AIMS Medical Science, 3(3): 248–271.
DOI: 10.3934/medsci.2016.3.248
Received 30 July 2016,
Accepted 5 September 2016,
Published 10 September 2016

http://www.aimspress.com/journal/medicalScience

Review article

A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics

Giuliano Crispatzu [1, 2], Alexandra Schrader[1, 2], Michael Nothnagel [3], Marco Herling [1, 2, *], and Carmen Diana Herling [1, *]

[1] Department of Internal Medicine I, Center for Integrated Oncology (CIO) Köln-Bonn, University of Cologne (UoC), Germany;
[2] Excellence Cluster for Cellular Stress Response and Aging-Associated Diseases (CECAD), UoC, Germany;
[3] Cologne Center for Genomics (CCG), Department of Statistical Genetics and Bioinformatics, UoC, Germany

* **Correspondence:** Email: carmen.herling@uk-koeln.de; marco.herling@uk-koeln.de;
Tel: +49-221-478-5969; Fax: +49-221-478-6383

Abstract: In cancer research, transcriptional aberrations are often deduced from mRNA-based gene expression profiling (GEP). Although transcriptome sequencing (RNA-seq) has gained ground in the recent past, mRNA-based microarrays remain a useful asset for high-throughput experiments in many laboratories. Possible reasons are the lower per-sample costs and the opportunity to analyze obtained GEP data in association with published data sets. There are established and widely used methods for the analysis of microarray data, which increase the comparability of different GEP data sets and facilitate data-mining approaches. However, analytic pitfalls, such as batch effects and issues of sample purity, e.g. by complex tissue composition, are often not properly addressed by these standard approaches. Moreover, most of these tools do not capitalize on the full range of public data sources or do not take advantage of the analytic possibilities for functional interpretation or of comprehensive meta-analyses. We present an overview of the most critical steps in the analysis of microarray-based GEP data. We discuss software and database query solutions that may be useful for

each step and for generally overcoming analytic challenges. Aside from machine-learning applications to classify and cluster samples, we describe clinical applications of GEP, including a novel exploratory algorithm to identify potential biomarkers of prognosis in small sample cohorts as demonstrated by exemplary data from lymphatic leukemias. Overall, this review and the attached source code provide guidance to both molecular biologists and bioinformaticians / biostatisticians to properly conduct GEP analyses as well as to evaluate the clinical / biological relevance of obtained results.

Keywords: Cancer genomics; gene expression profiling; microarray; RNA-Seq; survival analysis; CLL; T-PLL; leukemia; lymphoma; TCL1; contamination; SVM; random forest

1. Introduction

Traditionally, gene expression analysis includes reverse transcription of mRNA into cDNA and probing of gene transcripts of interest by specific primers designed for target PCR amplification (gold standard), followed by quantitative, semi-quantitative (e.g. qRT-PCR), or electrophoresis (e.g. Southern blotting) detection methods. Based on efforts provided by the Human Genome Project [1,2] and studies on expressed sequence tags (ESTs) in mammalian genomes, cDNA hybridization array chips have originally been designed to investigate deregulated mRNA expression of distinct and well-characterized gene transcripts in various diseases. Modern mRNA-microarray platforms apply one or two-color fluorescence labeling (i.e. Cyanine3 / Cy3 for green and Cy5 for red dye fluorescence) for one or two samples to be loaded on the chip, respectively, and allow the detection of more than 47 000 transcripts. In contrast to two-color arrays (e.g. HuA1 by Agilent Technologies, Santa Clara, CA, USA), one-color arrays, are most commonly used today (e.g. HG-U133 Plus 2.0 by Affymetrix, Inc., Santa Clara, CA, USA, or BeadArray HT-12v4, Illumina, Inc., San Diego, CA, USA) and represent the focus of this review.

The past few years have seen the advent of transcriptome sequencing (RNA-seq) based on the next-generation sequencing (NGS) technology using high-throughput platforms, such as the GA IIx or HiSeq2000 sequencer from Illumina. RNA-seq does not require the prior design of specific probes, rendering it a highly versatile approach for gene expression profiling (GEP). Accordingly, a number of publications on the genomic landscape of various neoplasms have applied RNA-seq to investigate gene-specific aspects such as differential splicing and exon usage [3], hidden viral transcripts [4], and cancer-specific fusion transcripts [5]. However, published reports using RNA-seq in cancer often lack statistical power for comprehensive gene expression analyses due to a limited sample size. In contrast, mRNA-based microarrays have remained the initial method of choice for high-throughput analyses of gene expression in many laboratories. Reasons for this include the associated lower per-sample costs as well as the availability of already published microarray-derived GEP data in

public databases. Many of these data sets were processed by established and widely used methods, thereby improving their comparability and the suitability for data-mining approaches.

Within this review, we present an overview of critical steps in the analysis of microarray-based GEP data (see overview in Figure 1) and the corresponding library and code information (summarized in Table 1 and 2). We will discuss step-by-step software and database query solutions that may be useful for data analysis, to avoid analytic pitfalls, and to provide an increased capability for clinical and biological interpretation of data. To illustrate the proposed analytic steps, we present analyses on exemplary data of previously published and own GEP data, all obtained in patients with B- and T-cell leukemias or lymphomas.

2. Quality Control can Greatly Differ by Platform

There are various possibilities to apply basic steps of quality control (QC) prior to or during preprocessing of GEP raw data. In order to avoid false estimates of background intensities and false inputs for normalization, removal of potential problematic samples and probes *before* data preprocessing is essential towards a correct interpretation of data. Problematic samples often present as outliers in density distributions or in an unsupervised cluster analysis on global gene expression values (*after* data preprocessing). The latter, e.g. in form of dendrograms (Code 1) or principal component analyses (PCA; Code 2), is created by using the R [7] library *arrayQualityMetrics* from Bioconductor [8] with its informative HTML report per array.

Numerous methods and libraries for R are available for more specific quality assessments for each of the three major microarray platforms. Affymetrix arrays can be analyzed using the *affyQCReport* and *simpleaffy* libraries (see Table 1 for all library references), which normalize expression values using housekeeping genes (e.g. calculating the actin3/actin5 ratio), while the *affyPLM* library allows calculation of important quality measures such as the normalized unscaled standard error (NUSE) and relative log expression (RLE) as well as their plotting across samples (Code 3). The quality of data obtained with Illumina chips can be assessed by statistical standard measurements (mean and standard deviation) or outlier detection using the *lumiQ* function within the *lumi* library (Code 4). Possible slide inhomogeneities (i.e. scratches) or contamination on two-color arrays may be detected with the *imageplot* function of the *limma* library. This package also allows the calculation of the RNA Integrity Number (RIN) as a measure of mRNA degradation with a subsequent option to remove samples below a given threshold.

