Practical Course Integrative Bioinformatics

– Report –

WS 2012/2013

Florian Aldehoff - Matrikel 3798908
Philipp Brachvogel - Matrikel 3808351
Sebastian Lehnerer - Matrikel 3801387

Eberhard Karls Universität Tübingen - May 18, 2013

Contents

1. Background

1.1. Cancer - on a molecular level

The term cancer describes a broad spectrum of more than 200 diseases with the common characteristic of uncontrolled cellular growth that may lead to tumor formation, invasion of surrounding tissues and metastasis. Cancer can be caused by individual chemical, physical and biological agents and any combinations thereof with only a few strong causative correlations scientifically proven for mutagens and carcinogens like e.g. highly energetic ionizing or non-ionizing radiation, tobacco smoke ingredients, alcohol, aerosolized asbestos, benzene vapor or aflatoxin B1 but also for physical inactivity, obesity, high-salt diet as well as viral infections by so-called oncoviruses like the human papillomavirus and Epstein-Barr virus. One or many of these causative agents, sometimes in combination with additional hereditary predispositions, can affect the expression of both oncogenes and tumor suppressor genes. Oncogene products generally promote cell proliferation and growth and include regulatory GTPases like Ras and transcription factors like myc, while tumor suppressor genes ("anti-oncogenes") generally control apoptosis and inhibit cell division, like the *TP53* transcript p53 which tightly controls the cell cycle and can initiate DNA repair as well as apoptosis in case of severe damage to the DNA.

Possible genetic transformations or mutations range from point mutations like single nucleotide changes, deletions or insertions (leading to premature stopping, reading frame extension or shift during translation), allele or copy number changes by duplication, inversion, deletion or chimerization of entire protein coding sequences or non-coding control regions as well as abnormal splice variants all the way to chromosomal translocations. The resulting differences both in protein expression levels – eg. an increased level of oncogene expression or a decreased level of tumor suppressor gene expression – and in protein activity across several interconnected molecular pathways can set of a chain reaction of self-amplifying, compounding errors that influence important cellular functions such as DNA damage repair, cell proliferation, apoptosis signaling, inter-cellular signaling (eg. hormone reception and secretion or antigen presentation) and adhesion to surrounding cells and the extracellular matrix.

1.2. DNA Microarrays

Changes in expression levels of genes between different samples or tissues can be detected in parallel by immobilizing thousands of microscopic spots of DNA with known sequences and positions (known as probes, reporters or oligos) to a glass or silicon surface (known as DNA chip, biochip or DNA microarray). Multiple identical chips (or – depending on the technology and amount of samples – the very same chip) are hybridized with different samples of fluoroscope-labeled complementary DNA (cDNA) targets generated by whole-RNA extraction, reverse transcription, labeling and fragmentation. A washing step then removes all partially complementary strands, leaving only complementary probe-target pairs on the chip. These can then be detected and quantified by the means of laser excitation and fluorescence scanning.

The Affymetrix Human Exon 1.0 ST Array in the transcript (gene) version (see http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL5175 for details) allows to distinguish between different isoforms of genes as it includes about four probes per exon and 40 probes for gene that are collectively called a probeset. This makes it a useful tool when investigating chimeric genes as those discovered in human melanoma cultures by Berger et al. [1].

1.3. Gene Ontology

The Gene Ontology initiative (GO, http://www.geneontology.org/), a part of the Open Biological and Biomedical Ontology Foundry (OBO, http://www.obofoundry.org/), attempts to unify the representation of genes and gene product attributes by the means of so-called GO terms which provide a controlled, species-neutral vocabulary that can be applied to both prokaryotic and eukaryotic organisms. The terms, each consisting of a unique numeric 7-digit accession with the prefix "GO:" (eg. GO:0008150), a name, a definition and various fields for cross-references and synonyms, fall into one of three domains or namespaces: cellular component, molecular function

and biological process. The entire ontology of GO terms is structured as directed acyclic graph (DAG) with defined relationships between GO terms within a namespace.

A selection of Python scripts called goatools (https://github.com/tanghaibao/goatools/) can be used for enrichment of GO terms. We are using the class GODag and specifically its method query_terms to query the GO term descriptions for the GO IDs translated by the BridgeDB client.

1.4. GSEA

Gene Set Enrichment Analysis (GSEA), a computational method developed by [2] and implemented in a freely available Java software package by [3] (available at http://www.broadinstitute.org/gsea/index.jsp), determines if a defined set of genes shows statistically significant expression level differences between two samples. Instead of focussing only on differences in expression levels of individual genes, their regulatory or functional relationships with other genes within a gene set are considered as well to allow detection of changes in entire metabolic pathways, transcriptional programs and stress responses which would be too subtle and distributed for other statistical tests. The method consists of the following steps:

1. calculation of enrichment scores (ES)

   Scoring gene sets by iterating over a ranked list of all genes and either increasing a running-sum statistic when a gene is found in the set or decreasing it when not, with the amount of the increment depending on the correlation of the gene with the phenotype.

2. estimation of significance level

   Instead of permuting genes, the phenotype (or in our case: tissue) labels are randomly permuted while maintaining the gene correlations and enrichment scores are recalculated for all permutations to generate an ES null distribution, which is then used to calculate empirical, nominal P values for the observed ES.

3. adjustments for Multiple Hypothesis Testing

   If testing against an entire database of gene sets, the ES values for all gene sets are normalized (NES) and false discovery rates (FDR) are calculated by comparing the tails the observed NES distribution and the null NES distribution calculated from the ES values in step 2.

The core genes within a gene set that account for most of the enrichment signal can be identified by analysis of so-called leading-edge graphs which show the leading-edge subset of genes in the ranked list at or before the maximum deviation of the running-sum statistic from zero. These biologically significant subsets of genes can be of special interest when examining curated gene sets consisting of multiple pathways.

1.5. Z-Scores

The standard score, or so-called z-score, describes how many standard deviations an data point is above or below the mean. Assuming a normal distribution (mean = median = 2. quantile) the z-score therefore indicates how large the distance of a data point from the population mean is. With the z-score it is possible to make standardized statements about a bunch of different data sets, like how many data points are equal, above or below the mean. Doing so, one can compare data without referring to the absolute values of the data.

1.6. p-Value

The p-value indicates the probability of obtaining a certain data point at least as extreme as actually observed assuming a null hypotheses (e.g. common normal distribution). By applying a threshold (0.05) one rejects the null hypotheses if the value lies below the threshold, indicating that such an event is very unlikely to happen under this distribution.

1.7. Fisher's exact test

In order to assign significance to the up regulation we used Fisher's exact test [4]. The test is based upon an contingency table counting the number of significant up regulated genes (according to their p-value with an threshold of 0.05) for each tissue. As Fisher's test uses categorical data (nominal) we labeled up regulated (down regulated) genes, holding an significant low p-value, with "TRUE" where not significant varied genes are labeled "FALSE". Basically Fisher's test computes a probability under the assumption of a hyper geometric distribution, for this exact contingency table under the null hypotheses that all samples are equally distributed. Low probability values therefore indicate that the underlying data set can hardly be explained by one common distribution, hence its likely that there are profound differences.

Prerequisite for Fisher's test is a large enough sample size (more the 100 samples) for each cell in the contingency table which is given for our data set (about 600 data points).

