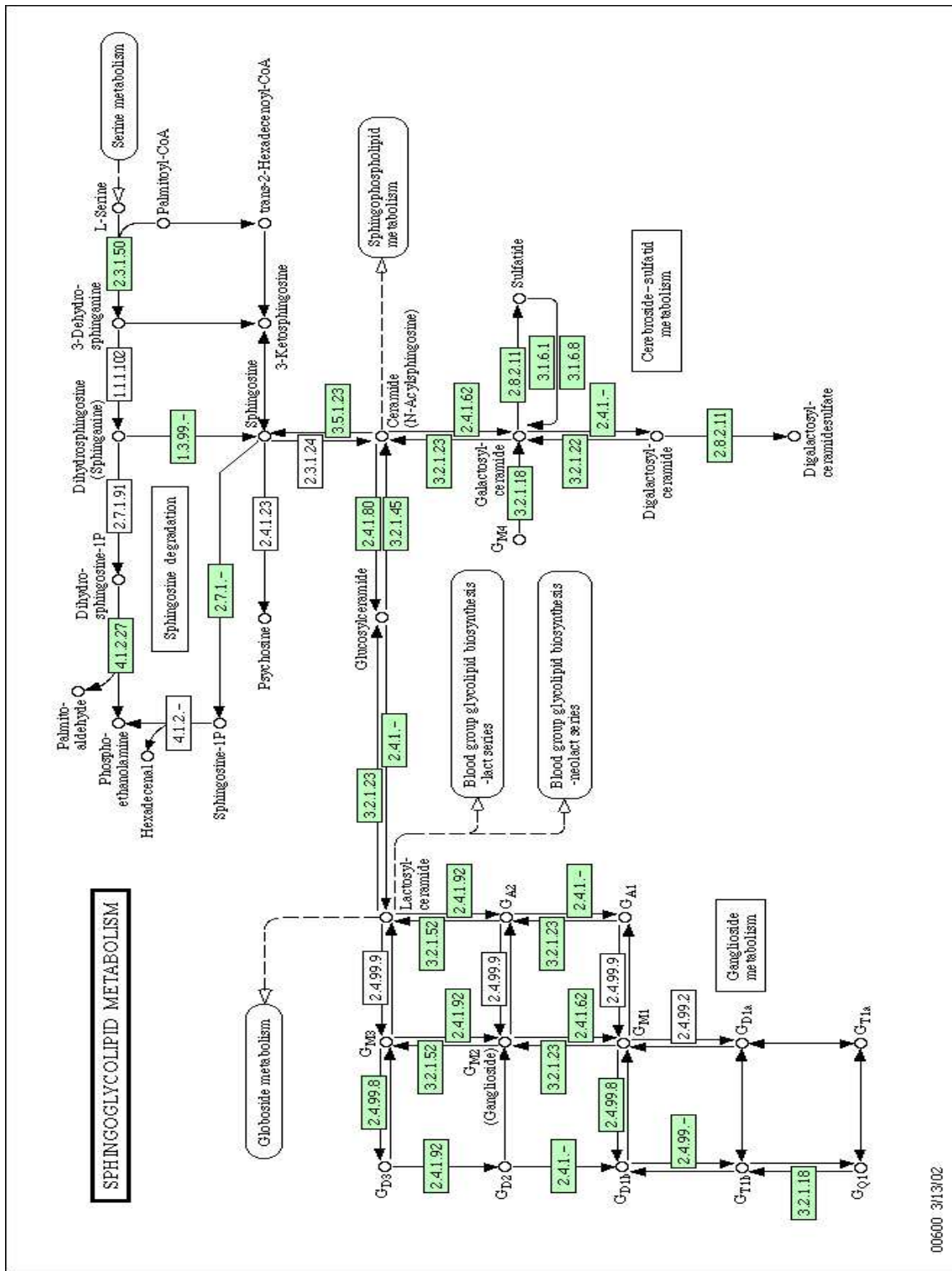


Figure 15: The KEGG pathway of the highest ranking gene from Table 6

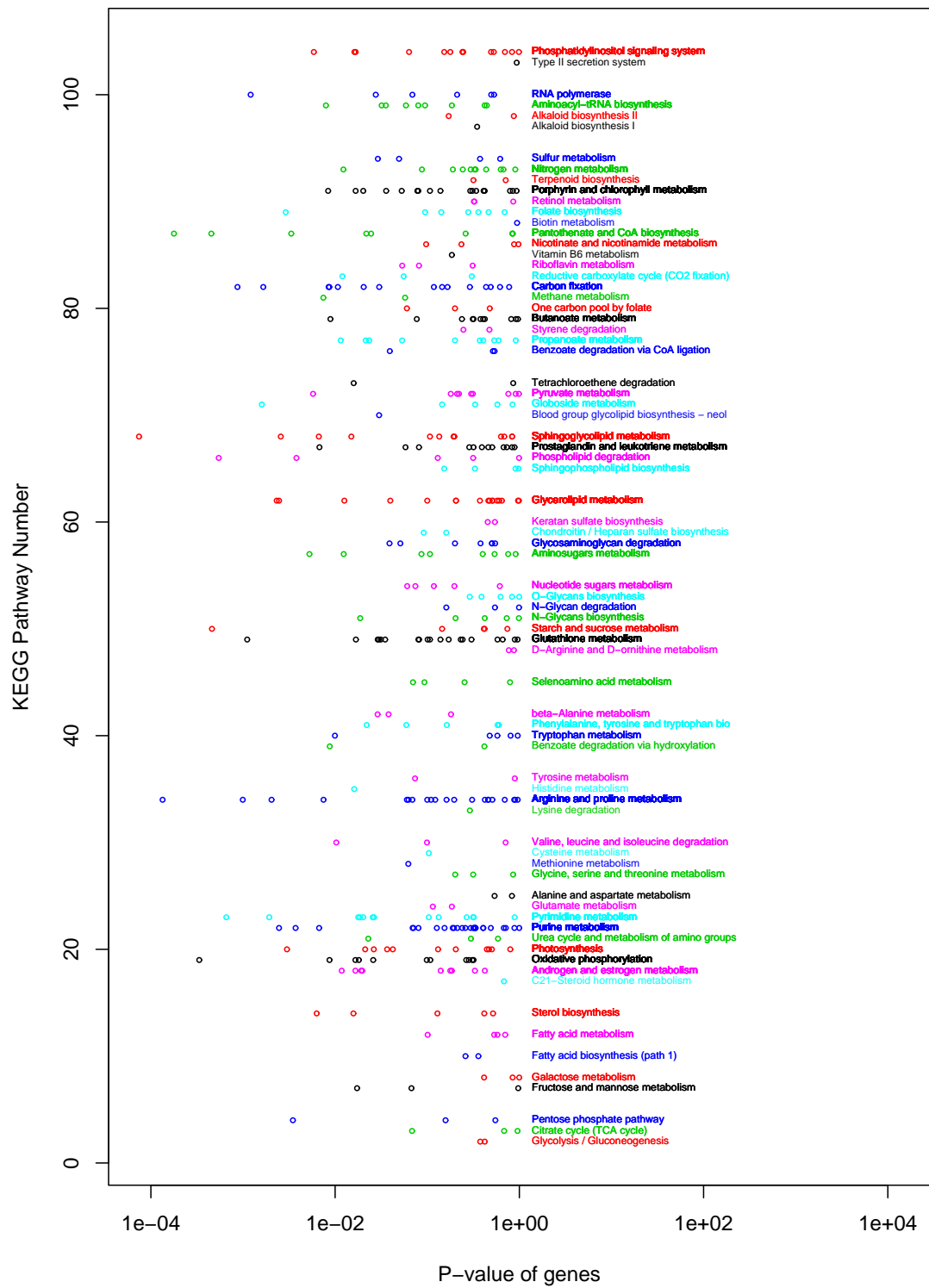


Figure 16: A list of all KEGG pathways in which genes were found. The x-axis shows the P-value of each gene assigned to each pathway. A P-value close to 1 means the gene is almost certain to be unchanged in the experiment. The smaller the P-value, the greater the probability of differential regulation. Pathways with differential expression should stand out from the background level.