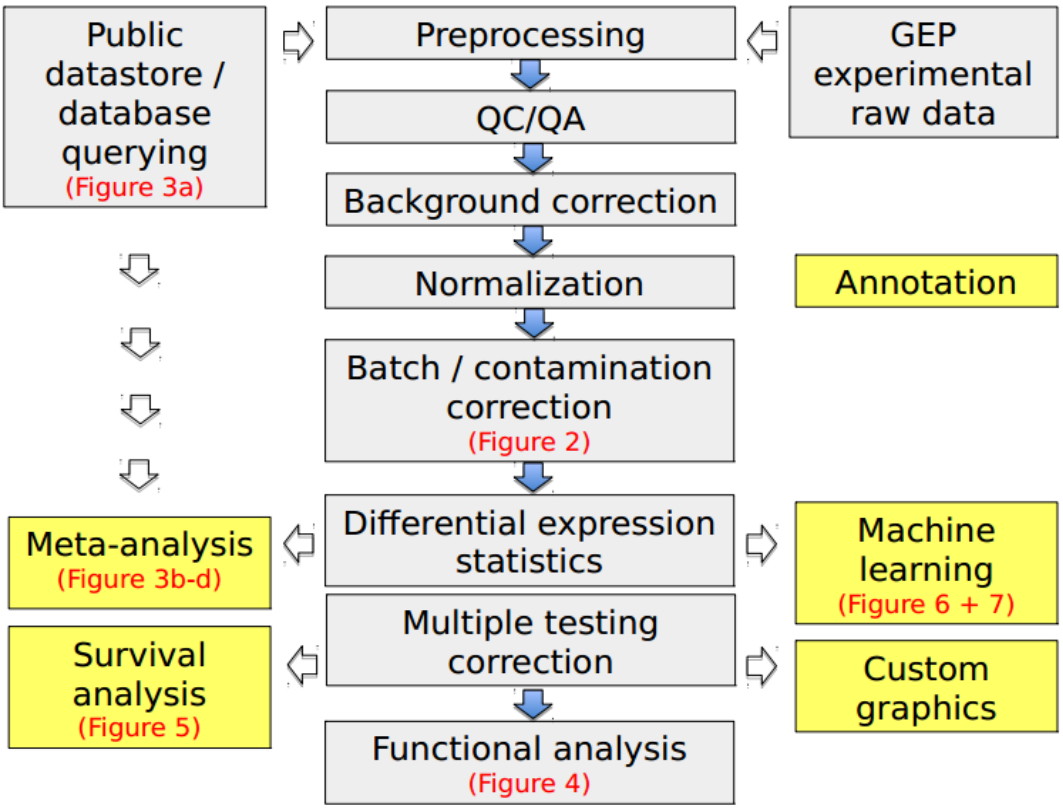


Figure 1. Flow chart describing a suggested GEP protocol. Steps in yellow boxes are modular and may function somewhere independently downstream of the steps in grey boxes. The red text refers to those figures of this review that illustrate the respective step.

3. Proper Preprocessing of Raw Data

A first step in the standard analysis protocol of cDNA microarrays usually is the conversion of hybridization image spots obtained by array scanners into raw gene expression values. For Affymetrix chips this is normally done either by using the freeware *Affymetrix Power Tools* or the R library *affy*. For Illumina's BeadChips the proprietary *GenomeStudio* software or manual decryption via the R library *beadArray* may be used. For two-color arrays, scanner output files, e.g. in TIF format, can easily be read with the *read.maimages* function from the *limma* R library.

In a second step, background correction is conducted by subtracting technical noise from biological variation. This is accomplished by using e.g. *RMA* [9] for Affymetrix arrays or the *bgAdjust* function from the *lumi* R library for Illumina arrays, which employs a similar algorithm as *GenomeStudio* (Code 5). In order to account for outliers and to remove systematic variation, normalization of expression values is required. The most common procedures include quantile-normalization, which preserves the rank, but may eliminate small differences in expression values, and LOESS (locally weighted scatterplot smoothing)-normalization, which does the opposite. Robust splice normalization (RSN) aims to combine the advantages of both methods through a

monotonic splice fit to one reference sample, while simple scaling normalization (SSN) forces samples to have the same scale and background. Both approaches are included in the *lumi* R library for Illumina arrays. For two-color arrays it may be essential to further account for dye biases in the normalization [10] and to normalize within the array itself (between both color-labeled samples) and between all two-color arrays of the cohort, e.g. by use of the *limma* R library. Variance-stabilizing normalization (VSN) constitutes another method for combining background correction and normalization [11], while preserving biological variation. It is implemented in the *vsn* (Code 6) library, applicable to arrays of all major platforms. Within the normalization process raw intensities are usually transformed, either into a log2 scale or glog in case of VSN, in order to smoothen extreme values.

4. Probe Annotation and Deconvolution

Frequent impediments for GEP data analysis are missing array annotations or outdated annotation files provided by the manufacturers (e.g. frequently old GenBank predictions are included). Data-mining tools such as *biomaRt* [12] can be used to acquire up-to-date probe information (Code 7). They may also be helpful in assigning probes to transcripts, thereby enabling filtering for redundancies of probes, which map primarily to transcripts that are prone to nonsense-mediated mRNA decay (NMD) or to unprocessed pseudogenes. Deconvolution of genes with known transcript variants of differential function into probed isoforms may also be important for extrapolations on biological relevance. An example is the apoptosis regulator *myeloid cell leukemia sequence 1 (MCL1),* of which the longer isoform (MCL1-001) has been reported to enhance survival by inhibiting apoptosis, while its shorter isoform (MCL1-002) acts as a pro-apoptotic molecule [13].

