Bioinformatics and Biology Insights

CASE REPORT

Conceptual Aspects of Large Meta-Analyses with Publicly Available Microarray Data: A Case Study in Oncology

Markus Schmidberger, Sabine Lennert and Ulrich Mansmann

Division of Biometrics and Bioinformatics, IBE, University of Munich, 81377 Munich, Germany.
Corresponding author email: mansmann@ibe.med.uni-muenchen.de

Abstract: Large public repositories of microarray experiments offer an abundance of biological data. It is of interest to use and to combine the available material to create new biological information and to develop a broader view on biological phenomena.

Meta-analyses recombine similar information over a series of experiments to sketch scientific aspects which were not accessible by each of the single experiments. Meta-analysis of high-throughput experiments has to handle methodological as well as technical challenges. Methodological aspects concern the identification of homogeneous material which can be combined by appropriate statistical procedures. Technical challenges come from the data management of large amounts of high-dimensional data, long computation time, as well as the handling of the stored phenotype data.

This paper compares in a meta-analysis of a large series of microarray experiments the interaction structure within selected pathways between different tumour entities. The feasibility of such a study is explored and a technical as well as a statistical framework for its completion is presented. Multiple obstacles were met during completion of this project. They are mainly related to the quality of the available data and influence the biological interpretation of the results derived.

The sobering experience of our study asks for combined efforts to improve the data quality in public repositories of high-throughput data. The exploration of the available data in large meta-analyses is limited by incomplete documentation of essential aspects of experiments and studies, by technical deficiencies in the data stored, and by careless duplications of data.

Keywords: meta-analysis, oncology, r, public microarray data, gene graphs

Bioinformatics and Biology Insights 2011:5 13–39

doi: 10.4137/BBI.S5537

This article is available from http://www.la-press.com.

Introduction

Increasing insights into the pathogenesis of malignant disorders and the detection of a rapidly rising number of molecular alterations gave rise to the hope that cancer specific genetic profiles might be generated that will define biologic subgroups as well as define targets for direct specific therapeutic agents.[1] The search for genomic alterations has revealed a huge heterogeneity not only within one histologically defined cancer entity but even within one individual tumour.[2] The heterogeneity of genomic mutations, however, becomes less complex since their functional effects merge in the alteration of a few, distinct pathways, only.[3] Hence, the understanding of cancer biology may be improved also by focusing on alterations in pathway activities across tumour entities.

Rhodes et al[4] developed meta-analytic tools to characterize a common transcriptional profile that is universally activated in most cancer types relative to the normal tissues from which they arise, likely reflecting essential transcriptional features of neoplastic transformation. In addition, they characterized a transcriptional profile that is commonly activated in various types of undifferentiated cancer, suggesting common molecular mechanisms by which cancer cells progress and avoid differentiation.

It is the goal of this study to explore the feasibility of a large cross-cancer meta-analysis based on high-throughput gene expression microarray data (GEMA) to compare the interaction structure between members of specific pathways across relevant tumour entities based on available gene expression microarray data from oncological studies. The challenges of this project are given by the quality of available data, the data management for the projected study, the biostatistics/bioinformatics tools available for the analysis, and finally the strategy for interpreting the computational results.

Editorial policies and the idea to reuse the high-throughput gene expression data for validation and new research questions triggered the creation of public repositories. The MIAME (Minimum information about a microarray experiment) criteria[5] formulate the necessary conditions for verifying and reproducing results of microarray data analyses. MIAME compliance assures a sensible reuse of public microarray data for the study of new questions: biological properties of the samples and phenotypes that were assayed need to be recorded along the data obtained from these assays.

At the moment there are three recommended international repositories to archive publication related functional genomic data:[6,7] ArrayExpress (AE),[8] Gene Expression Omnibus (GEO),[9] and the Center for Information Biology Gene Expression Database (CIBEX).[10] GEO is currently the largest fully public gene expression resource.

Meta-analytic tools for GEMA are developed by many authors[11] but mainly in the field of differential gene expression and profiling. To our knowledge this paper is the first trying to do a meta-analysis of pathway specific network structures across tumour entities. The structural comparison is motivated by the discovery of different relationships between cancer types.[12] There is evidence for familial associations between acute myeloic leukemia and colorectal cancer.[13] Men with family history of breast cancer also have an increased risk of prostate cancer.[14] Different leukemia derive from specified deregulation during the hematopoietic stem cell differentiation.[15]

Therefore, the interaction structure within genes annotated to specific pathways is explored and compared between eight human cancer entities. The cancer entities are grouped in eight tumour groups: four solid tumours (breast, colon, prostate, lung) and four haemic tumours (ALL, AML, CLL, Lymphoma). Thirteen different KEGG pathways which are organized into three groups are studied: Basic cellular signalling pathways (KEGG ID 04110: Cell cycle, 04115: p53 signalling pathway, 04210: Apoptosis, 04310: Wnt signalling pathway, 04512: ECM-receptor interaction),

Table 1. Number of experiments and samples in GEO (published data) and AE database (27/02/2009).

Database	Experiments	Samples	Experiments without FLEO	First data
GEO	11298	286645	4362 (39%)	Jan 2001
AE	7637	224947	1599 (21%)	Okt 2003

Abbreviation: FLEO, feature-level extraction output.

disease specific pathways (05210: Colorectal cancer, 05215: Prostate cancer, 05221: Acute myeloic leukaemia, 05223: Non-small cell lung cancer), and pathways related to DNA repair (04150: mTOR signalling pathway, 03410: Base excision repair, 03420: Nucleotide excision repair, 03430: Mismatch repair).

The exploration of the communication structure within a large set of genes is feasible by ignoring the dynamics of the complex biological system. The available micorarray measurements represent time averages of transcription dynamics. Conditional interaction graphs are used to infer their conditional correlation structure.[16] An interpretation of the edges of these graphs will not be given. The interest consists in assessing evidence that these graphs are different with respect to edges between cancer entities.

The paper is an explorative study on a strategy how to combine publicly available data repositories, bioinformatic tools, and statistical concepts to the defined task. Therefore, an analysis pipeline for the intended problems is described. The adaption of data management and related tools to assemble the data, to check its quality, and to perform the low and high level analysis for a very large set of microarrays is demonstrated. The results are presented. The paper is organized as follows: Section 2 describes material and methods, presents the data as well as the tools used for the low- and high-level analyses. Section 3 contains the results on global differences in the conditional correlation structure of thirteen pathways in eight cancer entities. We discuss our experiences and results in Section 4.

Materials and Methods

Microarray data set

Due to the weekly imports from GEO to AE, the data is taken from AE in order to facilitate the data management process. The repositories are dominated by experiments with Affymetrix Microarray data of the 'HG-U133A' and 'HG-U133 Plus 2.0' chip platforms. In order to work with a sample with an uniform laboratory work-up, we concentrate on data from the 'HG-U133A' Affymetrix GeneChip. In order to avoid bias due to specific pre-processing of the raw data, the feature-level extraction output (FLEO) files (CEL files) are used.[17]

All experiments from AE repository available on February 27, 2009 and satisfying the following selection criteria are included: FLEO data available, more than 10 arrays have chip type HGU133A, experiment has more than 20 arrays, 50% of the arrays belong to one of the eight cancer entities. Some experiments satisfying these criteria contain identical arrays. For example the arrays from the experiments 'E-GEOD-3910' and 'E-GEOD-3911' together are identical to the arrays from the super series experiment 'E-GEOD-3912'. These experiments are not included in the study to avoid duplicate arrays. Thereby 23 experiments are excluded.

A large cancer data set with more than 7000 microarrays is built from about 60 public available experiments in the AE database. An overview of the selected experiments is available in the Appendix. A detailed statistic of the data set is shown in Table 2. Data from cell line experiments and from human patients are grouped together. Furthermore, cancer subtypes are combined to one cancer entity (eg, childhood ALL is included in the ALL cancer entity group).

The R language[18] and the Bioconductor project[19] are chosen as the computational environment. In order to handle several thousand of microarrays for the low- and high-level analyses of our

Table 2. Statistic of available arrays for selected ArrayExpress experiments grouped by the eight cancer entities.

	Experiments	Arrays	HG-U133A	Deficient	Used
BREAST	20	3595	2454 (68%)	40 (1%)	1834 (51%)
ALL	12	1190	1140 (96%)	3 (0%)	916 (77%)
LUNG	7	537	398 (74%)	12 (2%)	386 (72%)
COLON	6	203	203 (100%)	6 (3%)	197 (97%)
PROSTATE	5	475	418 (88%)	2 (0%)	416 (88%)
AML	4	726	563 (78%)	29 (4%)	534 (74%)
LYMPHOMA	4	335	335 (100%)	4 (1%)	331 (99%)
CLL	3	194	182 (94%)	5 (3%)	177 (91%)
	61	7255	5693 (78%)	101 (1%)	4791 (66%)

data parallel computing is used.[20,21] A Bioconductor package called **affyPara**[22,23] implements parallel computing for pre-processing quality assessment of microarray data.

The tools 'boxplot' and 'MA-plot'[19] are used for quality assessment in the pre-processing step. If an array is deficient in both assessments, it is marked as 'deficient' and excluded. Sixty deficient arrays for solid cancer experiments and 41 deficient arrays for haemic cancer experiments are excluded, which is about 1% of the data. Due to duplicated arrays in different experiments (from one cancer entity) and several deficient arrays, the set of arrays used in the analysis is smaller than all available arrays. For the breast cancer experiments only 51% can be used in the analysis and about 66% of all arrays. Therefore, the meta-analysis is executed on 4791 microarrays: 2833 arrays for solid tumours and 1958 arrays for haemic disease tumours.

Phenotype data

No detailed data on the phenotypes of the patients included in the study was available due to lack of compliance with MIAME annotation rules. Even basic information like sex and age of the patients is not completely available. Detailed information on basic features of the tumours are generally missing. Our findings are summarized as follows: 47% of the patients are female and 18% male (36% missing), the median age is 55 (50% missing). All other variables, eg, tumour staging, are available for less than 20% of all arrays. Figure 1 shows the histogram of the age distribution for the haemic and solid cancer group. Since basic information on the tumours are not available the tumour entities may represent quite inhomogeneous groups.