2. Implementation

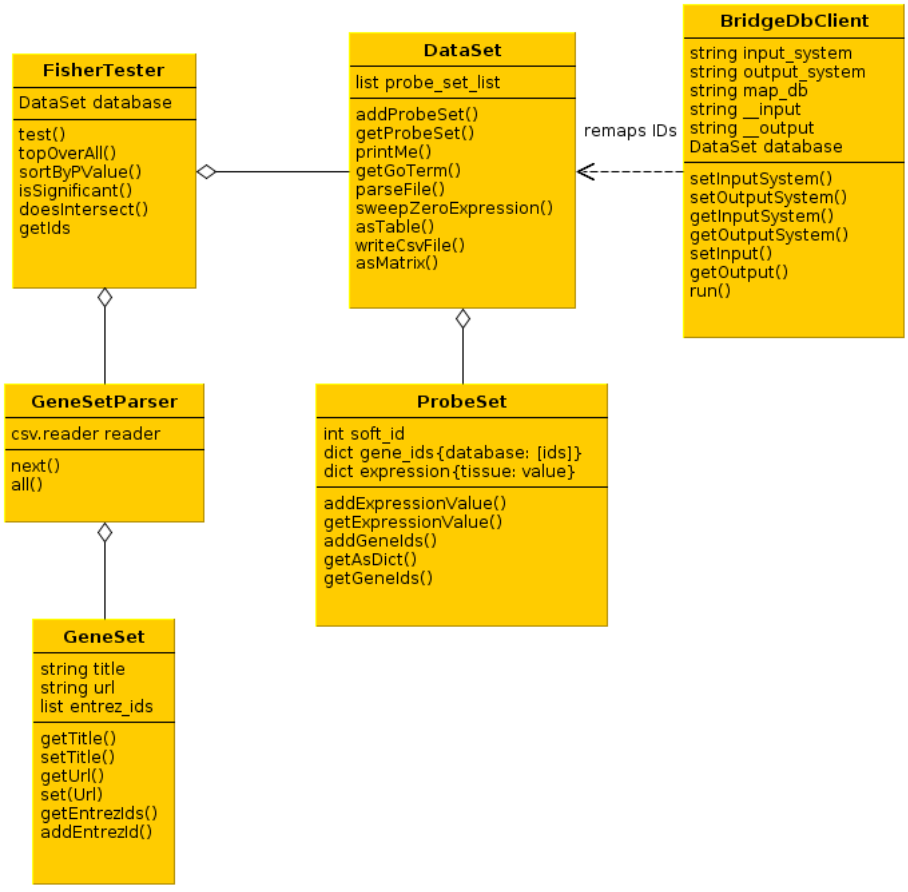

Figure 1: UML scheme

2.1. Subversion Repository

Python source code for all implemented scripts, UML schemata and the documentation (including this report) is hosted at https://xp-dev.com/svn/pibi using the Subversion (SVN) version control system.

2.2. Geo SOFT parser for Affymetrix MicroArray data

The classes ProbeSet and DataSet were designed to represent the expression values from an Affymetrix Human Exon 1.0 ST Array as used in [1]. The experiment-specific SOFT ID for each probeset serves as unique key or identifier for instances of the ProbeSet class, which also have a dictionary of gene identifiers by database and a dictionary of expression values by tissue as attributes. An instance of a DataSet is composed of a list of ProbeSets. Instead of the suggested GEO parser included in the Bio package we use our own parser to read the SOFT file obtained from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17593.

The parser method parseFile() processes a Geo SOFT file (eg. GSE17593_family.soft from aforementioned experiment) by first scanning the platform section of the Affymetrix array (identified by the ID code "GPL5175") for Ensemble and NCBI gene IDs associated with each SOFT ID and adding them to the corresponding dictionaries. The parser then proceeds to the sample section and reads all expression values from samples that were present on the Affymetrix array, thus excluding any GenomeAnalyzer II RNAseq and SNP array data.

In a final step, the method sweepZeroExpression() is called to remove all ProbeSets from the DataSet that do not contain expression values - these entries were created in the initial parsing step of the platform table but not populated in the second step. The resulting list of ProbeSets contains 17381 entries, matching the expected number of unique Affymetrix Human Exon 1.0 ST Array probesets that are included in the original SOFT file.

2.3. Wrapping of BridgeDB for translation and aggregation of gene IDs

The translation of Ensembl IDs to gene identifiers used in other repositories for genetic information like Entrez, UniProt and Gene Ontology requires access to the ID mapping framework like BridgeDb (http://www.bridgedb.org/), which is available as a local web service within the WSI network (http://pride:8183) and can be queried through HTTP GET requests to specific URIs as listed and explained at http://www.bridgedb.org/wiki/BridgeWebservice. As repeated HTTP requests for all IDs in the DataSet would likely fail due to network latency we use batchmapper.sh, a scripted version of BridgeDb available at /share/opt/noarc/BridgeDB/bridgedb−1.1.0/batchmapper.sh, to speed up the ID mapping process. The class BridgeDbClient has attributes for all parameters of the script and wraps its functionality in the remap() method by setting the input system code (EnHs = Ensembl Homo sapiens), the output system code (L = Entrez, S = UniProt or T = GeneOntology ID), the path to an Apache Derby database containing the mappings (/share/data/bridgedb/Hs_Derby_20110601.bridge) and calling the shell script via os.system() in the run() method. The temporary output file created by the script is then parsed by getOutput() and either directly appended to the DataSet (if provided as input) or returned as list.

2.4. Retrieval of GO terms using goatools

With GO IDs now being available in the DataSet a method to also query the corresponding GO term descriptions is needed. We use the goatools module available at https://github.com/tanghaibao/goatools/ to parse the OBO flat file format that stores all GO IDs and their descriptions (see http://www.geneontology.org/GO.format.obo-1_2.shtml for the v1.2 specification). A call of getGoTerm() on a DataSet instance passes the queried GO ID to the method query_term of a GODag instance working on a recent Gene Ontology OBO flat file (downloaded from http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo) and returns the GO term desciption as output string.

2.5. Visualization of expression values as heatmap

A heatmap of expression values was created using the R script heatmap.R provided by the tutors. It reads the CSV output of a DataSet (specifically columns 8 to 16, which contain the expression values by tissue sample) in line 1, determines the mean expression value of each probeset over all nine tissues (line 2) as well as the maximum deviation from the mean per probeset row (line 3), then filters out samples with less than 1.0 deviation from the mean (remaining probesets are at least twofold over- or underexpressed compared to the mean) in line 4. The apply() function in

line 5 first centers all filtered probesets by subtracting the means over all tissue columns from the individual expression values and then scaled by dividing them by their standard deviations (see http://stat.ethz.ch/R-manual/R-devel/library/base/html/scale.html). Lines 6 to 8 then create a JPEG file of the output of heatmap(), which produces a false-color representation of expression value z-scores on a yellow-red scale with additional dendrograms of distances on both the sample and the probeset axis.

```
1 d <-read.table("../data_out.csv", header=T, sep=",")
2 d_mean <- rowMeans(d[,8:16])
3 d_maxfc <- apply(d[,8:16],1,function(x) max(max(x)-mean(x), abs(min(x)-mean
      (x))))
4 d_filtered <- d[which(d_maxfc > 1.0),]
5 d_filtered_z <- apply(d_filtered[,8:16],1,scale)
6 jpeg('heatmap.jpg')
7 heatmap(d_filtered_z)
8 dev.off()
```

2.6. Calculation of z-scores and p-values

The z-score is calculated as follows:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the expected value and σ the standard deviation. An important distinction is whether to take the mean and deviation from the columns or the rows. In our case, columns stand for different tissues, rows for different genes. Therefore we used the mean and standard deviation for each row in order to compare the tissues regarding a certain gene.

In our implementation we calculated the p-values upon the z-scores obtained from the expression levels. For computation we used the scipy function ndtr() which returns the area under the standard Gaussian probability density function starting from a point x to infinity. We used a two-sided p-value, therefore the final function call was as follows:

```
1 2 * special.ndtr(numpy.absolute(z_score) * -1)
```

2.7. Parsing GMT file for gene sets

GMT are text-based files that contain one geneset per row. Each tab-delimited row has a geneset title string, a URL pointing to metadata on the geneset and a list of comma-separated Entrez IDs. The structure of these data was modeled in a basic GeneSet class with three attributes and the corresponding get and set methods. The class GeneSetParser is responsible for reading the passed GMT file and returning GeneSet entities - either individually or as list. As the GMT format is basically a CSV format with tab instead of comma delimiters the GeneSetParser simply wraps the csv.reader() class and next() method and splits the row items into title, URL and Entrez IDs. An additional all() method returns a list of all GeneSet instances when called.