5. Exploring Differentially Expressed Genes Considering the "Multiple Comparisons Problem"

Raw data preprocessing and QC is followed by the actual statistical analysis, usually in the form of probe-by-probe hypothesis tests for differential expression including: (1) two-group mean comparisons using a Student's t-test (parametric, i.e. presuming a known statistical distribution), (2) empirical Bayes / moderated t-tests (for low sample size; e.g. $n < 10$; parametric), (3) Mann-Whitney-U tests (for samples with low variability; non-parametric) (Code 8), (4) multiple-group tests by means of an analysis of variance (ANOVA; parametric) (Code 9), or (5), a Jonckheere test (trend test; non-parametric). However, statistical testing of all genes / transcripts detected by an array requires correction for multiple testing, in order to avoid a substantial number of false-positive findings [14,15]. For example, using a significance level of 0.05 for each of 10,000 tests would result in approximately 0.05 * 10,000 = 500 significant rejections by chance, even if all null hypotheses of no differential expression were true. To this end, we can either control the family-wise error rate (FWER) to curtail

the number of statistically significant results, e.g. by use of the (conservative) Bonferroni correction, in which the significance level for each probe-specific test equals the FWER (e.g. 0.05) divided by the total number of tested probes, or by some permutation / resampling approach. Furthermore, we can aim for controlling the false-discovery rate (FDR), i.e. the proportion of falsely rejected null hypotheses, e.g. using the Benjamini-Hochberg's procedure, q-values, or other approaches. It should be noted, however, that control of the FDR, while very helpful in limiting the number of erroneously followed-up probes, does not imply a notion of statistical significance. The procedures by Bonferroni and by Benjamini-Hochberg are implemented in the *multtest* library [16], while the *qvalue* library provides an implementation for the rank-preserving q-value calculation (Code 10).

Nominally differentially expressed probes (e.g. with a single-test level of $p < 0.05$) can also be filtered by multiple-testing correction, for example by applying a q-value / FDR cutoff (common cut-off, e.g. 0.1) to ensure a low proportion of false-positives in the set of probes to be subsequently followed up. To reduce time in the analysis, it may also be useful to exclude genes / probes that are not expected to be differentially expressed either due to biologically low variability in the investigated samples, or due to technically low detectability on the array. This can be achieved either by non-specific filtering of expression values restricted to a given range (e.g. the shortest interval containing half of the data by standard deviation (sd) or interquartile range) or by setting an empirical cut-off to the coefficient of variation (sd/mean), e.g. the top 10 percent or a fixed value of 0.6. Note, however, that this may increase the rate of false-negative findings (Code 11).

6. Pitfalls: Batch-correction and Contamination Estimation

When comparing GEP data obtained in the same laboratory, but with two or more different batches of arrays, the results will deviate from one another beyond the expected biological and array-specific technical variation. Batch correction addresses this issue. Two approaches commonly considered to be performing best [17] are mean-centering and a Bayesian framework named ComBat [18] (Figure 2a–c; Code 12).

A particular problem for cancer transcriptomics / genomics is the contamination of cancer tissues by normal cells (irrespective whether to consider them as actual milieu components) and vice versa. Even in lymphomas and lymphoid leukemias, such problems are encountered in lymph-node samples or in the seemingly 'pure' blood samples, as these are also of mostly multicellular composition. Tools like *ESTIMATE* [19] can weigh specific markers (e.g. indicating an immune or stromal cell origin) within gene expression profiles in the form of gene set enrichment analyses and thus evaluate the degree of purity. Unfortunately, due to intrinsic aberrations of 'immune cell' genes within tumor cells of leukemias / lymphomas, the immune gene set used within *ESTIMATE* is not reliable for the enrichment analysis within these malignancies (Figure 2d; Code 13). An alternative approach especially for leukemias / lymphomas might be *CellMix* [20] which uses gene sets from

specific immune cell subsets, e.g. CD4+ and CD8+ T-lymphocytes, CD14+ monocytes, CD19+ B-lymphocytes, CD56+ natural killer cells, and CD66b+ granulocytes.