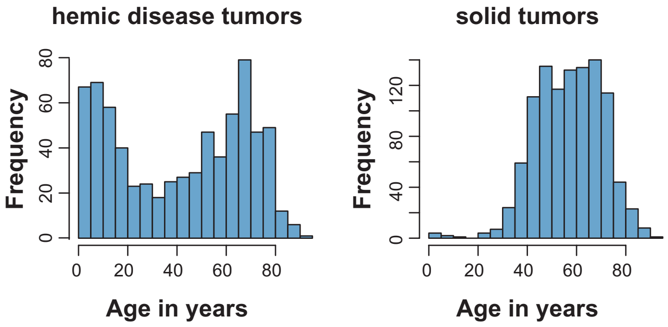

Figure 1. Histogram of the age distribution for the hemic and solid cancer group.

Data management

Sixty one single experiments and 7255 microarrays define the data set. Due to the large amount of CEL files and about 80 GB data volume, data management and storing is intricate. To make the data management feasible and reproducible, the raw data and processed data are saved in a general defined directory structure on the local hard disk. For every cancer entity, a directory containing the files is created. The file structure is optimized for the data processing with the R language and for re-usability of intermediate results.

The R package called **ArrayExpressDataManage** supports the data management of AE experiments at the local file system. It uses the Bioconductor package **ArrayExpress**[24] to download data from the AE database. Functions for different operations on the file structure are provided: Standard microarray processing steps (eg, rma parallel and serial preprocessing) as well as functions for data structure cleaning, creating overview tables.

The package creates automatically the data set generation script. Providing an R list structure object with the AE experiment IDs the complete data set can be regenerated from the AE data base. For the large cancer study the list object is available in the Appendix. Therefore, the data set of our analysis is not submitted as new super-series data set to one of the public repositories. It (raw data and phenodata) is already available in the AE database and can easily constructed by the analyst from the data set generation script. It is straightforward to add new experiments to the analysis. For more details see the vignette of the package or the help files of the package. The package is available at the R-forge repository: http://AEDataManage.R-forge.R-project.org/.

Low-level analysis

The data is pre-processed in one run using the R packages **ArrayExpressDataManage** and **affyPara**. After quality control, normalization is achieved by the Robust Multichip Average[25] [RMA] method. All analyses are parallelized and run on the 32 engine computer cluster at the IBE (LMU, Munich) offering a maximum of 128 processors. Each machine runs on four processors and eight GB main memory and they are connected with a 1 Gbit network. The complete RMA pre-processing of the 4791 HG-U133A CEL files took about 50 minutes computation time.

The data showed strong batch effects. Correction for batch effects[26,27] uses an empirical Bayes framework as proposed by Johnson et al.[27]

High-level analysis

The PC-Algorithm[28] is used to estimate the network structure (conditional correlation structure) within the set of genes annotated to each pathway for each cancer entities.

Estimating graphs

Multivariate gene expression data is characterised by its mean value structure as well as its dependence or correlation structure. While the first is concerned with the quantitative amount of transcription activity, the second focuses on the map of direct influences between genes: does the transcription activity of one gene influences the transcription activity of a second gene freezing all other genes annotated to the corresponding pathway on a fixed transcription level. An edge is drawn between two nodes (genes) if a direct influence is assessed. The PC-Algorithm estimates such a graph from observational data.[28,29]

Thorough validation studies[30] show advantages of the PC-Algorithm compared to competing approaches[31–35] especially for sparse graphs in terms of estimation quality (true and false discovery rates for edges) as well as computational speed.

The PC-Algorithm is run with $\alpha = 0.05$. This is a good choice for a graph with less then 20% of the maximal number of edges.[28] This is a plausible assumption for gene sets annotated to the KEGG pathways.

Comparing graphs

Graphs on the same set of nodes are compared by the Structural Hamming Distance (SHD). The SHD between two graphs is the number of edge insertions, deletions or flips in order to transform one graph to the other. The smaller the SHD the bigger is the similarity between the two graphs. The SHD is symmetric and can be calculated by SHD = # of different edges in both graphs—# common edges in both graphs.

The null-hypothesis of *no* structural difference between two tumour entities is tested by a permutation test. The test assesses if an observed SHD between two graphs is untypically large compared to the SHD distribution under the null-hypothesis. This distribution results from comparing two estimated graphs from two data sets which differ just by random fluctuations. The permutation test is carried out after standardizing the transcription values of genes annotated to the specific pathways. The mean value is substracted from the individual measurements and the difference is divided by the standard deviation in each set of the two cancer entities which are compared. The rejection of this null-hypothesis on a 5% significance level is considered as evidence the cell processes as captured by the specific set of pathway genes proves a differential dynamic between both tumour entities considered.

The resampling for the test procedure proceeds as follows:

- Choose the SHD to measure differential conditional correlation structure between both graphs.
- Estimate each graph by the PC-Algorithm with $\alpha = 0.05$ from the observed data and determine the *SH Dobs* between both graphs.
- For resampling step *i* permute the data units between both data sets, estimate both graphs and calculate the specific *SH Di* ($i = 1, \ldots, R$).
- Determine a permutation *P*-value by *pperm* = $\#\{SH\ Dobs < SH\ Di\}/R$.
- Reject the null-hypothesis if *pperm* is smaller then 0.05.

The data is resampled $R = 500$ times.

Permutation *P*-values below 0.01 are considered as evidence for a difference. Larger *P*-values are called marginal ($P \leq 0.05$) or not significant ($P > 0.05$).

Results

A total of 4791 microarrays was grouped into eight tumour entities (four solid tumours with a total of 1958 arrays and four haemic tumours with a total of 2833 arrays). The minimal sample sizes is 177 arrays for probes from CLL patients, the maximal sample size is 1834 arrays for breast cancer tissue (see Table 2). The phenotype information on the individual tumour probes is very sparse and is not considered in the following analysis.

Figure 2 shows the SHD for all six combinations of solid tumours (red triangles), all six combinations of haemic tumours (black triangles), and for all 16 haemic-solid combinations (blue triangles) when conditional independence graphs are estimated for each entity and compared by SHD.

There is no obvious evidence in any pathway that the SHD for a between group (haemic/solid) comparison is larger as the SHD for a within group (haemic/haemic or solid/solid) comparison.

The comparison within solid tumours can be summarized as follows. It holds that the breast-colon comparison (# of arrays: 1834/197) is only distinct for the Wnt signalling pathway (04310). The breast-lung comparison (# of arrays: 1834/386) results for most pathways in a pronounced difference except the AML pathway (05221) and the Mismatch repair pathway (03430). The breast-prostate comparison (# of arrays: 1834/416) shows marginal or non-significant differences for the p53 signalling pathway (04115), the ECM-receptor interaction pathway (04512), the AML

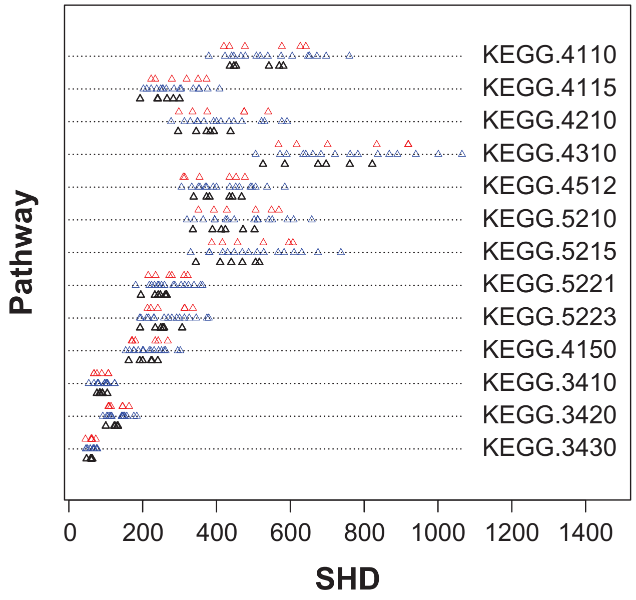

Figure 2. SHD in single pathways for comparisons within solid tumours (black), haemic tumours (red) and between group comparisons (blue).

Table 3. SHDs (permutation *P*-values) for different hemic and solid cancer entities and pathways.

Solid entities						
	BRE-COL	**BRE-LUN**	**BRE-PRO**	**COL-LUN**	**COL-PRO**	**LUN-PRO**
04110	577 (0.302)	642 (<0.002)	627 (<0.002)	435 (<0.002)	420 (<0.002)	477 (<0.002)
04115	319 (0.994)	373 (<0.002)	350 (0.046)	234 (<0.002)	223 (0.016)	279 (<0.002)
04210	475 (0.054)	540 (<0.002)	475 (<0.002)	335 (<0.002)	298 (0.02)	375 (<0.002)
04310	834 (<0.002)	919 (<0.002)	920 (<0.002)	617 (<0.002)	568 (<0.002)	701 (<0.002)
04512	435 (0.993)	477 (<0.002)	453 (0.614)	314 (0.686)	310 (0.592)	354 (0.802)
05210	506 (0.524)	569 (<0.002)	549 (<0.002)	393 (<0.002)	351 (<0.002)	428 (<0.002)
05215	527 (0.998)	607 (<0.002)	596 (0.002)	416 (<0.002)	387 (0.046)	457 (0.002)
05221	279 (0.993)	322 (0.044)	312 (0.84)	235 (0.136)	215 (0.374)	272 (0.008)
05223	314 (0.644)	336 (<0.002)	313 (0.162)	222 (0.038)	213 (0.028)	241 (0.022)
04150	234 (0.418)	242 (<0.002)	268 (<0.002)	172 (<0.002)	170 (<0.002)	180 (0.004)
03410	89 (0.744)	107 (<0.002)	107 (<0.002)	68 (<0.002)	68 (<0.002)	76 (<0.002)
03420	145 (0.35)	163 (<0.002)	146 (<0.002)	114 (<0.002)	109 (<0.002)	107 (<0.002)
03430	60 (0.993)	73 (0.054)	61 (0.918)	63 (0.006)	45 (0.366)	62 (0.188)
Hemic entities						
	ALL-AML	**ALL-CLL**	**ALL-LYM**	**AML-CLL**	**AML-LYM**	**CLL-LYM**
04110	581 (<0.002)	542 (<0.002)	570 (<0.002)	447 (<0.002)	453 (<0.002)	436 (<0.002)
04115	300 (<0.002)	283 (0.142)	266 (0.066)	241 (<0.002)	240 (<0.002)	193 (0.128)
04210	438 (<0.002)	373 (0.118)	393 (<0.002)	345 (<0.002)	383 (<0.002)	296 (0.004)
04310	822 (<0.002)	697 (0.05)	761 (<0.002)	585 (<0.002)	675 (<0.002)	526 (<0.002)
04512	468 (0.004)	435 (0.993)	443 (0.544)	373 (0.658)	381 (<0.002)	338 (0.52)
05210	503 (<0.002)	424 (0.644)	472 (<0.002)	389 (<0.002)	413 (<0.002)	336 (<0.002)
05215	516 (<0.002)	470 (0.994)	506 (0.006)	410 (0.23)	440 (<0.002)	344 (0.562)
05221	265 (0.002)	242 (0.994)	261 (0.008)	233 (0.386)	248 (<0.002)	195 (0.232)
05223	307 (<0.002)	255 (0.868)	258 (0.02)	234 (0.014)	249 (<0.002)	193 (0.038)
04150	241 (<0.002)	223 (0.06)	225 (<0.002)	192 (<0.002)	200 (<0.002)	162 (0.002)
03410	104 (<0.002)	85 (0.012)	91 (<0.002)	83 (<0.002)	91 (<0.002)	76 (<0.002)
03420	132 (<0.002)	123 (0.034)	131 (<0.002)	125 (<0.002)	133 (<0.002)	100 (<0.002)
03430	62 (0.002)	63 (0.758)	63 (<0.002)	61 (0.062)	59 (0.042)	48 (0.628)