2.8. Gene set enrichment

2.8.1. Fisher exact test

As described in section 1.7, the fisher's exact test is based on contingency tables. Therefore we implemented a method counting every significant gene (p-value < 0.05) for every tissue contained in a gene set. The contingency table is structured as follows:

$$\begin{bmatrix} \text{significant \& in geneset} & \text{significant \& not in geneset} \\ \text{not significant \& in geneset} & \text{not significant \& not in geneset} \end{bmatrix}$$

Each cell contains the gene count for a combination of gene set and tissue with the specified characteristics.

For every contingency table a fisher's exact test was computed using the scipy implementation. As scoring measure we used the p-value received from the test. As result we received a list of all

gene sets for every tissue with an attached p-value, indicating if the geneset is significantly up- or down regulated within the current tissue.

2.8.2. GSEA

For GSEA we used pure expression levels as provided in the dataset as GSEA will perform standardizing methods by its own (see [3] for details) For output we had to filter out those genes with no matching Entrez IDs, as GSEA matches those IDs to genesets. Furthermore GSEA filters out those genesets which have less then 15 or more then 500 hits within our data. All settings were left at their default values. As we did not have control groups, every tissue has been tested against all others, resulting in 18 different runs. HTML output and shows relevant plots were automatically created by GSEA.

3. Results

3.1. Heatmap

The expression values of all 17,381 probesets across all 9 samples were analysed by the heatmap.R script to produce a false-color representation of expression value z-scores on a yellow-red scale with additional dendrograms of distances on both the sample and the probeset axis (see figure 2). The filtering process implemented in R, which is similar to our own statistical analysis and filtering, results in 689 probesets above the threshold of 2fold over- or underexpression.

The heatmap provides a general overview of the dataset and confirms the functionality of all processing steps up to this point. Additionally it can be used to compare the output of our implementations for z-score and p value calculation as these were implemented separately in R and Python.

Unfortunately the very high number of probesets combined with the limited resolution of the output figure make it impossible to read the individual gene IDs at the bottom of the figure along the x axis, which are needed for interpretation of the dendrogram clusters depicted above figure. However, the dendrogram of the 9 tissues used in our analysis remains interpretable and shows tissue #4 (GSM433779) and #8 (GSM433783) as having a different expression pattern than the rest. In tissue #4 several clusters of genes (depicted in red, left half of heatmap) are up-regulated while a large contiguous area (depicted in yellow, third quarter of heatmap) is down-regulated. The pattern of #8 is a bit different from that of #4. Here the over-expressed and down-regulated regions are a bit shifted and shorter but roughly in the same regions. Neither of these special patterns within #4 and #8 can be found in any of the other tissues, which warrants a closer examination of tissue GSM433779 and GSM433783.

3.2. Exploration of enriched Gene Sets

3.2.1. Fisher's exact test

To all tissues the Fischer's exact test was applied to gene sets correlating to the KEGG pathways and a GO term gene set database. For both gene set classes we calculated a list of the 30 sets with the lowest p values among all tissues. Many cancer unrelated pathways were found in the list like ALZHEIMERS DISEASE, LYSOSOME, ECM RECEPTOR INTERACTION, TYPE I DIABETES MELLITUS, PARKINSONS DISEASE, KHUNTINGTONS DISEASE. Most of them were connected to tissue GSM433779. Some of the cancer related gene sets within the list were KEGG FOCAL ADHESION, PATHWAYS IN CANCER, SMALL CELL LUNG CANCER and MELANOGENESIS. All four were found in tissue GSM433783, GSM433782 and GSM433781, but only tissue GSM433783 contained all of them.

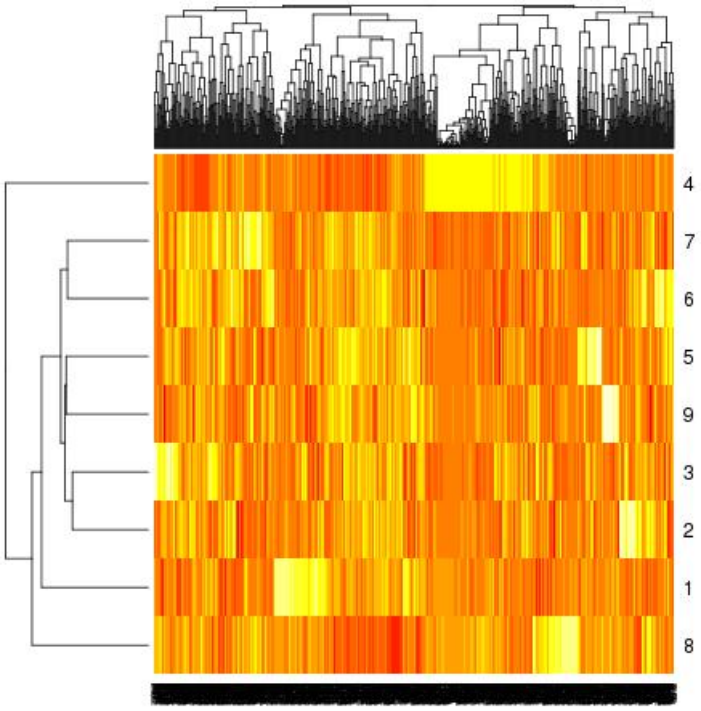

Figure 2: Heatmap of expression values; color values range from low (yellow) to high (red) expression; numbers correspond to the following tissue samples: 1 = GSM433776, 2 = GSM433777, 3 = GSM433778, 4 = GSM433779, 5 = GSM433780, 6 = GSM433781, 7 = GSM433782, 8 = GSM433783, 9 = GSM433784; the x-axis shows clustering of 689 probesets by similarity of expression values while the y-axis shows clustering of 9 tissues by similarity of expression values

Table 1: Top 30 KEGG pathways as produced by Fisher exact test for gene set enrichment analysis; the KEGG term KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC has been shortened to KEGG_ARVC to fit the page

tissue	**KEGG pathway**	**p value**
GSM433783	KEGG_FOCAL_ADHESION	$2.0477 \cdot 10^{-2}$
GSM433779	KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	$4.2654 \cdot 10^{-2}$
GSM433779	KEGG_ALZHEIMERS_DISEASE	$6.9883 \cdot 10^{-2}$
GSM433779	KEGG_LYSOSOME	$7.2256 \cdot 10^{-2}$
GSM433783	KEGG_ECM_RECEPTOR_INTERACTION	$8.2122 \cdot 10^{-2}$
GSM433783	KEGG_LYSOSOME	$1.1015 \cdot 10^{-1}$
GSM433783	KEGG_PATHWAYS_IN_CANCER	$1.2493 \cdot 10^{-1}$
GSM433779	KEGG_OOCYTE_MEIOSIS	$1.6001 \cdot 10^{-1}$
GSM433779	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	$1.6001 \cdot 10^{-1}$
GSM433779	KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	$1.6001 \cdot 10^{-1}$
GSM433779	KEGG_TYPE_I_DIABETES_MELLITUS	$1.6001 \cdot 10^{-1}$
GSM433779	KEGG_GRAFT_VERSUS_HOST_DISEASE	$1.6001 \cdot 10^{-1}$
GSM433779	KEGG_OXIDATIVE_PHOSPHORYLATION	$1.6211 \cdot 10^{-1}$
GSM433779	KEGG_HEMATOPOIETIC_CELL_LINEAGE	$1.6211 \cdot 10^{-1}$
GSM433782	KEGG_PATHWAYS_IN_CANCER	$1.8626 \cdot 10^{-1}$
GSM433783	KEGG_ADHERENS_JUNCTION	$1.9733 \cdot 10^{-1}$
GSM433779	KEGG_PARKINSONS_DISEASE	$1.9818 \cdot 10^{-1}$
GSM433779	KEGG_HUNTINGTONS_DISEASE	$1.9818 \cdot 10^{-1}$
GSM433779	KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	$2.0817 \cdot 10^{-1}$