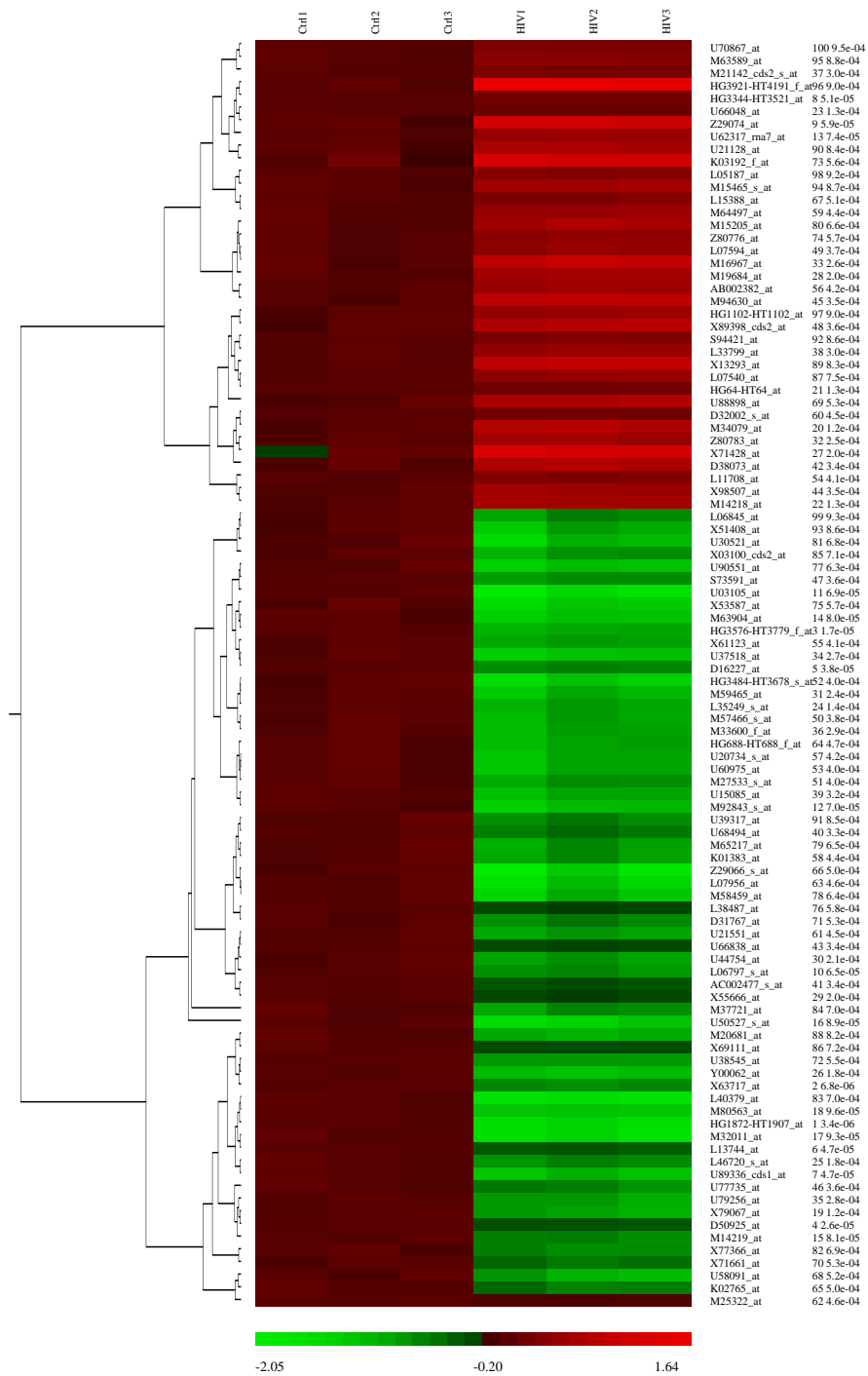


Figure 17: Hierarchical clustering of top ranking genes based on their vector angle distance. The color scale shows for each gene the logarithm of the fold change relative to the average expression in the first category. For each gene, the chip ID, the number referring to Table 2 or Table 3, as well as the P-value are given.