pathway (05221), Non-small cell lung cancer pathway (05223), and the Mismatch repair pathway (03430). The colon-lung comparison (# of arrays: 197/386) shows marginal or non-significant differences for the ECM-receptor interaction pathway (04512), the AML pathway (05221), and the Non-small cell lung cancer pathway (05223). The colon-prostate comparison (# of arrays: 197/416) shows marginal or non-significant differences for the p53 signalling pathway (04115), Apoptosis (04210), the ECM receptor interaction pathway (04512), Prostate cancer pathway (05215), the AML pathway (05221), Non-small cell lung cancer pathway (05223), and the mismatch repair pathway (03430). The lung-prostate comparison (# of arrays: 386/416) shows marginal or non-significant differences for ECM-receptor interaction pathway (04512), Non-small cell lung cancer pathway (05223), and the Mismatch repair pathway (03430).

The comparison within haemic tumours can be summarized as follows. The ALL-AML comparison (# of arrays: 916/534) shows for each pathway a distinct conditional correlation structure. The ALL-CLL comparison (# of arrays: 916/177) shows marginal or non-significant differences for all pathway except Cell cycle (04110). The ALL-LYM (# of arrays: 916/331) comparison shows marginal or non-significant differences for p53 signalling (04115), ECM-receptor interaction (04512), Non-small cell lung cancer (05223). The AML-CLL comparison shows marginal or non-significant differences for the ECM receptor interaction (04512), Prostate cancer (05215), AML (05221), Non-small cell lung cancer (05223), and mismatch repair (03430). Comparing AML-LYM (# of arrays: 534/331) shows only the Mismatch repair pathway (03430) as marginal significant. The CLL-LYM comparison (# of arrays: 177/331) shows marginal or non-significant differences for p53 signalling (04115), ECM-receptor interaction (04512), Colon cancer (05210), Prostate cancer (05215), AML (05221), Non-small cell lung cancer (05223), and Mismatch repair (03430).

Table 6 in the Appendix presents the SHD and *P*-values for the between groups comparisons. They result in distinctive conditional correlation structures in all pathways for most of the pairs. More than two marginal or non-significant *P*-values are found in the COL-CLL, COL-LYM, LUN-ALL comparisons (see Table 4). No clear evidence for a difference in

Table 4. Number of pathways with *no* evidence for difference in conditional correlation structure.

Solid tumors						Heamic tumors					
BRE COL	BRE LUN	BRE PRO	COL LUN	COL PRO	LUN PRO	ALL AML	ALL CLL	ALL LYM	AML CLL	AML LYM	CLL LYM
12	2	5	3	7	3	3	12	3	5	1	6

Mixed comparison															
BRE ALL	BRE AML	BRE CLL	BRE LYM	COL ALL	COL AML	COL CLL	COL LYM	LUN ALL	LUN AML	LUN CLL	LUN LYM	PRO ALL	PRO AML	PRO CLL	PRO LYM
0	0	0	0	1	2	10	7	7	0	0	0	2	0	2	0

the COL-CLL comparison is found for p53 signalling (04115), Apoptosis (04210), ECM-receptor interaction (04512), Prostate cancer (05215), AML (05221), Non-small cell lung cancer (05223), mTOR signalling (04150), Base excision repair (03410), Nucleotide excision repair (03420), and Mismatch repair (03430) pathway. No clear evidence for a difference in the COL-LYM comparison is found for p53 signalling (04115), ECM-receptor interaction (04512), Prostate cancer (05215), AML (05221), Non-small cell lung cancer (05223), mTOR signalling (04150), and Mismatch repair (03430). Finally, no clear evidence for a difference in the COL-CLL comparison is found for p53 signalling (04115), Apoptosis (04210), ECM-receptor interaction (04512), Prostate cancer (05215), AML (05221), Non-small cell lung cancer (05223), mTOR signalling (04150), Base excision repair (03410), Nucleotide excision repair (03420), and mismatch repair (03430) pathway.

We use the number of pathways with no evidence for differential conditional correlation structure as a measure for similarity between tumor entities. Table 4 and Figure 3 summarize the situation. Table 5 lists the number of comparisons between and within groups with a permutation *P*-value above 0.1. The highest ranked pathways with respect to no evidence for a difference are Mismatch repair (03430), Non-small cell lung cancer (05223), AML (05221), ECM-receptor interaction (04512), and p53 signalling (04115) pathway. The pathways Cell cycle (04110) and Wnt signalling (04310) show in all except one comparison a significant difference in conditional correlation structure.

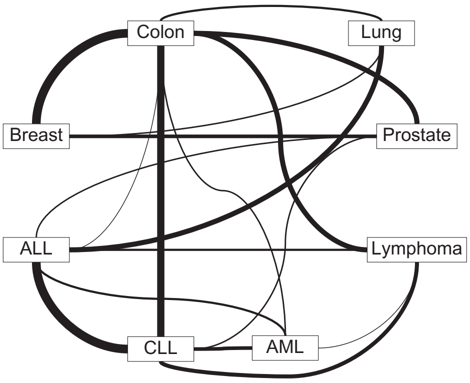

Figure 3. Similarity between tumours in terms of pathways with *no* evidence for a difference in conditional correlation structure.

The major similarity between the entity pairs are visualized in Figure 3. Every node represents an entity and the wide of the edges is the number of pathways with no evidence for a difference. Similar pathway structure can be found between ALL-CLL, BREAST-COLON, and COLON-CLL. In the haemic tumour entities there is a noticeable similarity between lymphatic tumours (ALL, CLL, Lymphoma). In the solid tumour entities there is a similarity between tumours (breast, colon, prostate) arising in gland tissues.

Discussion

Similarities as well as dissimilarities between cellular processes in different tumour entities are of general interest. They promise a better understanding of basic disease mechanism as well as therapeutic principles. To this end many studies are concerned with the comparison of transcription profiles and gene signatures.[4] Bioinformatic tools give easy access to a wide range of signatures[36] and help to validate them over a wide range of disease entities. For example the molecular program of normal wound healing plays an important role in cancer metastasis. Consistent features in the transcriptional response of normal fibroblasts to serum reveal links between wound healing and cancer progression in a variety of common epithelial tumours.[37]

Our studies does not focus on transcription profiles but on the structure of wiring between genes annotated to specific pathways and if these structures differ between tumor entities. We formalize the wiring between genes by conditional correlation graphs[28] where the genes of interest define the nodes and a direct influence between two genes is represented by an edge.

The following ideas motivated the study:

- Tumours of different tissues may have distinct features in the way the corresponding transcription of genes annotated to the pathways is regulated. Therefore, we compared haemic and solid tumours. Within both groups we defined two subgroups. Solid tumours were split in a group taken from gland (breast, colon, and prostate cancer) and lung tissue. Haemic tumours were split in lymphatic tumours (ALL, CLL, and Lympoma) and AML.
- To detect similar regulation structures in major pathways between cancers, we studied generic pathways which are crucial for the cell machinery

Table 5. Number of comparisons with *no* evidence (p *ge* 0.1) for a difference in conditional correlation structure (per pathway).

Pathway KEGG ID	Total 28 comparisons	Solid tumors 6 comparisons	Haemic tumors 6 comparisons	Mixed tumors 16 comparisons
3410	3	1	1	1
3420	4	1	1	2
3430	15	5	4	6
4110	1	1	0	0
4115	10	3	3	4
4150	5	1	1	3
4210	5	2	1	2
4310	1	0	1	0
4512	11	5	4	2
5210	2	1	1	0
5215	5	2	3	2
5221	13	5	3	5
5223	13	5	4	4

(Cell cycle, Apoptosis, p53 signalling, ECM-receptor, Wnt signalling), which are disease related (Colorectal cancer, Prostate cancer, Non-small cell lung cancer, AML), and finally pathways which concern DNA repair.

The study is designed as a meta-analysis of data available in public repositories. Uses microarrays from one specific technical platform (HG-U133A) to avoid unnecessary heterogeneity caused by differing technical and lab-specific conditions. The data collection was the most challenging part of the meta-analysis. Our activity focused on the ArrayExpress repository. Here we found 61 studies with a total of 5693 arrays (HG-U133A) arrays which contributed material to the tumours of interest. A total of 101 defect arrays and further 801 duplicates of microarrays were removed. The package **ArrayExpressDataManage** was developed to organise the data management of the 4791 remaining microarrays for the meta-analysis. The package allows an efficient and reproducible retrieval of large data sets from the ArrayExpress repository.