continued on next page

tissue	KEGG pathway	p value
GSM433783	KEGG_AXON_GUIDANCE	$2.1736 \cdot 10^{-1}$
GSM433781	KEGG_PATHWAYS_IN_CANCER	$2.3590 \cdot 10^{-1}$
GSM433783	KEGG_SMALL_CELL_LUNG_CANCER	$2.3938 \cdot 10^{-1}$
GSM433783	KEGG_ARVC	$2.6359 \cdot 10^{-1}$
GSM433779	KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	$2.7070 \cdot 10^{-1}$
GSM433779	KEGG_ASTHMA	$2.7070 \cdot 10^{-1}$
GSM433779	KEGG_AUTOIMMUNE_THYROID_DISEASE	$2.7070 \cdot 10^{-1}$
GSM433779	KEGG_ALLOGRAFT_REJECTION	$2.7070 \cdot 10^{-1}$
GSM433776	KEGG_WNT_SIGNALING_PATHWAY	$2.7829 \cdot 10^{-1}$
GSM433783	KEGG_WNT_SIGNALING_PATHWAY	$2.9020 \cdot 10^{-1}$
GSM433783	KEGG_MELANOGENESIS	$2.9020 \cdot 10^{-1}$

The p-Values for the GO terms were calculated as required, but we decided to disregard those results for the discussion as their usefulness is highly doubtful. The file `c5.all.v3.1.entrez.gmt` only provides a single GO term without any GO-ID or additional GO terms complementing the given single GO term. Any GO term for a set is either a **single** cellular component, biological process or molecular function. Thus a result like "NUCLEUS" without any additional terms might point to anything from nucleus transport protein over signal molecule to cancer associated proteins like p53. The only interesting result might have been to find a lot of biological processes or molecular functions associated with cancer like behavior, but all top scoring GO terms stood for not cancer associated cellular components or biological processes.

Table 2: Top 30 GO terms as produced by Fisher exact test for gene set enrichment analysis

tissue	GO term	p value
GSM433776	ANATOMICAL_STRUCTURE_DEVELOPMENT	$4.1484 \cdot 10^{-3}$
GSM433776	MULTICELLULAR_ORGANISMAL_DEVELOPMENT	$6.0841 \cdot 10^{-3}$
GSM433776	ORGAN_DEVELOPMENT	$1.0661 \cdot 10^{-2}$
GSM433776	SYSTEM_DEVELOPMENT	$1.4116 \cdot 10^{-2}$
GSM433779	NUCLEUS	$1.6135 \cdot 10^{-2}$
GSM433784	MEMBRANE	$1.6535 \cdot 10^{-2}$
GSM433776	EXTRACELLULAR_REGION	$2.4333 \cdot 10^{-2}$
GSM433779	NUCLEAR_PART	$3.2700 \cdot 10^{-2}$
GSM433776	ANATOMICAL_STRUCTURE_MORPHOGENESIS	$3.3341 \cdot 10^{-2}$
GSM433784	MEMBRANE_PART	$3.3632 \cdot 10^{-2}$
GSM433776	TRANSMEMBRANE_RECEPTOR_ACTIVITY	$3.6945 \cdot 10^{-2}$
GSM433784	PLASMA_MEMBRANE	$6.2403 \cdot 10^{-2}$
GSM433784	INTRINSIC_TO_MEMBRANE	$6.9048 \cdot 10^{-2}$
GSM433784	INTEGRAL_TO_MEMBRANE	$6.9048 \cdot 10^{-2}$
GSM433779	ORGANELLE_LUMEN	$7.2481 \cdot 10^{-2}$
GSM433779	ENVELOPE	$7.2481 \cdot 10^{-2}$
GSM433779	ORGANELLE_ENVELOPE	$7.2481 \cdot 10^{-2}$
GSM433779	MEMBRANE_ENCLOSED_LUMEN	$7.2481 \cdot 10^{-2}$
GSM433776	LIPID_METABOLIC_PROCESS	$7.5447 \cdot 10^{-2}$
GSM433784	PROTEIN_METABOLIC_PROCESS	$7.6362 \cdot 10^{-2}$
GSM433776	INTRINSIC_TO_MEMBRANE	$8.0530 \cdot 10^{-2}$
GSM433776	INTEGRAL_TO_MEMBRANE	$8.0530 \cdot 10^{-2}$
GSM433776	EXTRACELLULAR_REGION_PART	$8.1680 \cdot 10^{-2}$
GSM433779	MITOCHONDRION	$8.6676 \cdot 10^{-2}$
GSM433776	CELLULAR_LOCALIZATION	$9.2386 \cdot 10^{-2}$
GSM433784	CELLULAR_MACROMOLECULE_METABOLIC_PROCESS	$9.3255 \cdot 10^{-2}$
GSM433784	CELLULAR_PROTEIN_METABOLIC_PROCESS	$9.9635 \cdot 10^{-2}$
GSM433783	PROTEIN_DIMERIZATION_ACTIVITY	$9.9893 \cdot 10^{-2}$
GSM433776	ENDOPLASMIC_RETICULUM	$1.0221 \cdot 10^{-1}$
GSM433776	ESTABLISHMENT_OF_CELLULAR_LOCALIZATION	$1.0221 \cdot 10^{-1}$

3.2.2. Gene Set Enrichment Analysis

The Gene Set Enrichment Analysis initially rejects 174 of the 186 given gene sets, leaving 12 sets that show significant deregulation among the nine given cancer cell lines.

Table 3: List of pathways used in GSEA

list of selected pathways
KEGG_MAPK_SIGNALING_PATHWAY
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
KEGG_LYSOSOME
KEGG_AXON_GUIDANCE
KEGG_FOCAL_ADHESION
KEGG_ECM_RECEPTOR_INTERACTION
KEGG_CELL_ADHESION_MOLECULES_CAMS
KEGG_ADHERENS_JUNCTION
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON
KEGG_ALZHEIMERS_DISEASE
KEGG_PATHWAYS_IN_CANCER
KEGG_SMALL_CELL_LUNG_CANCER

The results of Fisher's exact test indicated that in tissue GSM433783 genes correlated with oncogenic pathways were deregulated compared to the other cell lines. For that reason the GSEA results for GSM433783 were checked in more detail. The GSEA output showed the focal adhesion, small cell lung cancer and pathways in cancer pathways to be relevant again. Their p-Values were among the lowest. Except the lysosome pathways all other findings were regulated down compared to the rest, as can be seen from the negative enrichment scores.