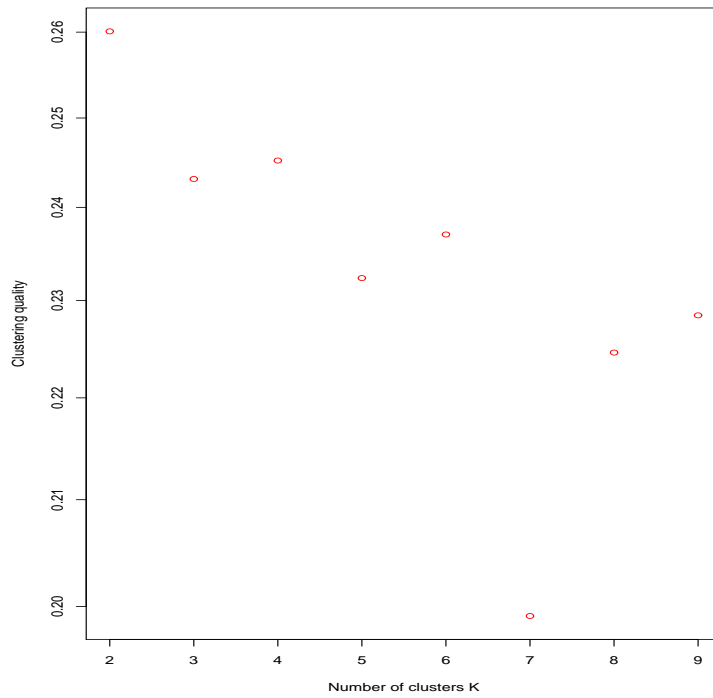


Figure 18: Optimization of the number of clusters K . The clustering quality was measured, for each value of K , as the ratio of between-cluster variance to within-cluster variance. The higher this ratio is, the better the separation into clusters is.