The quality of the downloaded data in terms of phenotype information was very low. It is astonishing that other authors do not report this fact. Phenotype information is generally missing and even incomplete in basic items to characterize the clinical staging of tumours. Our data was taken from clinical studies and only 64% of the studies provided the sex of the patients. Age was missing in 50% of the patients. Detailed information on the tumour (tumour grade was available in 20% of the solid cancer, only sparse information on metastatic disease) was not given.

All studies contributed to ArrayExpress declare formally compliance with the MIAME criteria.[38] MIAME also requests information on the biological properties of the samples and phenotypes that were assayed. Neglecting this information may produce an insensible mixture of probes and invalidates most of the public oncological microarray data for secondary research.

The missing phenotype information confronts us with a potentially inhomogeneous set and a biological mix of probes within a tumour entity which are in different stages of its development. This confounding may invalidate the interpretation of our results.

Furthermore, the observed microarray data quantifies time averages of a complex dynamics with many components. Additionally, the gene expression measured for a few selected genes annotated to a pathway is only a small observation window on the system. The unobserved components also may confound the conditional correlation between two observed genes.

All these restriction impose a severe restrictions on interpretability and only allow a very coarse conclusion if a difference between two conditional correlation graphs is assessed: The dynamics are somehow distinct. Techniques to locate differences between two graphs more precisely to specific subset of nodes of graphs are under study and applied in settings with a stricter control of the phenotype and the confounding.[16]

The estimation of conditional correlation structures was performed with the PC-Algorithm.[28] The difference between the estimated structures was quantified by the Structural Hamming Distance. Statistical significance was explored by resampling techniques. The computational challenge were handled by a parallelized computation environment created out of standard open access tools from R and the Bioconductor.[18,19] Additional software which was required to build up our pipeline for reproducible calculations is readily provided.

The quantitative results for the conditional correlation structures of the gene sets and tumour quantities studied can be summarized as follows: The inferred structures for the selected thirteen pathways look mostly similar between breast cancer and colon cancer as well as between ALL and CLL samples. The lag of strong differences in the structures between colon cancer and CLL samples never addressed to our knowledge in the literature. There are narrative reports on patients with two different primary tumours where the treatment directed two one of both only resulted in a response of the second and not the first (in our centre a patient with simultaneous colon cancer and AML). The pathways with the most similar conditional correlation of the annotated genes between tumour entities are Mismatch repair (03430), Non-small cell lung cancer (05223), and AML (05221). The wiring between genes annotated to the Wnt pathway (04512) seems to be more similar within as between the tumour entities (solid/haemic).

The study answers both heuristics which motivated the meta-analysis in a limited way. It defined the technical requirements to perform a meta-analysis with about 4800 microarrays. Also, the statistical methods are available which help to tackle the question posed. But, it was not possible to enrich the developed analysis pipeline with necessary biological and clinical data. The publicly available data generally lack relevant and important phenotype information. The sobering experience of our study asks for combined efforts to improve the data quality in public repositories for data from high-throughput technologies.

The dataset used for the presented results is from February 2009. Since then, only 7 new data sets became available which meet our inclusion/exclusion criteria: six for breast cancer and 1 for prostate cancer, all together about 550 arrays. Furthermore, it would have been possible to validate our findings on microarray data derived from microarray type HGU-133 Plus 2.0. We did not perform the validation since we are not able to assure the homogeneity between the populations used for the first step and the validation. In this case we would not be able to discuss possible deviations between the validation based on the HGU-133 Plus 2.0 and the calculation based on the HGU-133 arrays. The HGU-133 Plus 2.0 arrays are more recent as the HGU-133 arrays which may imply that the studies based on the HGU-133 Plus 2.0 array are performed with more specific questions on more specific populations.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Thomas J Hudson, Warwick Anderson, Axel Artez, et al. International Cancer Genome Consortium. International network of cancer genome projects. *Nature.* 2010 Apr;464(7291):993–8.
2. Baisse B, Bouzourene H, Saraga EP, Bosman FT, Ben-Hattar J. Intratumor genetic heterogeneity in advanced human colorectal adenocarcinoma. *Int J Cancer.* 2001 Aug;93(3):346–52.
3. Thomas Brabletz, Andreas Jung, Simone Spaderna, Falk Hlubek, Thomas Kirchner. Opinion: migrating cancer stem cells—an integrated concept of malignant tumour progression. *Nat Rev Cancer.* 2005 Sep;5(9):744–9.
4. Daniel R Rhodes, Jianjun Yu, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A.* 2004 Jun;101(25):9309–14.
5. Alvis Brazma. Minimum information about a microarray experiment (MIAME)—towards standards for microarray data. *Nature Genetics.* 2001; 29:365–71.
6. Catherine A Ball, Alvis Brazma, Helen Causton, et al. Submission of microarray data to public repositories. *PLoS Biology.* 2004 Aug;2(9):e317.
7. Gardiner-Garden M, Littlejohn TG. A comparison of microarray databases. *Brief Bioinform.* 2001 May;2(2):143–58.
8. Brazma Alvis, Parkinson Helen, Sarkans Ugis, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research.* 2003 Jan;31(1):68–71.
9. Barrett Tanya, Troup Dennis B, Wilhite Stephen E, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research.* 2009 Jan;37(Database issue):D885–90.
10. Kazuho Ikeo, Jun Ishi-i, Takurou Tamura, Takashi Gojobori, Yoshio Tateno. CIBEX: center for information biology gene expression database. *C R Biol.* 2003;326(10–11):1079–82.
11. Erin M Conlon. A bayesian mixture model for meta-analysis of microarray studies. *Funct Integr Genomics.* 2008 Feb;8(1):43–53.
12. Landgren O, Pfeiffer RM, Stewart L, et al. Risk of second malignant neoplasms among lymphoma patients with a family history of cancer. *International Journal of Cancer.* 2007;1099–1102(5):8–14.

13. Lynch H, Family with acute myelocytic leukemia, breast, ovarian, and gastrointestinal cancer. *Cancer Genetics and Cytogenetics*. 2009;137(1): 8–14.
14. Brandt A, Bermejo JL, Sundquist J, Hemminki K. Risk of second malignant neoplasms among lymphoma patients with a family history of cancer. *European Journal of Cancer*. 2009. [Epub ahead of print].
15. Arinobu Y, Mizuno SI, Chong Y, et al. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myelo-erythroid and myelolymphoid lineages. *Cell Stem Cell.* 2007;1(4): 416–27.
16. Ulrich Mansmann, Markus Schmidberger, Ralf Strobl, Vindi Jurinovic. *Statistical modelling and regression structures—festschrift in honour of Ludwig Fahrmeir,* chapter Indirect Comparison of Interaction Graphs. Physica, 2009:249–65.
17. Adaikalavan Ramasamy, Adrian Mondry, Chris C Holmes, Douglas G Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine.* 2008 Sep;5(9):e184.
18. R development core team. *R: A language and environment for statistical computing.* Vienna, Austria: R foundation for statistical computing; 2009. ISBN 3-900051-07-0.
19. Robert C Gentleman, Vincent J Carey, Douglas M Bates, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology.* 2004:5.
20. Markus Schmidberger, Martin Morgan, Dirk Eddelbuettel, Hao Yu, Luke Tierney, Ulrich Mansmann. State of the art in parallel computing with R. *Journal of Statistical Software.* 2009;31(1).
21. Markus Schmidberger. *Parallel Computing for Biological Data.* PhD thesis, 2009 Nov.
22. Markus Schmidberger, Ulrich Mansmann. Parallelized pre-processing algorithms for high-density oligonucleotide arrays. In *Proceedings IEEE International Symposium on Parallel and Distributed Processing IPDPS 2008.* 2008 Apr 14–18:1–7.
23. Markus Schmidberger, Ulrich Mansmann. affyPara—a bio-conductor package for parallelized preprocessing algorithms of affymetrix microarray data. *Bioinformatics and Biology Insights.* 2009;3:83–7.
24. Audrey Kauffmann, Tim F Rayner, Helen Parkinson, et al. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics.* 2009;25(16): 2092–4.
25. Robert C Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, 1st ed. Springer; 2005 Aug.
26. Rafael A Irizarry, Daniel Warren, Forrest Spencer, et al. Multiple-laboratory comparison of microarray platforms. *Nature Methods.* 2005 May;2(5): 345–50.
27. Evan Johnson W, Cheng Li, Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics.* 2007 Jan;8(1):118–27.
28. Markus Kalisch, Peter Buehlmann. Estimating high dimensional acyclic graphs with the PC-Algorithm. *Journal of Machine Learning Research.* 2007;8:613–36.
29. Peter Spirtes, Clark Glymour, Richard Scheines. *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: The MIT Press; 2001 Jan.
30. Fanny Villers, Brigitte Schaeffer, Caroline Bertin, Sylvie Huet. Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. *Statistical Applications in Genetics and Molecular Biology.* 2008;7(1):Article 14.
31. Juliane Schaefer, Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology.* 2005;4: Article 32.
32. Nicolai Meinshausen, Peter Buehlmann. High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics.* 2006;34:1436–62.
33. Martin J Wainwright, Pradeep Ravikumar, John D Lafferty. High dimensional graphical model selection using L1-regularized logistic regression. *Proceedings of Advances in Neural Information Processing Systems.* 2006; 9:1465–72.
34. Jerome Friedman, Trevor Hastie, Robert Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics.* 2007;9(3):432–41.
35. Banerjee Onureena, El Ghaoui Laurent, d'Aspremont Alexandre. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research.* 2008: 485–516.
36. Aedn C Culhane, Thomas Schwarzl, Razvan Sultana, et al. Genesigdb—a curated database of gene expression signatures. *Nucleic Acids Res.* 2010 Jan; 38(Database issue):D716–25.
37. Howard Y Chang, Dimitry SA Nuyten, Julie B Sneddon, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A.* 2005 Mar;102(10):3738–43.
38. Alvis Brazma. Minimum information about a microarray experiment (MIAME)—successes, failures, challenges. *Scientific World Journal.* 9:420–3.
39. Arnold Ludwig. *Random Dynamical Systems*, 2nd ed. Berlin, Germany: Springer-Verlag; 1998 Jan.