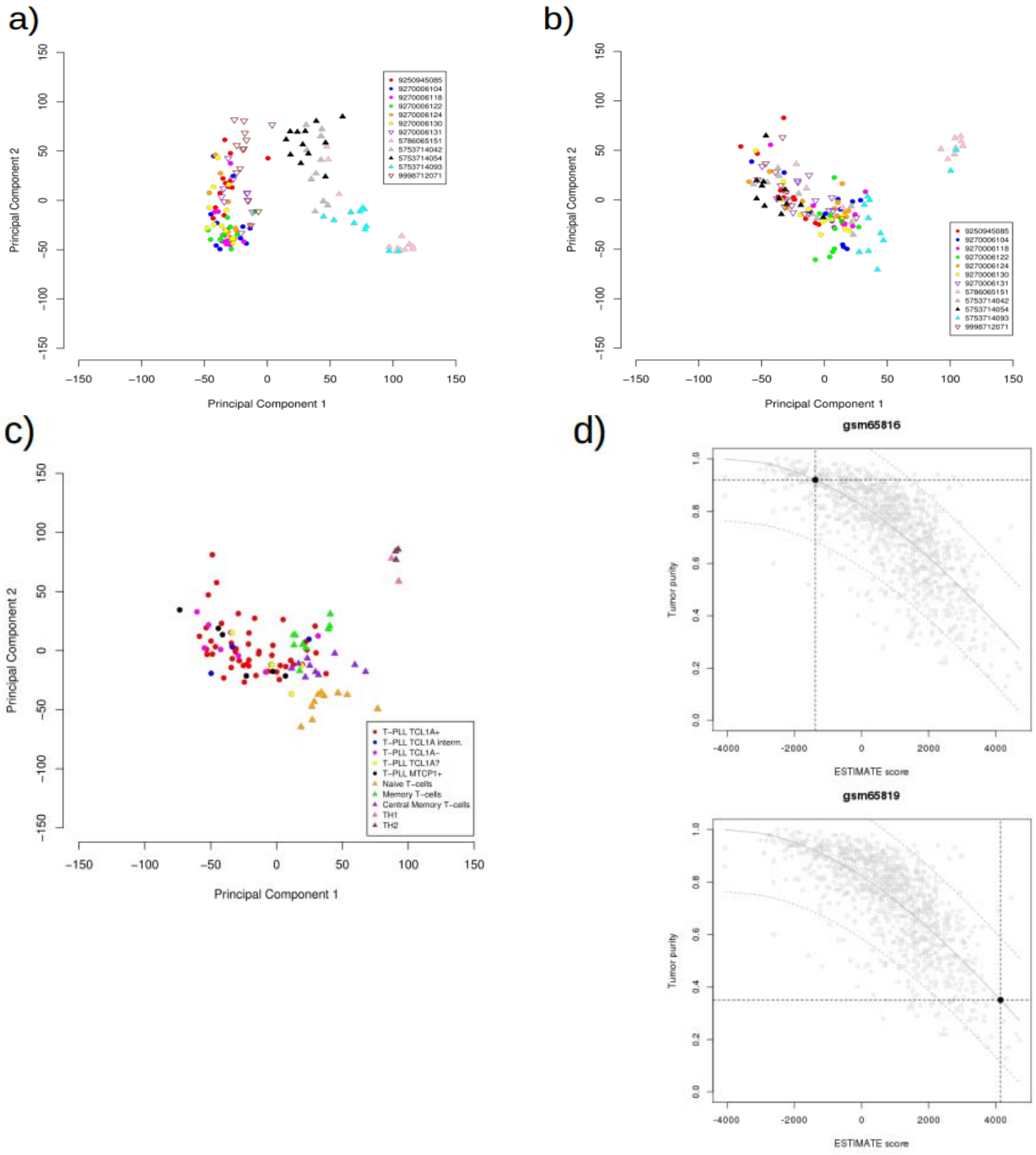


Figure 2. a) PCA (principal component analysis) of the 1000 most variable genes (by variation coefficient) within 12 distinct batches of our T-PLL (T-cell prolymphocytic leukemia) data set reveals batch-specific clustering. **b)** After batch correction samples do not cluster anymore due to technical bias, but rather due to biological information when annotated as in *c)*. **c)** Entity information can be included in ComBat (besides batch information) to fit batches. T-PLL samples (further divided by different oncogene protein status) and normal T-cells form a cloud,

while stimulated T-helper cells (TH1 and TH2) form another cloud. **d)** *ESTIMATE* plots of fitted purities from two samples within the publicly available breast cancer data set GSE2990 [48] (*n* = 189 invasive breast carcinomas; including 64 estrogen receptor (ER)-positive tumors, histologic grade 1 and 3 tumors; Affymetrix HG-U133A). **Upper panel:** When comparing the black dot to gray dots (all other samples), one can observe that the sample is among those with highest purity. **Lower panel:** sample among those with lowest purity.

7. Making Use of Public Databases

Two public databases are commonly used for the comparison of own microarray data with independent data sets, for example in a meta-analysis, namely the GEO (gene expression omnibus) database [21] (http://www.ncbi.nlm.nih.gov/geo) and the ArrayExpress database [22] (https://www.ebi.ac.uk/arrayexpress), with GEO featuring a larger number of integrated samples. Both platforms use distinct annotation / meta-data file systems. In GEO, samples are either described in MIAME Notation in Markup Language (MINiML; pronounced 'minimal') or SOFT formatted family files. In ArrayExpress, sample and data relationships (SDR) are described in the SDRF format, while protocol information is stored in the Investigation Description Format (IDF). Both databases offer processed numerical gene expression values (in the form of matrices) stored in regular text format (txt), or raw data in CEL or idat (for Affymetrix or Illumina chips) files. GEO and ArrayExpress also provide respective R libraries to automate queries and processing of differential expression analyses, namely *GEOquery* and *ArrayExpress*.

Analysis results for data sets within ArrayExpress are further integrated in the 'Gene expression atlas' of the EMBL / EBI (http://www.ebi.ac.uk/gxa). The latter provides information about gene and protein expression in animal and plant samples for different cell types, tissues, developmental stages, diseases, and other conditions from 1572 studies as of August 2015 [23]. The human data sets are currently exported into an RDF version accessible via a SPARQL Endpoint (http://www.ebi.ac.uk/rdf/services/atlas/sparql; accessed 02/21/2016).

Implemented queries include:

- “Query 1: Get experiments where the sample description contains diabetes”
- “Query 2: Get differentially expressed genes where factor is asthma”
- “Query 3: Show expression for ENSG00000129991 (TNNI3)”
- “Query 4: Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot (Federated query to http://sparql.uniprot.org/sparql)”
- “Query 5: For the genes differentially expressed in asthma, get the gene products associated to a Reactome pathway”
- “Query 6: Get all mappings for a given probe e.g. A-AFFY-1/661_at”

Query 2 and 5 can be further modified in order to compare gene dysregulation in other types of diseases, e.g. in lymphoid leukemias, such as chronic lymphocytic leukemia (CLL; Table 3). User's familiarity with the underlying ontologies (controlled vocabulary; [24]) is, however, necessary to construct queries.

8. Meta Analyses: Exploring Possible Phenotypic Markers across Different Conditions

For conceptualizing a pharmacologic compound (e.g. inhibitor) acting against a specific gene product or for designing specific gene-knockouts within a model organism, it may be particularly important to know in what conditions and disease subtypes expression of a distinct gene is up- or down-regulated and to which degree (basal or extreme). Integrative analyses of expression changes within a multitude of samples of the same entity, or model organism, or any other comparable biological system as well as across initially separately analyzed (and published) series (cohorts) are often called gene expression meta-analyses. In the following we describe multiple ways to conduct a meta-analysis of GEP data with their limitations and advantages.

The first approach includes construction and sending of specific queries to the EMBL / EBI RDF platform. Querying can further be semi-automated using the *SPARQL* R library, which allows the investigation of different data sets in a specific condition, e.g. comparisons of CLL vs. normal B-cells, or between distinct groups of tumor samples stratified by a characteristic of interest, e.g. immunoglobulin heavy chain (*IGHV*) gene mutated vs. unmutated CLL. Results are usually tabularized and fold-changes visualized within a heatmap (Figure 3a; Suppl. Table 1; Code 14).

Since not all 'ArrayExpress' data sets are yet integrated into the EMBL / EBI RDF platform and the GEO database contains additional data sets, the manual download, processing, and integration of such additional data is often necessary.

Therefore, a second, more hands-on approach to meta-analyses is a search by keyword, e.g. 'chronic lymphoid', within GEO and / or ArrayExpress (or any other public database). Once the data set has been picked, it is background-corrected and the annotated replicates can be combined with their original samples by calculating their mean. Afterwards all samples within the data set are normalized (e.g. quantile-normalized).