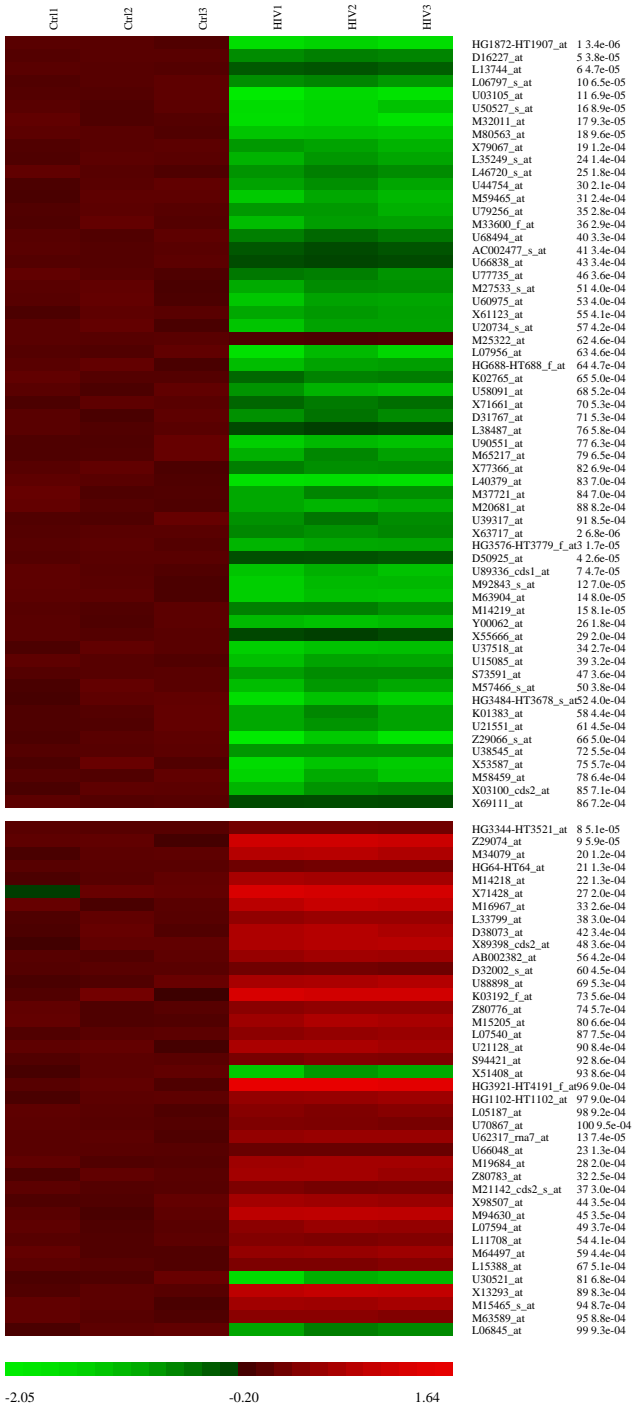


Figure 19: K-means clustering of top ranking genes based on their vector angle distance. The color scale shows for each gene the logarithm of the fold change relative to the average expression in the first category. For each gene, the chip ID, the number referring to Table 2 and Table 3, as well as the P-value are given. The number of clusters, 2, was selected by optimization.

3.10 Promoter analysis

From the K-means clustering the upstream regions were extracted from the genes of each cluster. The software program `saco_patterns`¹⁵ was run on each cluster to identify overrepresented patterns in the upstream regions. Table 7 shows the most overrepresented patterns for each cluster.

Table 7: Analysis of the upstream regions of the K-means clusters with `saco_patterns`. The occurrence of exact matches to each pattern is shown in the cluster (cluster size given in parenthesis) and in the background data set (set size given in parenthesis). The resulting (negative logarithm of the) probability of overrepresentation from the hypergeometric distribution is shown. For each pattern, the genes in which it was found are listed (up to 50 hits). If a pattern was found more than once in a gene, then that gene will appear more than once on the list. The sequence numbers refer to the numbers in the clustering and in the tables of up- and down-regulated genes.

Pattern	$-\log(P)$	In cluster	In bg (4409 genes)	Found in genes
Cluster number 1 (cluster size=60, upstream regions extracted=37)				
Cluster number 2 (cluster size=40, upstream regions extracted=20)				

An overrepresentation *per se* is not enough to signify biological relevance. To further substantiate a pattern, the patterns can be extracted from the upstream regions and aligned with context. If there is conservation in the regions surrounding the pattern then that further supports biological relevance. The final determination will come from biological verification using site-directed mutagenesis or bandshift methods.

The Gibbs sampler¹⁶ was run on the same clusters as `saco_patterns`. The Gibbs sampler looks for degenerate patterns which it tries to capture with a weight matrix description. In all sequences, the best match to this weight matrix is shown in the output. The alignment allows judgment of the degree of conservation. The results are shown below:

¹⁵Jensen, L.J. and S. Knudsen, (2000) Automatic Discovery of Regulatory Patterns in Promoter Regions Based on Whole Cell Expression Data and Functional Annotation. *Bioinformatics* 16:326-333.

¹⁶Lawrence, Altschul, Boguski, Liu, Neuwald & Wootton (1993) 'Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment', *Science* 262:208-214.

Table 8: Weight matrices describing gibbs patterns in upstream regions of K-means clusters. The hypergeometric sample statistics is given as the logarithm of the P-value, where i is the number of times the matrix matches the positive set above threshold, m is the number of times the matrix matches the negative set above threshold, and N and n are the sizes of the negative and positive sets, respectively. For each pattern, the genes in which it was found are listed (up to 50 hits).