Table 4: GSEA enrichment scores and p-values for sample GSM433783 compared to the remaining tissues

NAME	**ES**	**NOM p-val**
KEGG_FOCAL_ADHESION	-0.27030912	0.00998004
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.052738335
KEGG_ADHERENS_JUNCTION	-0.3143161	0.055666003
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.065737054
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.083657585
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.09343936
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.12627292
KEGG_ECM_RECEPTOR_INTERACTION	-0.22219512	0.16763006
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.37669903
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.6065259
KEGG_AXON_GUIDANCE	-0.16156015	0.77734375
KEGG_LYSOSOME	0.22209951	0.2090164

3.3. Visualization of selected pathways with BiNA

The three pathways identified both by Fischer's exact test and GSEA ("Pathways in cancer", "Small cell lung cancer" and "Focal adhesion") were further analyzed by using the raw expression values and mapping them onto the pathway illustrations using the tool BiNA (Bioinformatic Network Analysis). By visual inspection of the three pathways, it is possible to identify considerably down-regulated genes within GSM433783 in comparison to the other tissues. "Pathways in cancer" and "Small cell lung cancer" contain exactly the same two deregulated genes: *myc*

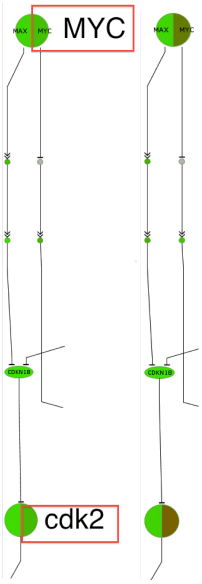

Figure 3: A small section from the KEGG pathway *"Pathways in cancer"* visualized by BiNA from the given expression values for the nine tissues. Left: in tissue GSM433783 the two proteins *myc* and *cdk2* are marked and display average expression levels as indicated by the green color of their nodes. Right: the same section but this time the average over all remaining samples. In average the same proteins have here significantly higher expression levels as can be deduced from the red color for their node. All other proteins are unchanged in comparison to GSM433783.

and *cdk2* while in "focal adhesion" *catenin*. All three genes are down-regulated in comparison to the remaining tissues. On the other hand we could not to identify any up-regulated genes for GSM433783 in any of these pathways.

4. Discussion

As seen already in the heat map (figure 2), the tissue expression values were quite heterogeneous especially the ones of samples GSM433783 and GSM433779. The Fischer's exact test provides additional evidence that GSM433779 has a unique pattern of gene expression, but a closer examination of the affected KEGG pathways shows that these differences are all related to non-carcinogenic pathways. As a consequence of these findings tissue GSM433779 has been excluded from further pathway analysis.

In contrast, a closer examination of differentially regulated pathways in sample GSM433783 yields several pathways containing oncogenes. "Pathways in cancer" and "Small cell lung cancer" are obvious reasons candidates, as is the "Focal adhesion" pathway due to the effect of intercellular contacts and adhesion to the extracellular matrix (ECM)) on the ability of cancerogenic cells to enter metastasis. The loss of self induced cell death upon losing connection to neighboring cells and the resulting motility of cancerogenic cells are important steps in oncogenesis. Contrary to this the last of the examined pathways – "Lysosome" – to our knowledge has no major impact on cancer proliferation.

The three cancer related pathways "Pathways in cancer", "Small cell lung cancer" and "Focal adhesion" were confirmed to be significantly deregulated by the GSEA method. But GSEA also provides additional information by exhibiting that these pathways are under-expressed in GSM433783 compared to all other tissues. To identify genes with the strongest contribution, pathway visualizations were created using BiNA which were manually inspected for interesting candidates (see figures 3, 4 and 5). The target genes *myc*, *cdk2* and *catenin* can be identified as significantly under-expressed compared to the average of all remaining tissues.

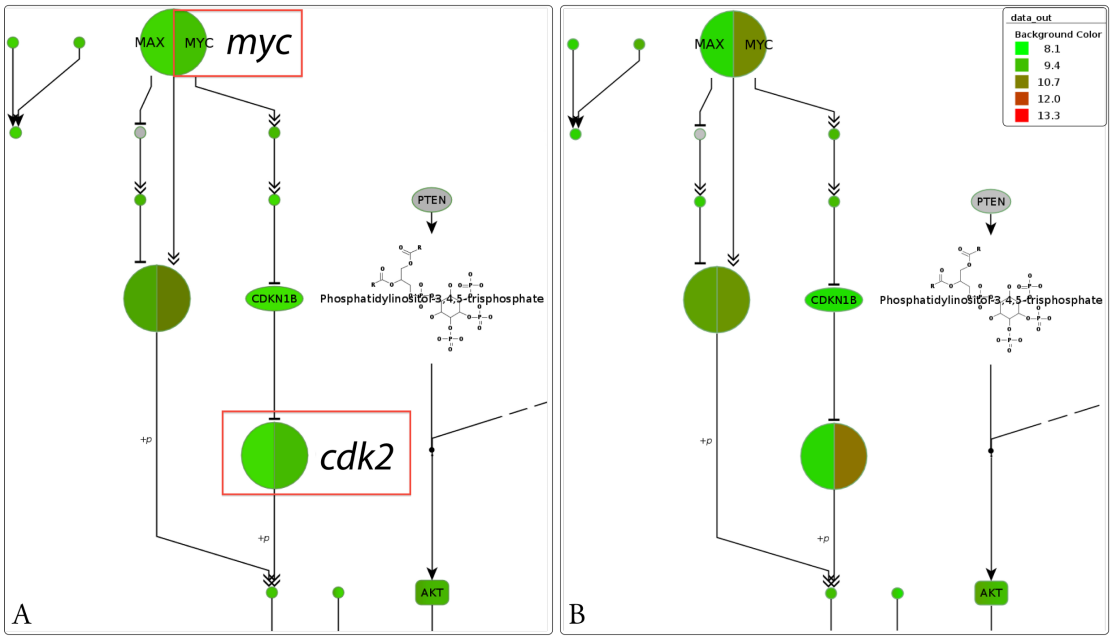

Figure 4: A section from the KEGG pathway *"Small cell lung cancer"* visualized by BiNA from the given expression values for the nine tissues. A: in tissue GSM433783 the two proteins *myc* and *cdk2* are marked and display average expression levels as indicated by the green color of their nodes. B: the same section but this time the average over all remaining samples. In average the same proteins have here significantly higher expression levels as can be deduced from the red color for their node. All other proteins are unchanged in comparison to GSM433783. This is basically the same as in the figure of *"Pathways in cancer"*, because we are looking at the same expression values, but this time in a slightly different context.

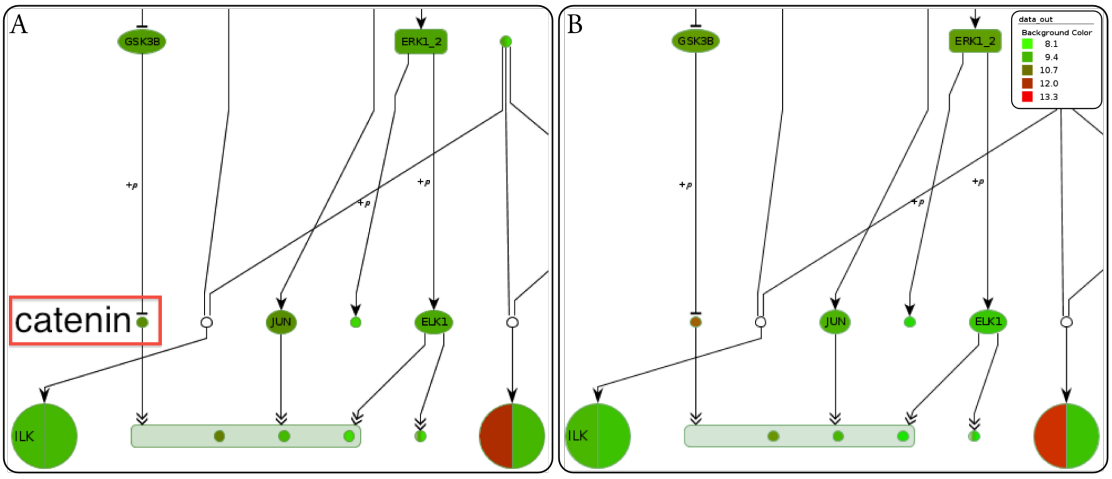

Figure 5: A section from the KEGG pathway *"Focal adhesion"* visualized by BiNA from the given expression values for the nine tissues. A: in tissue GSM433783 the protein *catenin* was marked. Its expression value is comparatively low as can be seen by the green coloring. B: the average expression values over all remaining samples. For almost every protein they are the same but only *catenin* is expressed stronger, as illustraed by the red coloring.