Probe sets of a gene which map to retained / dysfunctional transcripts (or which map to more retained / dysfunctional transcripts than other probe sets of the same gene) should be removed to obtain meaningful expression values (Suppl. Table 2). For example, *BCL2L11* on Affymetrix HG-U133 Plus 2.0 chips has two probes, one hybridizes two protein-coding and six NMD (nonsense-mediated decay) transcripts, the other one hybridizes two protein-coding and eight NMD transcripts. Thus, ambiguous expression values of this gene have to be evaluated with caution. The residual unambiguous probe sets assigned to a gene are then further summarized by calculation of average expression values per gene.

For further evaluation of the GEP meta-analysis, three different techniques for integration can be used to observe gene expression patterns and entity clustering:

1) The first method quantile-normalizes a matrix of average gene expressions across entities from different experiments and finally gives a visual approximation. If there is also a tumor suppressor gene (very low expression) and an oncogene (very high expression) in the gene set to be evaluated, one can expect an expression range similar to the whole transcriptome. It should be noted that in previous Affymetrix sets, such as HG-U133A, some genes (e.g. *BMF* and *BOK*) are not covered by specific probes on the array and, therefore, need to be imputed by the median of the respective data set. This guarantees that in the heatmap (or PCA) these genes are not visualized as up- or down-regulated; they in fact can be manually labeled (blackened). Expression values from all data sets are merged into one matrix and again quantile-normalized to account for variability in platform specifications and noise. A more suitable approach than normalizing on each gene set separately might be to normalize on the whole combined transcriptome (intersection of all probed genes). However, this would disregard genes not covered by all platforms used. The resulting heatmap (generated by function *heatmap.2*, library *gplots*; Figure 3b) shows the expression of selected genes and transcripts in their respective data set and can be additionally subdivided by the different entities (median across samples of an entity).

2) Batch effects cannot be entirely excluded by using method *1)* as may be observed by a bias in clustering of samples from the same experiment. Therefore, we recommend a novel method called *inSilicoMerge* [25], which combines data sets and removes their batch effect with a choice of various methods, such as the empirical Bayes method ComBat (Figure 3c).

Unfortunately, data sets from different platforms can only be combined gene-wise, meaning that e.g. *MCL1* would not be deconvolutable into its isoforms MCL-001 / MCL1-long and MCL-002 / MCL1-short.

3) For an advanced evaluation, one can further perform differential expression analysis for data sets with different control samples (of varying quality, number, and specificity) available for comparison, such as ’normal‘ non-malignant cells or bulk tissue specimens. Fold-changes with a p-value < 0.1 (trend) or < 0.05 (significant) are extracted to compare normal-matched gene expression between different experiments and probe targets representing different gene transcripts or protein isoforms. The results are again visualized by a heatmap, either in the order obtained by hierarchical clustering (using Euclidean distance) or in order of rows sorted by gene name.

As exemplified by illustration of expression levels of *Death-Associated Protein Kinase* (*DAPK*) gene family members in subsets of CLL and normal B-cells (Figure 3d), this method allows different disease vs. ’normal‘ comparisons and facilitates the evaluation of which genes are exclusively down- or up-regulated and which show no clear pattern or which are specific to small subgroups. In the meta-analysis itself every differential expression analysis is further evaluated by statistical testing. Default setting is the Student’s t-test, except for low variation or non-normal distributions, for which the non-parametric Wilcoxon rank sum test is recommended.