Appendix

Methodological considerations for the interpretation of our results

A short outline of ideas is given on how to relate biological ideas to the results of the described microarray analyses.

There are two problems: measuring time averages on a narrow set of dynamic processes and confounding of interaction within genes annotated to a pathway by unobserved quantities.

In order to understand the problem of time average, the cell is considered as a large dynamic system which is in a stable state. The multivariate process $(X_t)_{t\in[0,T]}$ describes the dynamic of all entities within the cell over a time period of length T. The data read from a microarray is the average over many cells caught at different time points within the dynamics of the cell-specific system. If the dynamic system is ergodic, its behaviour when it is allowed to run for a long time can be read from the cross-sectional measurements of many simultaneously trajectories at a fixed time point. This is expressed through ergodic theorems which assert that, under certain conditions, the time average of a function along the trajectories exists almost everywhere and is related to the space average.[39]

Therefore, if a large amount of cells is assayed we get an average of all measurements from each cell which is identical of the average over time of $(X_t)_{t\in[0,T]}$:

$$W = \frac{1}{T}\int X_t \ dt$$

The covariance matrix of the high-dimensional vector W consists of integrals of covariance (in the diagonal) as well as cross-covariance functions (off diagonal). The term cross-covariance refers to the covariance $cov(X,Y)$ between two random vectors X and Y. In order to distinguish that concept from the "covariance" of a random vector X, which is understood to be the matrix of covariance between the scalar components of X.

If W is observed in the tissue of several individuals, it is possible to estimate the conditional correlation structure of W by a conditional correlation graph. If material from a different biological condition exists, one could estimate a second graph and compare both. In case of a difference between both graphs the conclusion is allowed that the corresponding covariance matrices are different and that also some time averages of cross-covariance functions between components of the cell dynamics are different. This would allow the vague statement that the dynamic system under one condition is different from the dynamic system which governs the alternative condition. A deeper insight given such data is not possible.

The next problem is that we do not see the complex dynamic system. Transcription measurements are only available for a restricted set of genes (defined by the pathway). We do not see the whole of W but only a small subset of it. Let U be the components of W which are observable from the data and V be the components of W which are not measured (protein concentration, gene expression of genes which are not part of our pathway, …).

Often a random dynamical system is considered as a complex Gaussian process and the framework of multivariate normal distributions is available to do a thorough formal analysis. Thereby, the precision matrix of W which is the inverse of the covariance matrix plays the central role to understand direct interactions between the components.

Using this approach allows to understand how the unobserved components influence the correlation structure of the observed data by confounding. Confounding means that the conditional correlation between two genes is also the effect of unobserved components in the system and possibly not a real biological feature that is shared by both genes. This makes it difficult to interpret a conditional correlation graph of genes annotated to a pathway. It may show effects caused from outside and not by biological activity within the pathway.

The precision matrix of W is given by Q and can be partitioned in parts Q_{UU} and Q_{VV} which describe the conditional correlation structure in the observed part (U) and the unobserved part (V) of the system. The parts Q_{UV} and Q_{VU} ($= Q_{UV}^T$) describe the conditional correlation structure between the observed and unobserved components of the system.

$$Q = \begin{pmatrix} Q_{UU} & Q_{UV} \\ Q_{VU} & Q_{VV} \end{pmatrix}$$

The precision matrix of the marginal distribution of the observed components U is given by

$$Q_{UU}^{m\arg} = Q_{UU} + Q_{UV} \cdot Q_{VV}^{-1} Q_{VU}$$

This formalizes the idea that in the worst case an observed conditional correlation between two pathway genes is not caused by activity within the pathway but transmitted by conditional correlation of the genes with components of the system which are not observed. The following example demonstrates a practical consequence of the formal consideration.

A transcription factor regulates the expression of gene $G1$. Transcription factors belong to the unobserved V components of W. The concentration of a transcription factor may be regulated by some other protein which is also an unobserved V component of W. This protein is regulated by the transcriptional products of gene $G2$. The conditional correlation structure of both proteins is an element of Q_{VV} while the interaction of the transcription factor with G1 and the protein regulation by gene $G2$ are represented by elements of Q_{UV}. This may imply a non-zero element in $Q_{UU}^{m\arg}$ without the need for direct interaction of $G1$ and $G2$ within the pathway.

Result of permutation test between cancer groups

Table 6 presents the SHD and P-values for the between groups comparisons. The results in the solid and haemic tumors are available in Table 3 in the paper.

Data set generation script

The data set was created using the new developed **ArrayExpressDataManage** package (available at http://AEDataManage.R-forge.R-project.org/). The following commands (and experiments) were used to generate the locale data set structure:

```
R> library(ArrayExpressDataManage)
R> # Solid tumors
R> path_solid <- createDataStruct(path=‘/home/cancdat’,
+  data=list(
+   breast=c(‘E-GEOD-6532’, ‘E-GEOD-4922’, ‘E-GEOD-1456’,
+    ‘E-GEOD-11121’, ‘E-GEOD-7390’, ‘E-GEOD-12093’, ‘E-GEOD-2603’,
+    ‘E-GEOD-5462’, ‘E-GEOD-9936’, ‘E-GEOD-5847’, ‘E-MTAB-7’,
+    ‘E-GEOD-1561’, ‘E-TABM-43’, ‘E-MEXP-440’, ‘E-GEOD-11965’,
+    ‘E-GEOD-6772’, ‘E-GEOD-9574’, ‘E-GEOD-4917’, ‘E-GEOD-3494’,
+    ‘E-GEOD-2990’),
+     #E-GEOD-6883 to small
+     #E-TABM-244 included in E-MTAB-7
+   prostate=c(‘E-GEOD-8218’, ‘E-TABM-26’, ‘E-TABM-90’, ‘E-MEXP-1327’,
+    ‘E-GEOD-2443’),
+    colon=c(‘E-MTAB-57’, ‘E-GEOD-4045’, ‘E-MEXP-383’, ‘E-MEXP-101’,
+    ‘E-GEOD-2742’, ‘E-MEXP-833’),
+    lung=c(‘E-GEOD-4824’, ‘E-GEOD-10072’, ‘E-GEOD-6253’, ‘E-GEOD-7670’,
+    ‘E-MEXP-231’, ‘E-TABM-15’, ‘E-GEOD-4127’)
+  ),
+  name=‘solid’)
R> #Hemic tumors
R> path_hemic <- createDataStruct(path=‘/home/cancdat’,
+   data=list(
+   cll=c(‘E-GEOD-11038’, ‘E-GEOD-8835’, ‘E-GEOD-6691’),
+   #E-GEOD-9992 included in E-GEOD-11038 (super series)
+   aml=c(‘E-GEOD-12417’,’E-GEOD-1159’,’E-GEOD-9476’, ‘E-GEOD-1729’),
+   #E-GEOD-8970 defect annotation file
+    all=c(‘E-GEOD-12995’,’E-GEOD-635’, ‘E-GEOD-10255’, ‘E-GEOD-2351’,
+    ‘E-GEOD-3912’, ‘E-MEXP-313’, ‘E-GEOD-14618’, ‘E-TABM-125’,
+    ‘E-GEOD-4698’, ‘E-GEOD-8879’, ‘E-MEXP-120’, ‘E-GEOD-1577’),
+    #E-GEOD-14613 included in E-GEOD-14618
+    #E-GEOD-3910 and E-GEOD-3911 included in
+    #E-GEOD-3912 (super series)
+    #E-GEOD-643-660 included in E-GEOD-635 (super series)
+    lymphoma=c(‘E-GEOD-4475’,’E-TABM-346’,’E-TABM-117’, ‘E-GEOD-8388’)
+    #E-GEOD-4176 is to small
+    ),
+   name=‘hemic’)
```

Legend for KEGG graphs

Figure 4 describes the arrows in the KEGG graphs.

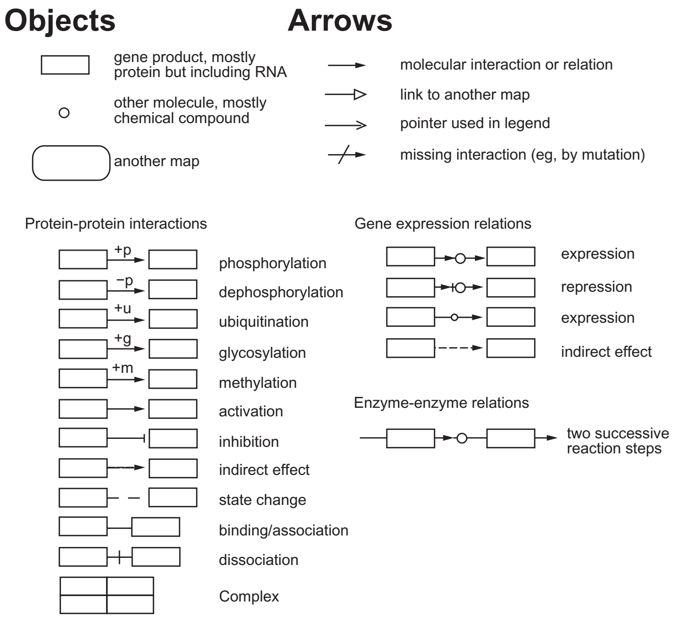

Figure 4. Legend for KEGG edges (http://www.genome.jp/kegg/document/symbols.gif).

ArrayExpress experiments selected for the meta-analysis

Tables 7–17 list all selected AE experiments used in the meta-analysis case study (grouped by cancer entities).

Table 6. SHDs (permutation P-values) for different hemic and solid cancer entities and pathways.