The protein *Cdk2* is of major importance to promote the cell cycle and can be an oncogen if over expressed[5]. High expression levels would result in enhanced proliferation. *Myc* is also a known oncogene[6] with a central role in oncogenesis. It regulates the expression of several cell cycle promoting proteins, one example being shown in the BiNA pathway of "Pathways in cancer" (figure 3), where *myc* promotes *cdk2*. Over-expression of *myc* is a characteristic feature of tumor cells leading to over-proliferation. Within the "Focal adhesion" (figure 5) pathway *catenin* was found as being down-regulated. Over-expression of *catenin* is strongly associated with cancer[7] due to the genes' nuclear localization and possible interaction with multiple transcription factors inducing proliferation.

We conclude that due to the significantly lower expression values for several oncogenes and related pathways cancer cells of tissue GSM433783 may have a lower proliferation rate than the other eight cell lines. Although the aforementioned genes play major roles in cancer, their lower expression may not necessarily have a very strong effect on the proliferation of these specific cancer cells. Enhanced proliferation is only one of many aspects of cancer and normally many alternative pathways exist in the cellular regulatory network to compensate down- and up-regulation in individual genes or pathways. Other important properties of cancer cells such as independence of correct cell contacts, motility and immunity to apoptosis may not be affected at all.

4.1. Biological significance

The given dataset contained nine melanoma tissues but no healthy control tissue. Thus, all analysis steps implemented in the course could only be applied to compare expression levels and regulation of pathways between the nine different cancer tissues, but not with a healthy control tissue. Such a comparison is likely to reveal that the overall difference between the nine cell lines is of minor magnitude. The outcome of our analysis therefore has limited biological relevance and the differences in pathway regulation are expected to be mostly insignificant.

References

[1] M. F. Berger, *et al.*, *Genome Res* **20**, 413 (2010).

[2] V. K. Mootha, *et al.*, *Nat Genet* **34**, 267 (2003).

[3] A. Subramanian, *et al.*, *Proc Natl Acad Sci U S A* **102**, 15545 (2005).

[4] R. A. Fisher, *Journal of the Royal Statistical Society* **85**, 87 (1922). JSTOR 2340521.

[5] J. Du, *et al.*, *Cancer Cell* **6**, 565 (2004).

[6] C. V. Dang, K. A. O'donnell, T. Juopperi, *Cancer Cell* **8**, 177 (2005).

[7] B. T. MacDonald, K. Tamai, X. He, *Dev Cell* **17**, 9 (2009).

A. Supplements

Table 5: Top 10 GO terms as produced by Fisher exact test for gene set enrichment analysis

tissue	GO term	p value
GSM433776	ANATOMICAL_STRUCTURE_DEVELOPMENT	$4.1484 \cdot 10^{-3}$
	MULTICELLULAR_ORGANISMAL_DEVELOPMENT	$6.0840 \cdot 10^{-3}$
	ORGAN_DEVELOPMENT	$1.0660 \cdot 10^{-2}$
	SYSTEM_DEVELOPMENT	$1.4116 \cdot 10^{-2}$
	EXTRACELLULAR_REGION	$2.4333 \cdot 10^{-2}$
	ANATOMICAL_STRUCTURE_MORPHOGENESIS	$3.3340 \cdot 10^{-2}$
	TRANSMEMBRANE_RECEPTOR_ACTIVITY	$3.6945 \cdot 10^{-2}$
	LIPID_METABOLIC_PROCESS	$7.5447 \cdot 10^{-2}$
	INTRINSIC_TO_MEMBRANE	$8.0530 \cdot 10^{-2}$
	INTEGRAL_TO_MEMBRANE	$8.0530 \cdot 10^{-2}$
GSM433777	CYTOPLASM	$1.7783 \cdot 10^{-1}$
	POSITIVE_REGULATION_OF_BIOLOGICAL_PROCESS	$1.9401 \cdot 10^{-1}$
	CELL_DEVELOPMENT	$2.2169 \cdot 10^{-1}$
	POSITIVE_REGULATION_OF_CELLULAR_PROCESS	$2.2169 \cdot 10^{-1}$
	NEGATIVE_REGULATION_OF_BIOLOGICAL_PROCESS	$2.2169 \cdot 10^{-1}$
	PLASMA_MEMBRANE	$2.2657 \cdot 10^{-1}$
	MEMBRANE	$2.2909 \cdot 10^{-1}$
	MACROMOLECULAR_COMPLEX	$2.3689 \cdot 10^{-1}$
	PROTEIN_COMPLEX	$2.3689 \cdot 10^{-1}$
	NEGATIVE_REGULATION_OF_CELLULAR_PROCESS	$2.5309 \cdot 10^{-1}$
GSM433778	ESTABLISHMENT_OF_LOCALIZATION	$3.4466 \cdot 10^{-1}$
	CELL_DEVELOPMENT	$3.6827 \cdot 10^{-1}$
	NEGATIVE_REGULATION_OF_BIOLOGICAL_PROCESS	$3.6827 \cdot 10^{-1}$
	MACROMOLECULAR_COMPLEX	$3.8483 \cdot 10^{-1}$
	PROTEIN_COMPLEX	$3.8483 \cdot 10^{-1}$
	TRANSPORT	$3.8483 \cdot 10^{-1}$
	NEGATIVE_REGULATION_OF_CELLULAR_PROCESS	$4.0208 \cdot 10^{-1}$
	BIOPOLYMER_MODIFICATION	$4.2931 \cdot 10^{-1}$
	REGULATION_OF_DEVELOPMENTAL_PROCESS	$4.3876 \cdot 10^{-1}$
	PROTEIN_MODIFICATION_PROCESS	$4.3876 \cdot 10^{-1}$
GSM433779	NUCLEUS	$1.6135 \cdot 10^{-2}$
	NUCLEAR_PART	$3.2700 \cdot 10^{-2}$
	ORGANELLE_LUMEN	$7.2480 \cdot 10^{-2}$
	ENVELOPE	$7.2480 \cdot 10^{-2}$
	ORGANELLE_ENVELOPE	$7.2480 \cdot 10^{-2}$
	MEMBRANE_ENCLOSED_LUMEN	$7.2480 \cdot 10^{-2}$
	MITOCHONDRION	$8.6676 \cdot 10^{-2}$
	MITOCHONDRIAL_PART	$1.2294 \cdot 10^{-1}$
	NUCLEAR_LUMEN	$1.2294 \cdot 10^{-1}$
	NUCLEOPLASM	$1.6001 \cdot 10^{-1}$
GSM433780	EXTRACELLULAR_REGION	$1.0572 \cdot 10^{-1}$
	MEMBRANE	$1.1636 \cdot 10^{-1}$
	PLASMA_MEMBRANE	$1.3267 \cdot 10^{-1}$
	INTRINSIC_TO_MEMBRANE	$1.4670 \cdot 10^{-1}$
	INTEGRAL_TO_MEMBRANE	$1.4670 \cdot 10^{-1}$
	EXTRACELLULAR_REGION_PART	$1.8887 \cdot 10^{-1}$
	MEMBRANE_PART	$2.0860 \cdot 10^{-1}$
	PLASMA_MEMBRANE_PART	$2.2309 \cdot 10^{-1}$
	MULTICELLULAR_ORGANISMAL_DEVELOPMENT	$2.3013 \cdot 10^{-1}$
	RECEPTOR_ACTIVITY	$2.3163 \cdot 10^{-1}$
GSM433781	EXTRACELLULAR_REGION	$1.1547 \cdot 10^{-1}$
	MEMBRANE	$1.3395 \cdot 10^{-1}$
	PLASMA_MEMBRANE	$1.4803 \cdot 10^{-1}$
	POSITIVE_REGULATION_OF_CELLULAR_PROCESS	$1.6546 \cdot 10^{-1}$