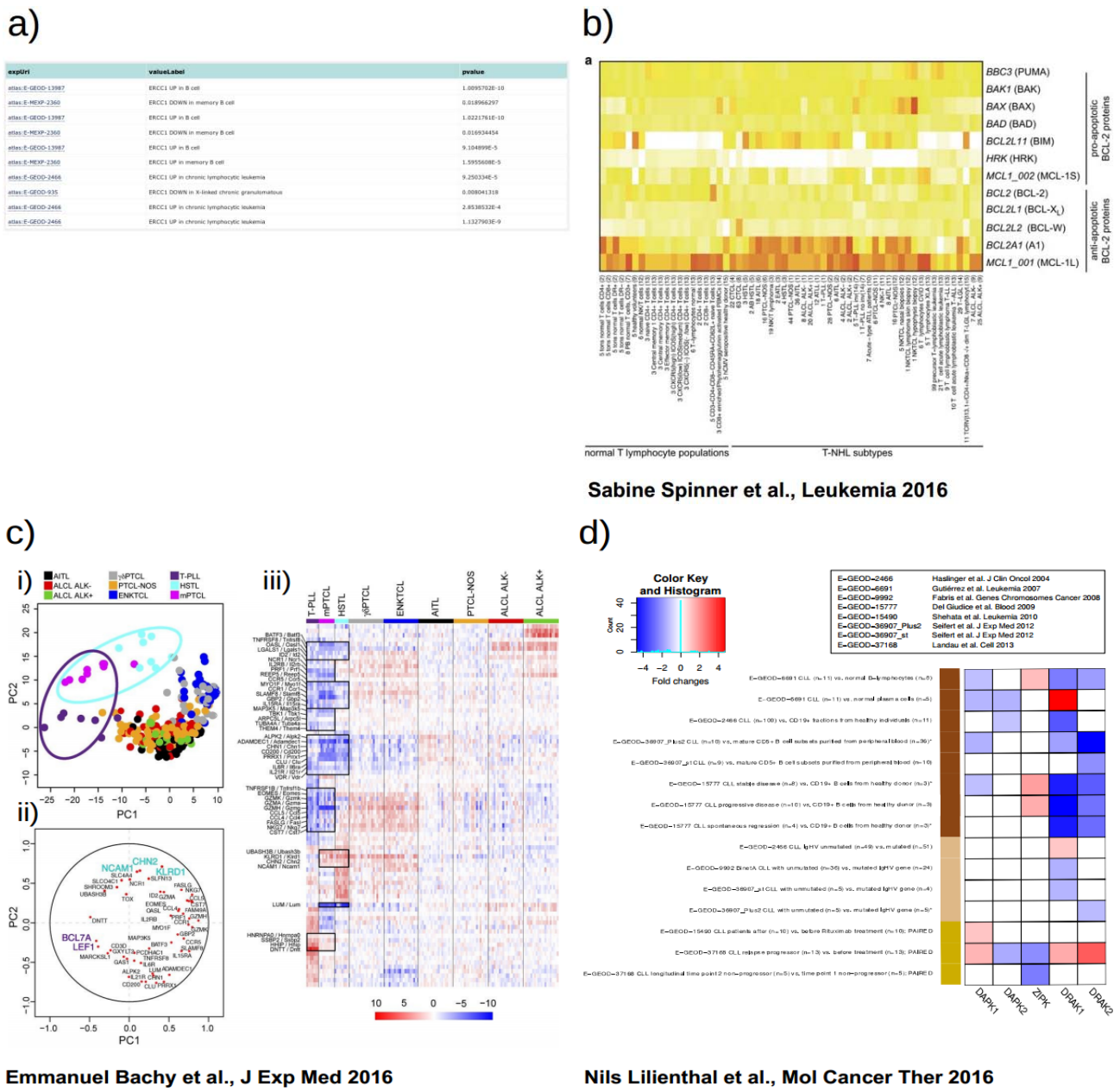


Figure 3. a) Potential ERCC1 deregulations in normals B-cells, B-cell lymphomas / leukemias (mantle-cell lymphoma, chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML)) and chronic conditions are queried within EMBL / EBI Gene Expression Atlas RDF (see Table 3 for exact query). The output, in table format, can be further exported into e.g. csv format. Fold-changes can be further visualized as in *c)*. **b)** Example taken from [49] (Fig. 1a): mature T-cell lymphomas and normal T-cell subsets are grouped by expression of pro- and anti-apoptotic *BCL2* family genes / isoforms. The long *MCL1* isoform seems to be used throughout malignant and benign T-cells, while *BCL2A1* and *BCL2L11* seem to be especially upregulated in malignant T-cells. Samples were quantile-normalized on the basis of 12 markers. **c)** Example taken from [50]. *i+iii)* illustrating different unsupervised clustering results (principal component analysis and heatmap) as *CD1d*-restricted murine

natural killer T-cell lymphoma seems to be most similar to T-cell prolymphocytic leukemia (T-PLL) and hepatosplenic T-cell lymphoma (HSTL). *ii)* Variables factor maps (produced by libraries like *FactoMineR*) show what marker contributes (or correlates) the most to each principal component and thus carries the highest specificity. Platform overlap was reduced to gene level, then batch-corrected using ComBat and quantile-normalized. **d)** Example taken from [51]. Fold-changes were calculated according to labeled comparisons for each *Death-Associated Protein Kinase (DAPK)* gene family member, then the range was cut off and results were visualized. Color bars used for 3 distinct comparisons: (1) CLL vs. normal B cells (various subtypes); (2) CLL with *IGHV* unmutated vs. mutated gene status; (3) CLL with post-to-pretreatment and other clinical comparisons.

9. Functional Analyses: the More the Merrier

In the abundance of genes obtained as significantly dysregulated, the role or function of a specific gene is often unknown and it is therefore encouraged to group them functionally by software tools often coined as ’pathway analysis‘ or ’enrichment‘ tools. One of the most user-friendly, however, costly tools is QIAGEN’s Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity/). Users can upload their differential expression results in the format of Excel tables into the Java GUI (graphical user interface). Annotation in the form of chip design or symbol identifiers (such as Gene Symbol, Ensembl ID or GenBank ID) can be selected for a given column as well as statistical parameters in separate columns, such as p-values, fold-changes, q-values / FDRs or simply expression values (fluorescence in microarrays or FPKM (fragments per kilobase of exon per million reads mapped) for RNA-seq). The list can be further restricted to a given range (e.g. p-value < 0.05). The selected genes are subsequently assembled into manually curated biological or toxicological / pharmacological pathways provided with an E-value (chance of a random hit). One advantage of IPA compared to other tools is the easy visualization of results by intuitive geometric forms, i.e. nodes / genes are drawn as distinct geometric symbols and edges / protein modifications in distinct line types. Similar graphs can be drawn with *igraph* in R, but are restricted to users that are more experienced in bioinformatics.

Other user-friendly and open-source alternatives include DAVID [26], gene set over-representation analysis (GSOA) by ConsensusPathDB [27] (Suppl. Figure 2), and gene set enrichment analysis (GSEA; Figure 4a) by the Broad Institute [28]. All three tools can be operated from web GUIs, while the first two options also offer an R implementation or in the case of GSEA, also a JAVA desktop application.

For more advanced users and those seeking to work with protein identifiers (complementary to above mentioned tools) *STRINGdb10* [29] is a potential alternative. Within the R library PPI (protein-protein interaction) graphs (nodes colored according to fold-change and also reachable via web link) and enrichments (including p-values and number of observed and expected interactions)

are calculated (Figure 4b; Code 15). Therein, inputs are the corresponding proteins of the most significantly dysregulated probes in different gene expression comparisons. Edges between proteins are colored according to evidence level, e.g. co-expression, literature mining, or experimental assays such as yeast2hybrid (y2h). The same R library can also be used for KEGG and GO (gene ontology) enrichment analyses (Code 16). RNA-to-protein inference can however only be approximate due to different half-lifes and decay rates as well as due to variable post-transcriptional and post-translational modifications.