Mix entities 1

	BREAST- ALL	BREAST- AML	BREAST- CLL	BREAST- LYM	COLON- ALL	COLON- AML	COLON- CLL	COLON- LYM
04110	760 (<0.002)	697 (<0.002)	652 (<0.002)	672 (<0.002)	575 (<0.002)	478 (<0.002)	379 (0.004)	423 (<0.002)
04115	408 (<0.002)	376 (<0.002)	353 (<0.002)	352 (<0.002)	301 (<0.002)	249 (<0.002)	202 (0.016)	211 (0.532)
04210	591 (<0.002)	577 (<0.002)	522 (<0.002)	530 (<0.002)	412 (<0.002)	366 (<0.002)	277 (0.114)	329 (0.002)
04310	1065 (<0.002)	1001 (<0.002)	890 (<0.002)	940 (<0.002)	783 (<0.002)	661 (<0.002)	506 (<0.002)	590 (<0.002)
04512	585 (<0.002)	537 (<0.002)	496 (<0.002)	492 (<0.002)	452 (<0.002)	370 (<0.002)	305 (0.492)	333 (0.764)
05210	658 (<0.002)	609 (<0.002)	542 (<0.002)	592 (<0.002)	512 (<0.002)	429 (<0.002)	320 (0.094)	364 (<0.002)
05215	737 (<0.002)	675 (<0.002)	583 (<0.002)	631 (<0.002)	528 (<0.002)	448 (<0.002)	330 (0.99)	380 (0.472)
05221	362 (<0.002)	357 (<0.002)	324 (<0.002)	339 (<0.002)	283 (0.002)	244 (0.01)	181 (0.986)	224 (0.034)
05223	379 (<0.002)	374 (<0.002)	326 (<0.002)	345 (<0.002)	279 (<0.002)	258 (<0.002)	192 (0.632)	211 (0.054)
04150	295 (<0.002)	302 (<0.002)	258 (<0.002)	262 (<0.002)	231 (<0.002)	202 (<0.002)	154 (0.164)	164 (0.046)
03410	124 (<0.002)	124 (<0.002)	97 (<0.002)	105 (<0.002)	101 (<0.002)	97 (<0.002)	54 (0.038)	80 (<0.002)
03420	176 (<0.002)	184 (<0.002)	151 (<0.002)	157 (<0.002)	143 (<0.002)	147 (<0.002)	92 (0.068)	106 (<0.002)
03430	76 (<0.002)	76 (0.004)	75 (<0.002)	69 (0.006)	66 (0.028)	66 (0.028)	47 (0.538)	51 (0.464)

Mix entities 2

	LUNG- ALL	LUNG- AML	LUNG- CLL	LUNG- LYM	PROSTATE- ALL	PROSTATE- AML	PROSTATE- CLL	PROSTATE- LYM
04110	648 (<0.002)	539 (<0.002)	466 (<0.002)	518 (<0.002)	605 (<0.002)	508 (<0.002)	441 (<0.002)	447 (<0.002)
04115	337 (0.316)	285 (<0.002)	254 (<0.002)	255 (<0.002)	304 (0.02)	264 (<0.002)	223 (<0.002)	238 (<0.002)
04210	469 (0.02)	447 (<0.002)	348 (<0.002)	400 (<0.002)	434 (<0.002)	392 (<0.002)	313 (<0.002)	349 (<0.002)
04310	868 (<0.002)	762 (<0.002)	635 (<0.002)	683 (<0.002)	837 (<0.002)	721 (<0.002)	572 (<0.002)	642 (<0.002)
04512	506 (<0.002)	436 (<0.002)	351 (0.002)	391 (<0.002)	462 (<0.002)	400 (<0.002)	355 (<0.002)	373 (<0.002)
05210	551 (<0.002)	502 (<0.002)	395 (<0.002)	425 (<0.002)	511 (<0.002)	448 (<0.002)	339 (<0.002)	395 (<0.002)
05215	610 (<0.002)	510 (<0.002)	430 (<0.002)	470 (<0.002)	567 (<0.002)	487 (<0.002)	381 (<0.002)	417 (<0.002)
05221	314 (0.77)	287 (<0.002)	232 (<0.002)	251 (<0.002)	304 (0.018)	259 (<0.002)	218 (0.004)	241 (<0.002)
05223	313 (0.322)	292 (<0.002)	230 (<0.002)	231 (<0.002)	300 (<0.002)	269 (<0.002)	195 (0.144)	216 (<0.002)
04150	243 (0.028)	218 (<0.002)	176 (<0.002)	200 (<0.002)	251 (<0.002)	222 (<0.002)	176 (<0.002)	188 (<0.002)
03410	103 (0.008)	103 (<0.002)	80 (<0.002)	78 (<0.002)	103 (<0.002)	97 (<0.002)	68 (<0.002)	78 (<0.002)
03420	147 (0.09)	151 (<0.002)	114 (<0.002)	116 (<0.002)	146 (<0.002)	144 (<0.002)	103 (<0.002)	111 (<0.002)
03430	77 (0.54)	75 (<0.002)	58 (0.006)	66 (<0.002)	77 (0.004)	67 (<0.002)	46 (0.224)	58 (0.004)

Table 7. Selected ArrayExpress experiments of cancer entity 'all'—part 1.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-10255	all	Gene expression in primary acute lymphoblastic leukemia (ALL) associated with methotrexate treatment response	Genome-wide assessment of gene expression in primary acute lymphoblastic leukemia cells was performed to identify genomic determinants of MTX{\ ¯A}{\textcent}{\ ^A }?{\ ^A }?s antileukemic effects. Reduction of circulating leukemia cells after in vivo methotrexate treatment served as a measure MTX's antileukemic effects. Experiment Overall Design: Gene expression in diagnostic primary acute lymphoblastic leukemia cells from bo …	161/161	
E-GEOD-12995	all	Expression data for diagnosis acute lymphoblastic leukemia samples	We studied a cohort of 221 high-risk pediatric B-progenitor ALL patients that excluded known high risk ALL subtypes (BCR-ABL1 and infant ALL), using Affymetrix single nucleotide polymorphism microarrays, gene expression proling and candidate gene resequencing. A CNA poor outcome predictor was identified using a semi-supervised principal components approach, and tested in an independent validatio …	175/175	
E-GEOD-14618	all	Microarray analyses of induction failure in T-ALL	The clinical and cytogenetic features associated with T-cell acute lymphoblastic leukemia (T-ALL) are not predictive of early treatment failure. Based on the hypothesis that microarrays might identify patients who fail therapy, we used the Affymetrix U133 Plus 2.0 chip and prediction analysis of microarrays (PAM) to profile 50 newly diagnosed patients who were treated in the Children's Oncology Gr …	42/92	17495134
E-GEOD-1577	all	T-ALL and T-lymphoblastic lymphoma	T-cell acute lymphoblastic leukemia (T-ALL) and T-cell lymphoblastic lymphoma (T-LL) and are often thought to represent a spectrum of a single disease. The malignant cells in T-ALL and T-LL are morphologically indistinguishable, and they share the expression of common cell surface antigens and cytogenetic characteristics. However, despite these similarities, differences in the predominant sites …	29/29	16358311
E-GEOD-2351	all	Chemotherapy cross-resistance and treatment response in childhood acute lymphoblastic leukemia	Acute lymphoblastic leukemia (ALL) can be cured with combination chemotherapy in over 75% of children, but the cause of treatment failure in the remaining patients is unknown. We determined the sensitivity of ALL cells to individual antileukemic agents in 441 patients, and used a genome-wide approach to identify 45 genes differentially expressed in ALL exhibiting cross-resistance to prednisolone, …	129/129	15837626
E-GEOD-3912	all	First bone marrow relapse with or without initial diagnosis	This SuperSeries comprises the following subset Series:; GSE3910: 35 patients at diagnosis and relapse; GSE3911: 60 samples obtained at relapse Experiment Overall Design: Refer to individual Series	113/113	16822902
E-GEOD-4698	all	Molecular characterization of very early relapsed childhood ALL	Purpose: In childhood acute lymphoblastic leukemia (ALL), approximately 25% of patients suffer from relapse. In recurrent disease, despite intensified therapy, overall cure rates of 40% remain unsatisfactory and survival rates are particularly poor in certain subgroups. The probability of long-term survival after relapse is predicted from well-established prognostic factors, ie, time and site of …	60/60	16899601

Table 8. Selected ArrayExpress experiments of cancer entity 'all'—part 2.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-635	all	Identification of novel genomic determinants of cellular drug resistance in acute lymphoblastic leukemia.	Cellular drug resistance is associated with an unfavorable prognosis in pediatric acute lymphoblastic leukemia (ALL). To identify genes conferring resistance to antileukemic agents, we analyzed the expression of >12,700 genes in sensitive and resistant ALL cells obtained at diagnosis from 174 patients. This revealed 42, 59, 54 and 22 genes (P{\~A}{\textcent}{\^A}?{\^A}{\textcurrency}0.001) that were differentially expressed in B-lineag …	173/173	15295046
E-GEOD-8879	all	Gene expression profiling of atypical T-ALL	Despite improved therapy, approximately one-fifth of children with acute T-lymphoblastic leukemia (T-ALL) succumb to the disease, suggesting unrecognized biologic heterogeneity that may contribute to drug resistance. We studied leukemic cells, collected at diagnosis, to identify features that could define this high-risk subgroup. A total of 139 patients with T-ALL were treated consecutively from 1 …	55/55	
E-MEXP-120	all	Transcription profiling of bone marrow samples of 31 children with acute lymphoblastic leukemia to identify changes in gene expression that are associated with the current risk assignment, irrespective of the genetic subtype	We analyzed bone marrow samples of 31 children with acute lymphoblastic leukemia to identify changes in gene expression that are associated with the current risk assignment, irrespective of the genetic subtype	31/31	15257931
E-MEXP-313	all	CIT-TALL-SIGAUX	104 samples; Affymetrix U133A micro-arrays. Ninety two patients with T-ALL were diagnosed and treated at Saint-Louis hospital, Paris. Seven patients were studied at diagnosis and relapse (total 99 T-ALL samples). There were 56 children (median age 9 years old; range 1 to 16), and 36 adults (median age 27; range 17 to 66). Informed consent was obtained from the patients and/or relatives. T …	104/104	15774621
E-TABM-125	all	Translating microarray data for diagnostic testing in childhood leukaemia	We examined published microarray data from 104 acute lymphoblastic leukaemia patient specimens, that represent six different subgroups defined by cytogenetic features and immunophenotypes. Using the decision-tree based supervised learning algorithm Random Forest (RF), we determined a small set of genes for optimal subgroup distinction and subsequently validated their predictive power in an indepen …	68/68	17002788