continued on next page

tissue	GO term	p value
	NEGATIVE_REGULATION_OF_BIOLOGICAL_PROCESS	$1.6546 \cdot 10^{-1}$
	MACROMOLECULAR_COMPLEX	$1.7910 \cdot 10^{-1}$
	PROTEIN_COMPLEX	$1.7910 \cdot 10^{-1}$
	NEGATIVE_REGULATION_OF_CELLULAR_PROCESS	$1.9383 \cdot 10^{-1}$
	EXTRACELLULAR_REGION_PART	$2.0162 \cdot 10^{-1}$
	REGULATION_OF_DEVELOPMENTAL_PROCESS	$2.2684 \cdot 10^{-1}$
GSM433782	PROTEIN_METABOLIC_PROCESS	$1.1930 \cdot 10^{-1}$
	MACROMOLECULAR_COMPLEX	$1.3518 \cdot 10^{-1}$
	PROTEIN_COMPLEX	$1.3518 \cdot 10^{-1}$
	CELLULAR_MACROMOLECULE_METABOLIC_PROCESS	$1.4914 \cdot 10^{-1}$
	CELLULAR_PROTEIN_METABOLIC_PROCESS	$1.6043 \cdot 10^{-1}$
	BIOPOLYMER_MODIFICATION	$1.7002 \cdot 10^{-1}$
	PROTEIN_MODIFICATION_PROCESS	$1.7796 \cdot 10^{-1}$
	BIOPOLYMER_METABOLIC_PROCESS	$2.1323 \cdot 10^{-1}$
	BIOSYNTHETIC_PROCESS	$2.3367 \cdot 10^{-1}$
	PLASMA_MEMBRANE	$2.5923 \cdot 10^{-1}$
GSM433783	PROTEIN_DIMERIZATION_ACTIVITY	$9.9892 \cdot 10^{-2}$
	ENDOPLASMIC_RETICULUM	$1.1014 \cdot 10^{-1}$
	ION_BINDING	$1.3386 \cdot 10^{-1}$
	REGULATION_OF_MOLECULAR_FUNCTION	$1.4753 \cdot 10^{-1}$
	INTRINSIC_TO_PLASMA_MEMBRANE	$1.6695 \cdot 10^{-1}$
	INTEGRAL_TO_PLASMA_MEMBRANE	$1.6695 \cdot 10^{-1}$
	SYSTEM_DEVELOPMENT	$1.6695 \cdot 10^{-1}$
	CATION_BINDING	$1.7912 \cdot 10^{-1}$
	INTRINSIC_TO_MEMBRANE	$1.9492 \cdot 10^{-1}$
	INTEGRAL_TO_MEMBRANE	$1.9492 \cdot 10^{-1}$
GSM433784	MEMBRANE	$1.6535 \cdot 10^{-2}$
	MEMBRANE_PART	$3.3632 \cdot 10^{-2}$
	PLASMA_MEMBRANE	$6.2403 \cdot 10^{-2}$
	INTRINSIC_TO_MEMBRANE	$6.9048 \cdot 10^{-2}$
	INTEGRAL_TO_MEMBRANE	$6.9048 \cdot 10^{-2}$
	PROTEIN_METABOLIC_PROCESS	$7.6361 \cdot 10^{-2}$
	CELLULAR_MACROMOLECULE_METABOLIC_PROCESS	$9.3255 \cdot 10^{-2}$
	CELLULAR_PROTEIN_METABOLIC_PROCESS	$9.9635 \cdot 10^{-2}$
	PLASMA_MEMBRANE_PART	$1.0642 \cdot 10^{-1}$
	INTRINSIC_TO_PLASMA_MEMBRANE	$1.4753 \cdot 10^{-1}$

Table 6: Top 10 KEGG pathways as produced by Fisher exact test for gene set enrichment analysis; the KEGG term KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC has been shortened to KEGG_ARVC to fit the page

tissue	KEGG pathway	p value
GSM433776	KEGG_WNT_SIGNALING_PATHWAY	$2.7829 \cdot 10^{-1}$
	KEGG_CALCIUM_SIGNALING_PATHWAY	$3.0733 \cdot 10^{-1}$
	KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	$3.0733 \cdot 10^{-1}$
	KEGG_LEISHMANIA_INFECTION	$3.0733 \cdot 10^{-1}$
	KEGG_COLORECTAL_CANCER	$3.3938 \cdot 10^{-1}$
	KEGG_MAPK_SIGNALING_PATHWAY	$3.5171 \cdot 10^{-1}$
	KEGG_LYSOSOME	$3.5171 \cdot 10^{-1}$
	KEGG_VIRAL_MYOCARDITIS	$3.7469 \cdot 10^{-1}$
	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	$4.1362 \cdot 10^{-1}$
	KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	$4.1362 \cdot 10^{-1}$
GSM433777	KEGG_PATHWAYS_IN_CANCER	$2.9833 \cdot 10^{-1}$
	KEGG_MAPK_SIGNALING_PATHWAY	$4.8494 \cdot 10^{-1}$
	KEGG_LYSOSOME	$4.8494 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$5.8712 \cdot 10^{-1}$

continued on next page

tissue	KEGG pathway	p value
	KEGG_ALZHEIMERS_DISEASE	$5.8712 \cdot 10^{-1}$
	KEGG_AXON_GUIDANCE	$6.0603 \cdot 10^{-1}$
	KEGG_SMALL_CELL_LUNG_CANCER	$6.2552 \cdot 10^{-1}$
	KEGG_OXIDATIVE_PHOSPHORYLATION	$6.6631 \cdot 10^{-1}$
	KEGG_WNT_SIGNALING_PATHWAY	$6.6631 \cdot 10^{-1}$
	KEGG_MELANOGENESIS	$6.6631 \cdot 10^{-1}$
GSM433778	KEGG_FOCAL_ADHESION	$4.2931 \cdot 10^{-1}$
	KEGG_PATHWAYS_IN_CANCER	$4.4840 \cdot 10^{-1}$
	KEGG_ECM_RECEPTOR_INTER.	$5.8055 \cdot 10^{-1}$
	KEGG_MAPK_SIGNALING_PATHWAY	$6.1881 \cdot 10^{-1}$
	KEGG_LYSOSOME	$6.1881 \cdot 10^{-1}$
	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	$6.1881 \cdot 10^{-1}$
	KEGG_CELL_ADHESION_MOLECULES_CAMS	$6.3208 \cdot 10^{-1}$
	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	$6.7349 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$7.0247 \cdot 10^{-1}$
	KEGG_ALZHEIMERS_DISEASE	$7.0247 \cdot 10^{-1}$
GSM433779	KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	$4.2654 \cdot 10^{-2}$
	KEGG_ALZHEIMERS_DISEASE	$6.9883 \cdot 10^{-2}$
	KEGG_LYSOSOME	$7.2256 \cdot 10^{-2}$
	KEGG_OOCYTE_MEIOSIS	$1.6001 \cdot 10^{-1}$
	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	$1.6001 \cdot 10^{-1}$
	KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	$1.6001 \cdot 10^{-1}$
	KEGG_TYPE_I_DIABETES_MELLITUS	$1.6001 \cdot 10^{-1}$
	KEGG_GRAFT_VERSUS_HOST_DISEASE	$1.6001 \cdot 10^{-1}$
	KEGG_OXIDATIVE_PHOSPHORYLATION	$1.6211 \cdot 10^{-1}$
	KEGG_HEMATOPOIETIC_CELL_LINEAGE	$1.6211 \cdot 10^{-1}$
GSM433780	KEGG_ECM_RECEPTOR_INTER.	$3.6091 \cdot 10^{-1}$
	KEGG_CELL_ADHESION_MOLECULES_CAMS	$4.2329 \cdot 10^{-1}$
	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	$4.7674 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$5.1592 \cdot 10^{-1}$
	KEGG_ALZHEIMERS_DISEASE	$5.1592 \cdot 10^{-1}$
	KEGG_FOCAL_ADHESION	$5.4660 \cdot 10^{-1}$
	KEGG_ARVC	$5.8055 \cdot 10^{-1}$
	KEGG_OXIDATIVE_PHOSPHORYLATION	$6.0377 \cdot 10^{-1}$
	KEGG_HEMATOPOIETIC_CELL_LINEAGE	$6.0377 \cdot 10^{-1}$
	KEGG_PARKINSONS_DISEASE	$6.2789 \cdot 10^{-1}$
GSM433781	KEGG_PATHWAYS_IN_CANCER	$2.3590 \cdot 10^{-1}$
	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	$4.2139 \cdot 10^{-1}$
	KEGG_CELL_ADHESION_MOLECULES_CAMS	$4.3780 \cdot 10^{-1}$
	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	$4.9079 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$5.2948 \cdot 10^{-1}$
	KEGG_AXON_GUIDANCE	$5.4990 \cdot 10^{-1}$
	KEGG_FOCAL_ADHESION	$5.6710 \cdot 10^{-1}$
	KEGG_SMALL_CELL_LUNG_CANCER	$5.7109 \cdot 10^{-1}$
	KEGG_ARVC	$5.9306 \cdot 10^{-1}$
	KEGG_OXIDATIVE_PHOSPHORYLATION	$6.1583 \cdot 10^{-1}$
GSM433782	KEGG_PATHWAYS_IN_CANCER	$1.8626 \cdot 10^{-1}$
	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	$3.6584 \cdot 10^{-1}$
	KEGG_CELL_ADHESION_MOLECULES_CAMS	$3.8248 \cdot 10^{-1}$
	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	$4.3686 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$4.7719 \cdot 10^{-1}$
	KEGG_ALZHEIMERS_DISEASE	$4.7719 \cdot 10^{-1}$
	KEGG_FOCAL_ADHESION	$4.8761 \cdot 10^{-1}$
	KEGG_AXON_GUIDANCE	$4.9868 \cdot 10^{-1}$
	KEGG_SMALL_CELL_LUNG_CANCER	$5.2110 \cdot 10^{-1}$
	KEGG_ARVC	$5.4449 \cdot 10^{-1}$
GSM433783	KEGG_FOCAL_ADHESION	$2.0477 \cdot 10^{-2}$
	KEGG_ECM_RECEPTOR_INTER.	$8.2121 \cdot 10^{-2}$
	KEGG_LYSOSOME	$1.1014 \cdot 10^{-1}$