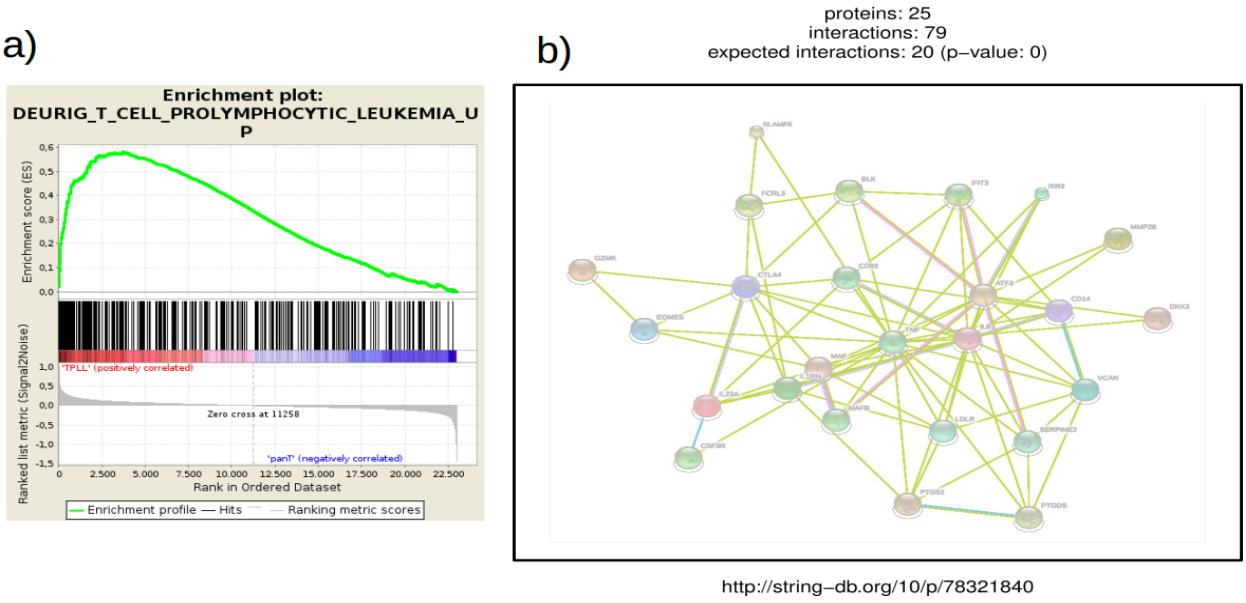


Figure 4. Results of differential expression analysis of 70 samples of T-cell prolymphocytic leukemia (T-PLL) and normal CD3+ T-cells from 10 healthy donors were further functionally annotated. a) Enrichment plot of Broad GSEA (gene set enrichment analysis) of the most deregulated (|fc|>1.5; $q < 0.05$) genes between T-PLL and normal CD3+ T-cells shows strong correlation (hit accumulation at the front of enrichment profile in dark and peak in green) to the results of a previous T-PLL gene expression data set [52]. **b)** Example of a PPI (protein-protein interaction) graph output from STRINGdb_v10 with a significant enrichment (59 more PPIs than expected). URL at the bottom is automatically generated and serves as an archive for the output.

10. Standard Survival Analysis and An Exploratory / Heuristic Approach

Besides parameters of more established nature (routinely tested), e.g. in CLL those from clinical chemistry, such as β_2 microglobulin [30] or from immunophenotyping, such as ZAP70 [31], the expression of a single gene or a gene set detected by microarray-based GEP can also serve as a marker, or a scored combination of them, that predict clinical outcomes. Such prognostic estimations are predominantly measured in subgroup differences of time-to-event metrics like overall survival (OS; from date of diagnosis or less correctly from first day of treatment or study randomization to last follow-up (FU) or death) or progression-free survival (PFS; from first day of treatment or randomization to disease progression or death). Other measurements include time-to-treatment (TTT; from diagnosis or randomization to first day of treatment), time-to-next-treatment (TTNT; end of first to beginning of next treatment), time-to-treatment-failure (TTF; time from diagnosis or randomization to treatment dismissal), or event-free survival (EFS; time from diagnosis or randomization to disease progression, death or treatment dismissal). These parameters are either right-censored (date of death or progression after study window, thus unknown) or left-censored (study entry is unknown) to deal with missing time points or events (death or progression). Here we focus on right-censored data.

An univariate analysis compares time-to-event parameters for two subgroups divided by a gene expression or other marker status (see [32] for an introduction). For multivariate analysis, multiple genes or markers are considered for a competing subset comparison (see [33] for an introduction). For the former there are standard methods implemented within the R library *survival* with functions *survdiff* to test the differences of survival times with the log-rank test [34] and *survfit* to plot the survival times with the Kaplan-Meier estimator [35] (Code 17). A multivariate analysis allows ranking of the most significant markers contributing to an adverse prognosis. It is usually conducted with the Cox Proportional Hazards [36] (CoxPH) model.

As evidence provided by different data sources and methods strengthens a given hypothesis, it is important to validate identified markers of prognosis in an independent patient cohort. However, this is often difficult due to a limited availability of reasonably-sized data sets for comparison. Possible causes may be a low disease incidence (e.g. notorious for mature T-cell lymphomas) or general difficulties in obtaining primary tumor samples (e.g. due to the need of invasive procedures to be consented by the patient). Another factor imposing limitations on sample size is the uniformity of received treatments, which must apply to a given patient cohort in order to reliably predict related outcomes. For GEP studies in such scenarios, we propose an alternative algorithm for the identification of prognostic gene expression signatures, which we demonstrate by the example of GEP data generated from peripheral blood tumor samples of patients with T-cell prolymphocytic leukemia (T-PLL) and CLL. We obtained gene expression profiles from 49 T-PLL samples with available OS status and from 58 chemoimmunotherapy-treated CLL patients with available PFS data, both from Illumina HumanHT-12 v4.0 Expression BeadChips.

In a first training set of 10 T-PLL, 5 patients with longest OS (time from diagnosis to death of disease, > 800 days) were compared to those with shortest OS (< 300 days, n = 5) using the 'Significance analysis of microarrays' (SAM) analysis in survival mode via the R library *samr* [37]. We only considered expression profiles from patients in whom corresponding samples had been obtained within 6 months from diagnosis (ensuring similarities between specimen and clinical data) and who had presented with similar lymphocyte doubling times as an indicator of disease kinetics at the time of sample. From an initial most informative index-set of 5 differentially expressed probes (*RAB25*, *KIAA1211L*-probe1, *KIAA1211L*-probe2, *GIMAP6*, *FXYD2;* FDR < 0.1), linear regression [38] and removal of one outlier by setting OS<200 days, resulted in a 2nd training set of nine cases. Another subsequent SAM (survival mode) resulted in a 2-gene / 3-probe set as the most robust combined predictor of OS. These probe sets were used to calculate an expression index via an additive model fit using Tukey's median polish procedure [39] (*medpolish* function within the standard *stats* library) on a test set of 40 uniformly treated T-PLL (the nine training cases excluded) fulfilling the criteria of available array data and OS information. Kaplan-Meier curves (log-rank tests for differences) were created based on stratified per patient-values of this "2-gene / 3-probe prognostic expression index" (*RAB25* and the two *KIAA1211LL* transcripts either merged or separated; Figure 5a). Ranking the cases solely based on these expression indices, the five T-PLL cases with the lowest values indeed showed significantly superior OS over those five cases with highest or 35 cases with higher (Figure 5b; Suppl. Figure 3a) expression index values (index fold-change (fc) = −2.37; Figure 5b; index fc = −1.62; Suppl. Figure 3a). A similar approach was used to identify signature genes associated with PFS in chemoimmunotherapy-treated CLL (Figure 5c; Suppl. Figure 3b; Code 18) resulting in a predictive 4-gene / 7-probe index (including *GPD1L, TNFSF12*, *JHDM1D, TBCD, AARS2, MTG1*, and *TNIP*). In both cohorts, the detected differential expression of signature genes and their association with clinical outcome requires further validation, e.g. by qRT-PCR, in independent samples before considering them further as valid markers.