Table 9. Selected ArrayExpress experiments of cancer entity 'breast'—part 1.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-11121	breast	The humoral immune system has a key prognostic impact in node-negative breast cancer	Estrogen receptor (ER) expression and proliferative activity are established prognostic factors in breast cancer. In a search for additional prognostic motives we analyzed the gene expression patterns of 200 tumors of patients who were not treated by systemic therapy after surgery using a discovery approach. After performing hierarchical cluster analysis, we identified co-regulated genes related t …	200/200	18593943
E-GEOD-11965	breast	Contribution of HSD17B12 for estradiol biosynthesis in human breast cancer	17beta-hydroxysteroid dehydrogenase type 12 (HSD17B12) has been demonstrated to be involved in regulation of in situ biosynthesis of estradiol (E2). HSD17B12 expression was reported in breast carcinomas but its functions have remained unknown. Therefore, we examined the correlation between mRNA expression profiles determined by microarray analysis and tissue E2 concentrations obtained from 16 postm …	16/32	
E-GEOD-12093	breast	The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy	Classification of tamixifen-treated breast cancer patients into high and low risk groups using the 76-gene signature Experiment Overall Design: 136 breast cancer samples that were treated with tamoxifen were classified using the 76-gene signature	136/136	18821012
E-GEOD-1456	breast	Gene expression of breast cancer tissue in a large population-based cohort of Swedish patients	Tissue material was collected from all breast cancer patients receiving surgery at Karolinska Hospital from 1994–1996. Material was frozen immediatley on dry ice or in liquid nitrogen and stored in –70{\~A}?{\^A}{\textdegree}C freezers. This series contains expression data for n = 159 tumors from which RNA could be collected in sufficient amounts and quality for analysis. Experiment Overall Design: All tumor specimens we …	159/318	16280042
E-GEOD-1561	breast	EORTC 10994 clinical trial	EORTC 10994 is a phase III clinical trial comparing FEC with ET in patients with large operable, locally advanced or inflammatory breast cancer (www.eortc.be). Frozen biopsies were taken at randomisation. RNA was extracted from 100 µm thickness of 14G core needle biopsies. Adjacent sections were tested by H&E to confirm >20% tumour cell content. 100 ng total RNA per chip were amplified using the Af …	49/49	15897907
E-GEOD-2603	breast	Subpopulations of MDA-MB-231 and primary breast cancers	Subpopulations of MDA-MB-231 that exhibit different metastatic tropisms when injected into immuno-deficient mice. Also, a cohort of primary breast cancers surgically resected at the Memorial Sloan-Kettering Cancer Center (MSKCC).	121/121	16049480
E-GEOD-2990	breast	Gene Expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis	Background: Histologic grade in breast cancer provides clinically important prognostic information. However, 30%–60% of tumors are classified as histologic grade 2. This grade is associated with an intermediate risk of recurrence and is thus not informative for clinical decision making. We examined whether histologic grade was associated with gene expression profiles of breast cancers and whether …	189/189	16478745

Table 10. Selected ArrayExpress experiments of cancer entity 'breast'—part 2.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-3494	breast	An expression signature for p53 in breast cancer predicts mutation status, transcriptional effects, and patient survival	The biological tumor samples (ie, breast tumor specimens) consisted of freshly frozen breast tumors from a population-based cohort of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989. Estrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. An experienced pathologist determ …	251/502	16141321
E-GEOD-4917	breast	Time course microarray data following GR activation in MCF10A-Myc breast cells	This series contain time course microarray data from MCF10A-Myc cells treated with either ethanol or Dexamethasone for 30 min, 2 hr, 4 hr, and 24 hr. This series contains three biological replicates that were analyzed as independent replicate experiments.	24/24	16690749
E-GEOD-4922	breast	Transcription profiling of human breast cancer tumor samples from Uppsala and Singapore cohorts	Histological grading of breast cancer defines morphological subtypes informative of metastatic potential, although not without considerable inter-observer disagreement and clinical heterogeneity particularly among the moderately differentiated grade **II** (G2) tumors. We posited that a gene expression signature capable of discerning tumors of grade **I** (GI) and grade **III** (G3) histology might provide a …	289/578	17079448
E-GEOD-5462	breast	Letrozole (Femara) early response to treatment	In the present investigation, we have exploited the opportunity provided by neoadjuvant treatment of a group of postmenopausal women with large operable or locally advanced breast cancer (in which therapy is given with the primary tumour remaining within the breast) to take sequential biopsies of the same cancers before and after 10–14 days treatment with letrozole. RNA extracted from the biopsie …	116/116	17885619
E-GEOD-5847	breast	Tumor and stroma from breast by LCM	Tumor epithelium and surrounding stromal cells were isolated using laser capture microdissection of human breast cancer to examine differences in gene expression based on tissue types from inflammatory and non-inflammatory breast cancer Experiment Overall Design: We applied LCM to obtain samples enriched in tumor epithelium and stroma from 15 IBC and 35 non-IBC cases to study the relative contribu …	95/95	17999412
E-GEOD-6532	breast	Transcription profiling of human breast cancers to define clinically distinct molecular subtypes in estrogen receptor positive breast carcinomas using genomic grade	Purpose: A number of microarray studies have reported distinct molecular profiles of breast cancers (BC): basal-like, ErbB2-like and two to three luminal-like subtypes. These were associated with different clinical outcomes. However, although the basal and the ErbB2 subtypes are repeatedly recognized, identification of estrogen receptor (ER)-positive subtypes has been inconsistent. Refinement of t …	327/741	17401012
E-GEOD-6772	breast	Comparison of gene expression data from human and mouse breast cancers	This SuperSeries is composed of the following subset Series; GSE6581: Expression data from mammary glands of transgenic mice; GSE6596: Comparison of gene expression data from human and mouse breast cancers: Identification of conserved breast tumor genes Experiment Overall Design: Refer to individual Series	26/38	17410534

Table 11. Selected ArrayExpress experiments of cancer entity 'breast'—part 3.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-7390	breast	Strong time dependence of the 76-gene prognostic signature	Background: Recently a 76-gene prognostic signature able to predict distant metastases in lymph node-negative (N-) breast cancer patients was reported. The aims of this study conducted by TRANSBIG were to independently validate these results and to compare the outcome with clinical risk assessment. Materials and Methods: Gene expression profiling of frozen samples from 198 N- systemically untreate …	198/198	17545524
E-GEOD-9574	breast	Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients	Normal-appearing epithelium of cancer patients can harbor occult genetic abnormalities. Data comprehensively comparing gene expression between histologically normal breast epithelium of breast cancer patients and cancer-free controls are limited. The present study compares global gene expression between these groups. We performed microarrays using RNA from microdissected histologically normal term …	29/29	18058819
E-GEOD-9936	breast	Expression data from human breast cancer cells (MCF-7) co-expressing ERalpha and Erbeta, treated with phytoestrogens	We used microarrays to detail the global transcriptional response mediated by ERalpha or ERbeta to the phytoestrogen genistein in the MCF-7 human breast cancer cell model. Experiment Overall Design: MCF-7 human breast cancer cells expressing endogenouse Estrogen Receptor Alpha (ERalpha) were infected with adenovirus carrying either estrogen receptor beta (AdERb) or no insert (Ad) at multiplicity o …	105/105	
E-MEXP-440	breast	Gene expression changes associated with lapatinib treatment of breast cancer cell lines	Dose and time course response of lapatinib in breast cancer cell lines.	36/36	17513611
E-MTAB-7	breast	MAGETAB worked examples DC 2008	A cell line comparison experiment for breat cancer 51 cancer cell lines	51/51	
E-TABM-43	breast	CIT-HS-BREAST-CANCER-CHEMOTHERAPY-RESPONSE	37 samples hybridized on Affymetrix HG-U133A arrays. Analysis of advanced breast cancers treated with a dose-intense epirubicin/cyclophosphamide regimen followed by mastectomy; Validation of TP53-related genes in breast and bladder cancers. We found that a complete response to chemotherapy was only observed in TP53 mutant tumours. We further show that, among TP53 mutant tumours, high basalcytoker …	37/37	17388661

Table 12. Selected ArrayExpress experiments of cancer entity 'lung'.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-10072	lung	Transcription profiling of human lung adenocarcinoma and non-tumors from former, current and never smoking individuals	Tobacco smoking is responsible for over 90% of lung cancer cases, and yet the precise molecular alterations induced by smoking in lung that develop into cancer and impact survival have remained obscure. We performed gene expression analysis using HG-U133A Affymetrix chips on 135 fresh frozen tissue samples of adenocar-cinoma and paired noninvolved lung tissue from current, former and never smoker …	107/107	
E-GEOD-4127	lung	Anticancer drug clustering in lung cancer based on gene expression profiles and sensitivity database	Anticancer drug clustering in lung cancer based on gene expression profiles. We performed gene expression analysis in lung cancer cell lines, (used: Affymetrix GeneChip Human Genome U133 Array Set HG-U133A). We also examines the sensitivity of these cell lines to commonly used anti-cancer agents (docetaxel, pacli-taxel, gemcitabine, vinorelbine, 5-FU, SN38, cisplatin, and carboplatin) via MTT assay …	29/29	
E-GEOD-4824	lung	Analysis of lung cancer cell lines	These arrays are used for various projects Experiment Overall Design: HG-U133A and HG-U133B data are combined and analyzed together with other U133A and B or with HG-U133plus2 samples. No replicates were performed. Controls are human bronchial epithelial cells (HBECs)	79/164	16843264
E-GEOD-6253	lung	A gene expression signature predicts survival of patients with stage i non-small cell lung cancer	We applied a meta-analysis of datasets from seven different microarray studies on lung cancer for differentially expressed genes related to survival time (under 2 y and over 5 y). Systematic bias adjustment in the datasets was performed by distance-weighted discrimination (DWD). We identified a gene expression signature consisting of 64 genes that is highly predictive of which stage I lung cancer …	18/72	17194181
E-GEOD-7670	lung	Expression data from Lung cancer	Detection, treatment, and prediction of outcome for lung cancer patients increasingly depend on a molecular understanding of tumor development and sensitivity of lung cancer to therapeutic drugs. The application of genomic technologies, such as microarray, is widely used to monitor global gene expression and has built up invaluable information and knowledge, which is essential to the discovery of …	66/66	17540040
E-MEXP-231	lung	Normal Lung + Lung adenocar-cinoma microarray	Gene transcription in a set of 49 human primary lung adenocarcinomas and 9 normal lung tissue samples was examined using Affymetrix GeneChip technology. We aimed to investigate differential gene expression between the two tissue types. A total of 3,442 genes, called the set MAD, were found to be either up- or down-regulated by at least two fold between the two phenotypes. Genes assigned to a parti …	58/58	15653641
E-TABM-15	lung	Transcription profiling of cancerous and non cancerous lung ade-nocarcinoma tissue. Tumour and normal samples from human lung carcinoma from 18 patients plus tumour only from 5 patients	Comparison of gene expression of cancerous and non cancerous lung adenocarci-noma tissue. Tumour and normal samples from 18 patients plus tumour only from 5 patients.	41/41	