Base	1	2	3	4	5	6	7	8	9	10	11
Cluster number 1 (cluster size=60, upstream regions extracted=37)											
HYP -2.698010 $i=11$, $m=941$, $N=4446$, $n=37$											
Consensus: GAGGCGGAGGC											
Found in genes 41 5 5 5 71 6 6 6 88 88 62 62 51 51 14 14 29 55 55 82											
A	3	87	17	0	0	3	10	100	0	0	0
C	3	0	0	0	93	17	3	0	0	7	67
G	93	13	83	100	0	40	87	0	100	93	0
T	0	0	0	0	7	40	0	0	0	0	33
Cluster number 2 (cluster size=40, upstream regions extracted=20)											
HYP -2.475133 $i=6$, $m=806$, $N=4429$, $n=20$											
Consensus: GGAGGCTGAGG											
Found in genes 49 49 49 22 89 27 27 44 44 44 44											
A	5	0	100	0	0	0	10	5	75	0	0
C	0	0	0	0	5	95	5	0	0	0	0
G	90	100	0	100	95	0	20	95	15	90	100
T	5	0	0	0	0	5	65	0	10	10	0

The transcription factor binding sites in Transfac¹⁷ were checked against the same clusters. All eukaryotic factors were matched and the results are shown below:

¹⁷ Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. "TRANSFAC: transcriptional regulation, from patterns to profi les. Nucleic Acids Res. 2003 Jan 1;31(1):374-8.

Table 9: Analysis of the upstream regions of the K-means clusters with Transfac. The occurrence of matches to each Factor is shown in the cluster (cluster size given in parenthesis). More information about the Factors can be found by looking them up in the public version of Transfac at www.gene-regulation.de. For each pattern, the genes in which it was found are listed (up to 50 hits). If a pattern was found more than once in a gene, then that gene will appear more than once on the list.

Factor name	Found in sequences
Cluster number 1 (cluster size=60, upstream regions extracted=37)	
Sp1@human	63
HSF@fruit	41 41 41 5 71 10 10 63 63 63 63 63 6 6 6 6 24 24 76 76 76 15 15 15 15 88 88 88 88 62 62 62 62 62 51 51 51 84 84 84 78 78 78 78 31 31 31 79 79 18
HSF@yeast	41 5 5 5 5 10 10 10 63 63 24 24 24 24 76 88 62 84 84 84 78 31 31 31 31 31 79 79 39 57 61 34 30 43 46 29 55 55 86 86 70 82
CREB@human	61
CRE-BP1/c-Jun@mouse	61
Sox-5@mouse	78
ADR1@yeast	63 76 51 84 57 34 30 30 68 43 86 82 19 19 66
MZF1@human	71 34 55 55
CdxA@chick	41 15 47
CdxA@chick	71 10 6 15 18 61
Bcd@fruit	18
Lyf-1@mouse	41 6 88 14
NIT2@Neurospora	10 24 79 39 57 29 29 66
SRY@mouse	41 5 5 10 63 24 15 88 62 78 78 31 31 18 61 30 30 30 30 30 70 66
HSF@fruit	34
cap@unknown	10
Cluster number 2 (cluster size=40, upstream regions extracted=20)	
HSF@fruit	56 56 56 56 60 60 60 60 60 42 73 73 98 98 98 99 87 87 87 49 49 54 54 54 67 33 33 33 33 95 59 59 59 59 59 90 90 90 90 13 89 89 27 27 27 27 27 27
HSF@yeast	56 60 42 42 42 73 98 49 54 54 54 33 95 59 90 90 90
Sox-5@mouse	54
ADR1@yeast	73 73 99 87 67 67 59 89 27
E2F@mouse	42
CdxA@chick	56 99 90
CdxA@chick	73 33 90
Lyf-1@mouse	49 89 44

NIT2@Neurospora	56
SRY@mouse	60 98 54 44
cap@unknown	95

3.11 Correspondence Analysis

A correspondence analysis was performed on the 50 top ranking genes to look for strong associations between genes and experiments (Figure 20. If there are only two categories, this association does not reveal any new information.) Genes and experiments are each projected into the same two-dimensional space. A gene that is far removed from the center of the plot (0,0) is associated with an experiment if that experiment is also far removed from the center of the plot in the same direction.