continued on next page

tissue	KEGG pathway	p value
	KEGG_PATHWAYS_IN_CANCER	$1.2492 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$1.9733 \cdot 10^{-1}$
	KEGG_AXON_GUIDANCE	$2.1736 \cdot 10^{-1}$
	KEGG_SMALL_CELL_LUNG_CANCER	$2.3938 \cdot 10^{-1}$
	KEGG_ARVC	$2.6359 \cdot 10^{-1}$
	KEGG_WNT_SIGNALING_PATHWAY	$2.9020 \cdot 10^{-1}$
	KEGG_MELANOGENESIS	$2.9020 \cdot 10^{-1}$
GSM433784	KEGG_PATHWAYS_IN_CANCER	$3.1630 \cdot 10^{-1}$
	KEGG_MAPK_SIGNALING_PATHWAY	$5.0220 \cdot 10^{-1}$
	KEGG_LYSOSOME	$5.0220 \cdot 10^{-1}$
	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	$5.6709 \cdot 10^{-1}$
	KEGG_ADHERENS_JUNCTION	$6.0243 \cdot 10^{-1}$
	KEGG_AXON_GUIDANCE	$6.2089 \cdot 10^{-1}$
	KEGG_SMALL_CELL_LUNG_CANCER	$6.3988 \cdot 10^{-1}$
	KEGG_FOCAL_ADHESION	$6.7449 \cdot 10^{-1}$
	KEGG_OXIDATIVE_PHOSPHORYLATION	$6.7951 \cdot 10^{-1}$
	KEGG_WNT_SIGNALING_PATHWAY	$6.7951 \cdot 10^{-1}$

Remaining 8 result tables from the GSEA analysis:

Table 7: GSEA enrichment scores and p-values for sample GSM433776 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.006185567
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.051229507
KEGG_ADHERENS_JUNCTION	-0.3143161	0.065891474
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.07707911
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.07100592
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.11394892
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.08383234
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.13404255
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.39849624
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.5968064
KEGG_AXON_GUIDANCE	-0.16156015	0.7637795
KEGG_LYSOSOME	0.22209951	0.17938931

Table 8: GSEA enrichment scores and p-values for sample GSM433777 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.0118811885
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.041198503
KEGG_ADHERENS_JUNCTION	-0.3143161	0.06626506
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.08659794
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.0776699
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.10453649
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.095918365
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.16082475
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.36382115
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.6046967
KEGG_AXON_GUIDANCE	-0.16156015	0.75456387
KEGG_LYSOSOME	0.22209951	0.25369978

Table 9: GSEA enrichment scores and p-values for sample GSM433778 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.0021367522
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.055666003
KEGG_ADHERENS_JUNCTION	-0.3143161	0.08829175
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.07114624
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.08471075
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.08383234
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.083333336
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.14897959
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.38342968
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.58882236
KEGG_AXON_GUIDANCE	-0.16156015	0.7386831
KEGG_LYSOSOME	0.22209951	0.20856611

Table 10: GSEA enrichment scores and p-values for sample GSM433779 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.0077220076
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.041036718
KEGG_ADHERENS_JUNCTION	-0.3143161	0.06990291
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.07495069
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.091796875
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.08167331
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.083333336
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.15891473
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.3601695
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.65742576
KEGG_AXON_GUIDANCE	-0.16156015	0.7345679
KEGG_LYSOSOME	0.22209951	0.22823985

Table 11: GSEA enrichment scores and p-values for sample GSM433780 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.00591716
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.053497944
KEGG_ADHERENS_JUNCTION	-0.3143161	0.079918034
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.0655106
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.08206107
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.08266129
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.105788425
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.16765286
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.36293435
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.6215686
KEGG_AXON_GUIDANCE	-0.16156015	0.7962578
KEGG_LYSOSOME	0.22209951	0.22393823

Table 12: GSEA enrichment scores and p-values for sample GSM433781 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.003968254
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.046421662
KEGG_ADHERENS_JUNCTION	-0.3143161	0.05168986
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.08159393
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.06972112
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.087128714
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.07942974
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.17382812
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.384
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.62765956
KEGG_AXON_GUIDANCE	-0.16156015	0.72633743
KEGG_LYSOSOME	0.22209951	0.22309197

Table 13: GSEA enrichment scores and p-values for sample GSM433782 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.003968254
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.044989776
KEGG_ADHERENS_JUNCTION	-0.3143161	0.052837573
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.06831119
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.09437751
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.09622642
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.094412334
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.16603054
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.366
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.62674654
KEGG_AXON_GUIDANCE	-0.16156015	0.72154474
KEGG_LYSOSOME	0.22209951	0.19685039

Table 14: GSEA enrichment scores and p-values for sample GSM433784 compared to the remaining tissues

NAME	ES	NOM p-val
KEGG_FOCAL_ADHESION	-0.27030912	0.014403292
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.28983307	0.04
KEGG_ADHERENS_JUNCTION	-0.3143161	0.06963249
KEGG_MAPK_SIGNALING_PATHWAY	-0.25829262	0.10261569
KEGG_SMALL_CELL_LUNG_CANCER	-0.31651652	0.09090909
KEGG_PATHWAYS_IN_CANCER	-0.20688379	0.0726257
KEGG_ALZHEIMERS_DISEASE	-0.29491496	0.09018036
KEGG_ECM_RECEPTOR_INTER.	-0.22219512	0.14345992
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTER.	-0.20512007	0.37425742
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.16316189	0.62249
KEGG_AXON_GUIDANCE	-0.16156015	0.7741273
KEGG_LYSOSOME	0.22209951	0.22269808