Table 13. Selected ArrayExpress experiments of cancer entity 'colon'.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-2742	colon	Genomic strategies identify the antitumor agent apratoxin A as a potent antagonist of FGF signaling and STAT3 activation	Total RNA was extracted from apratoxin A or vehicle treated HT29 cells using the RNeasy Mini Kit (Qiagen). Probe values from CEL files were condensed to probe sets using Rosetta Resolver software. Resolver ANOVA analysis was then performed between groups. Experiment Overall Design: 2 doses were compared to vehicle at 3 time points. Each time point had its own vehicle control	27/27	16474387
E-GEOD-4045	colon	Classification of serrated colorec-tal tumors	Serrated adenocarcinomas are morphologically different from conventional adeno-carcinomas. The serrated pathway has recently been proposed to represent a novel mechanism of colorectal cancer (CRC) formation. However, whether they are biologically different and truly form a distinct subclass of CRC, is not known. This study shows that the gene expression profile of serrated and conventional CRCs dif …	37/37	16819509
E-MEXP-101	colon	Narayanan Lab RKO SIM2s antisense	RKO Colon Carcinoma Cells were treated with 100 nM of either SIM2s Control or Antisense oligos in a timecourse (10, 14, 18, 24 hr) dependent manner.	32/32	16129820
E-MEXP-383	colon	Colorectal cancer: UICC II versus UICC III	Analyze differential expression between stage UICC II and UICC III colorectal cancer	36/36	16721809
E-MEXP-833	colon	Croner_CRC_GBP-I	The gene expression profile of 24 human colorectal carcinoma patients either expressing high (n = 12) or low (n = 12) levels of human guanylate binding protein-1 (GBP-1) were compared in order to identify coregulated genes.	24/24	
E-MTAB-57	colon	Transcription profiling of colon cancer tumor and normal biopsies from a series of patients to identify molecular biomarkers	Expression profiling studies on colon cancer comparing tumoral and normal biopsies from a series of patients in order to identify molecular biomarkers.	47/47	16919171

Table 14. Selected ArrayExpress experiments of cancer entity 'prostate'.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-2443	prostate	Prostate cancer—comparison of androgen-dependent and—independent microdissected primary tumor	Affymetrix U133A comparison of two groups (10 samples each): untreated (androgen-dependent) primary prostate cancer (Gleasons 5–9) and androgen-independent primary prostate cancer. All samples were microdissected for tumor cells only.	20/20	16203770
E-GEOD-8218	prostate	Transcription profiling of human prostate cancer samples	Prostate cancer gene expression profiles were studied in this project. A total RNA from 148 prostate sample with various amount of different cell types were hybridized to Affymetrix U133A arrays. The percentage of different cell types vary considerably among samples and were determined by pathologist. Cell type specific genes can be determined by linear regression using the methods of Stuart et al …	148/148	
E-MEXP-1327	prostate	Selenium vitamin E trial in Prostate Cancer	85 radical prostatectomy specimens (where 16 samples are in Placebo group (PL), 15 are in Selenium group (SE), 25 are in Vitamin E group (VE) and 27 are in Vitamin E and Selenium group vs. Treatment groups: 1-selenomethionine, 400 ug + placebo (vitamin E); vitamin E, 400 IU + placebo (1-selenomethionine); 1-selenomethionine, 400 ug + vitamin E, 400 IU; placebos) were subjected to laser capture micr …	85/85	
E-TABM-26	prostate	CSM-Prostate-Cancer-Samples	Microarray studies of Prostate tissues obtained from multiple Institutions. Analysis done during Aug. 2002 to June 2004.	57/114	16618720
E-TABM-90	prostate	Transcription profiling of irradiated human lymphyocytes from prostate carcinoma patients following curative radiotherapy to study late radiation toxicity	For a case-control study, we selected 54 prostate carcinoma patients with no evidence of disease 2 years after curative	108/108	

Table 15. Selected ArrayExpress experiments of cancer entity ‘aml’.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-1159	aml	Expression profiles of acute myeloid leukemia patient samples	Expression profiles of acute myeloid leukemia patient samples. Blasts and mononuclear cells were purified from bone marrow or peripheral blood aspirates of acute myeloid patients. Samples contained 80–100 percent blast cells. Total RNA was extracted by lyses with guanidium isothiocyanate followed by cesium chloride gradient purification	293/293	17910043
E-GEOD-12417	aml	Prognostic gene signature for normal karyotype AML	Patients with cytogenetically normal acute myeloid leukemia (CN-AML) show heterogeneous treatment outcomes. We used gene expression profiling to develop a gene signature that predicts overall survival (OS) in CN-AML. Based on data from 163 patients treated in the German AMLCG 1999 trial and analyzed on oligonucleotide microarrays, we used supervised principal component analysis to identify 86 prob …	163/326	18716133
E-GEOD-1729	aml	Gene expression profile of acute myeloid leukemia	Gene expression profile of acute myeloid leukemia. Bone marrow (BM) samples from 43 adult patients with newly de novo diagnosed AML. All samples contained more than 80% blast cells. Total RNA was extracted using Trizol reagent (Life Technologies, Gaithersburg, MD) and purified with RNeasy Mini Kit (Quiagen, Valencia, CA). The RNA integrity was assessed using Agilent 2100 Bioanalyzer (Agilent, Palo …	43/43	15674361
E-GEOD-9476	aml	Abnormal expression changes in AML	Acute myeloid leukemia (AML) is one of the most common and deadly forms of hematopoietic malignancies. We hypothesized that microarray studies could identify previously unrecognized expression changes that only occur only in AML blasts. We were particularly interested in those genes with increased expression in AML, believing that these genes may be potential therapeutic targets. Experiment Overa …	64/64	17910043

Table 16. Selected ArrayExpress experiments of cancer entity 'cll'.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-11038	cll	Molecular and transcriptional characterization of chromosome 17p loss in chronic lymphocytic leukemia	Distinct genetic abnormalities such as TP53 deletion at 17p13.1, have been identified as having an adverse prognostic relevance in B-cell chronic lymphocytic leukemia (B-CLL). Conventional cytogenetic studies have shown that TP53 deletion in B-CLL is associated predominantly with 17p loss resulting from complex chromosomal rearrangements. We performed genome-wide DNA (SNPs arrays), fluorescence in …	60/72	18521849
E-GEOD-6691	cll	Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglob-ulinemia.	The tumoral clone of Waldenstrom{\^A}?s macroglobulinemia (WM) shows a wide morphological heterogeneity which ranges from B-lymphocytes (BL) to plasma cells (PC). By means of genome-wide expression profiling we have been able to identify genes exclusively deregulated in BL and PC from WM, but with a similar expression pattern in their corresponding cell-counterparts from CLL and MM, as well as normal i …	56/56	17252022
E-GEOD-8835	cll	Chronic lymphocytic leukemia cells induce changes in gene expression of CD4 and CD8 T cells.	To examine the impact of tumors on the immune system, we compared global gene expression profiles of peripheral blood T cells from previously untreated patients with B cell chronic lymphocytic leukemia (CLL) with those from age-matched healthy donors. Although the cells analyzed were not part of the malignant clone, analysis revealed differentially expressed genes, mainly involved in cell differen …	66/66	15965501

Table 17. Selected ArrayExpress experiments of cancer entity 'lymphoma'.

ID	Entity	Title	Description	Arrays	PubMed ID
E-GEOD-4475	lymphoma	A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling	The distinction between the Burkitt lymphoma and diffuse large B-cell lymphoma is imprecise using current diagnostic criteria. We applied transcriptional and genomic profiling to molecularly define Burkitt lymphoma. Gene expression profiling employing Affymetrix GeneChips (U133A) was performed in 220 mature aggressive B-cell lymphomas, including a core group of eight Burkitt lymphomas, which fulfi …	221/221	16760442
E-GEOD-8388	lymphoma	Epigenetic upregulation of B-cell inappropriate genes induces extinction of B-cell program in classical Hodgkin lymphoma	A unique feature of the tumour cells (Hodgkin/Reed-Sternberg (HRS)) of classical Hodgkin lymphoma (cHL) is the loss of their B-cell phenotype despite their B-cell origin. Several lines of evidence suggest that epigenomic events, especially promoter DNA-methylation, are involved in this silencing of many B-cell associated genes. Here we show that DNA-demethylation alone or in conjunction with histo …	24/24	
E-TABM-117	lymphoma	CIT_LYMPHOMA_ALCL_ALK_SUBTYPES	Affymetrix UU133A gene expression data for a series of 32 cases of systemic anaplastic Large Cell Lymphoma (ALCL) and 5 ALCL cell lines; used to 1) confirm that tumors expressing Anaplastic Lymphoma Kinase (ALK+ ALCL) and ALK- ALCLs are different entities, 2) identify most significantly differentially expressed genes between ALK+ and ALK- samples, 3) generate a molecular signature of ALK- A …	37/37	17077326
E-TABM-346	lymphoma	CIT-HS-DLBCL-KLY	53 samples hybridized on Affymetrix HG-U133A GeneChips arrays, for 53 patients with diffuse large B-cell lymphoma (DLBCL); patients are treated with CHOP (cyclophosphamide, doxorubicin, vincristine, prednisone) or Ritxumab (R)-CHOP in the Groupe d{\^A}?Etude des Lymphomes de l{\^A}?Adulte (GELA) clinical centers.	53/53	