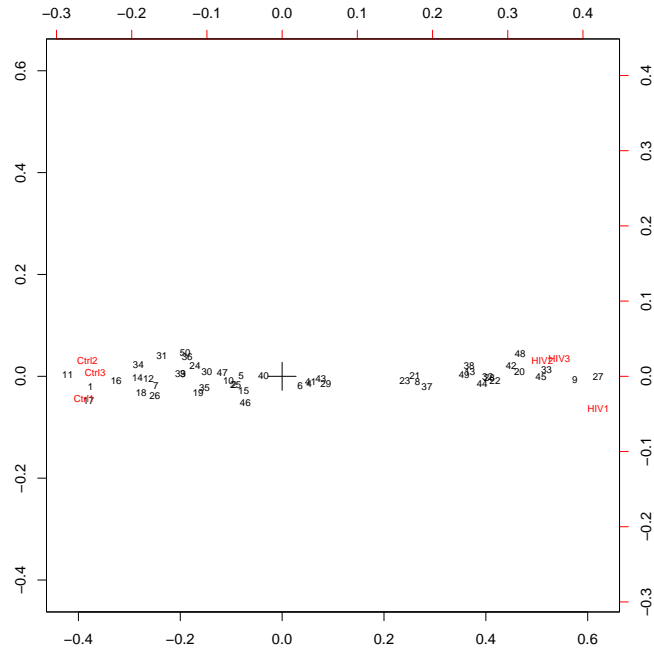


Figure 20: Correspondence analysis of the top 50 ranking genes and the experiments. Genes are shown in one color and experiments are shown in a different color. Gene numbers refer to Table 2 or Table 3.

4 Appendix A: parameters used in this report

Table 10: Parameters set in parameter file.

Parameter	Value (options in parenthesis)
Name of file	dataunorm4.txt
Header	FALSE (TRUE FALSE) Is there a header in the first line of the file?
Columns	1 2 3 4 5 6 7 8 9 10 11 12 13 14
Descriptor	ID AN N N N A A A B B B C C C
File names	day_7a_&CEL day_7b_&CEL day_7c_&CEL day_7a_HIV_&CEL day_7b_HIV_&CEL day_7c_HIV_&CEL
Categories	A A A B B B
Chip Type	HU6800 (HG_Focus HU6800 HG_U95Av2 HG-U133A MG_U74Av2 RG_U34A DrosGenome1 YG_S98 Ecoli Pae_G1a AG Other)
Compressed CEL files	FALSE (TRUE FALSE)
Experiment name	HIV Infection of Human T cells
Author	Steen Knudsen
Organism	hsa (bsu rno pae eco sce dro mmu pae)
A	Ctrl
B	HIV
Category Names	Ctrl1 Ctrl2 Ctrl3 HIV1 HIV2 HIV3
Normalization method	qspline (qspline quantile constant loess contrasts none)
Expression index	li.wong (li.wong avdiff medianpolish)
Remove outliers	FALSE (TRUE FALSE affects only li.wong calculation)
Background correction	bg.adjust (FALSE bg.adjust subtractmm)
Statistical analysis	parametric (parametric)
Paired t-test	FALSE (TRUE FALSE) (if TRUE experiments must appear in the order they are paired)
Minimum cutoff for logfold calculation	1 (1-20)
Show results on X display	FALSE (TRUE FALSE)
Max number of genes to analyze further	100
Bonferroni cutoff (max number of false pos.)	10
Logfold	log2 (log2 log10 hlog)
Color scheme	red-green (blue-yellow)
Include table of all genes as well	NO (YES NO)