11. Sample Classification by Supervised (Machine Learning) Approaches

When dealing with large data sets (e.g. a gene expression matrix) that incorporate different clinical or molecular information ('features'), and if a group status ('class') of clinical or biological interest (e.g. treatment responder vs. non-responder) is known, the application of discrimination (or supervised learning) methods can be considered. Such methods aim to train classifiers (logistic, linear, or non-linear) that are able to predict the status of future samples based on certain features (e.g. treatment response). In general, it is important to validate classification rules obtained from training data in an independent test set, preferably obtained from another set of patients from a different laboratory / trial group, in order to avoid a biased data interpretation. When there is no independent set available, an internal cross-validation can be performed. Therein, the available patient samples are repeatedly separated into a training set and a test set, while subsequently

observing the average classification performance by the number of false positives and false negatives obtained through the classifier.

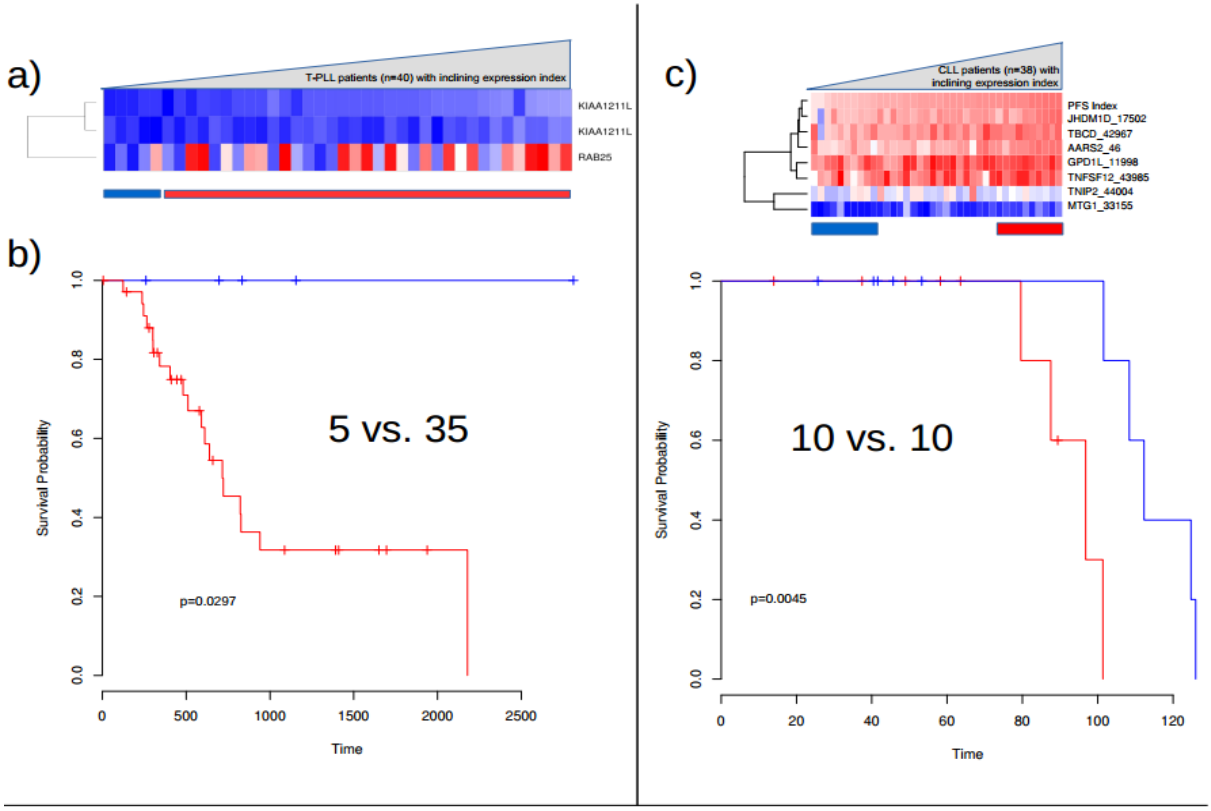


Figure 5. We explored alternative approaches to obtain prognostic values in a 49-case cohort of T-cell prolymphocytic leukemia (T-PLL) (Schrader, Crispatzu et al. submitted) with available overall survival (OS) data as well as in a chemoimmunotherapy-treated cohort of chronic lymphocytic leukemia (CLL) (Herling et al. unpublished; *n* = 58 with available progression-free survival (PFS) status). **a-b)** The five T-PLL patients with each the highest and lowest OS (without censored / alive ones) were considered for a 'Significance analysis of microarrays' (SAM) analysis in survival mode. The resulting probe sets / transcripts were used to calculate an expression index **a)** (via additive model fit using Tukey's median polish procedure) on the test set of residual cases. Kaplan-Meier (log rank; time in days) curves were created based on stratified values per patient of this 'prognostic expression index'. **b)** Five patients with lowest index expression vs. residual 35 patients of test set (see Suppl. Figure 3 for 5 vs. 5). **c)** The same approach was used for ten chemoimmunotherapy-treated CLL with the highest and lowest PFS. The index was calculated on probe set / transcript level and again evaluated in especially indolent and aggressive patient samples (here ten with lowest and highest index expression) within the test set. In both cohorts, of T-PLL and CLL, a high index expression was linked to an adverse prognosis.

A popular supervised learning approach are support vector machines [40] (SVM; R libraries *gmum.r* or *e1071*). They try to separate classes by projecting features and their interactions into high-dimensional space and subsequently by searching for either linear (Figure 6a-b) or non-linear (Figure 6c; Suppl. Figure 4) separating hyperplanes in the original feature space (Code 19).

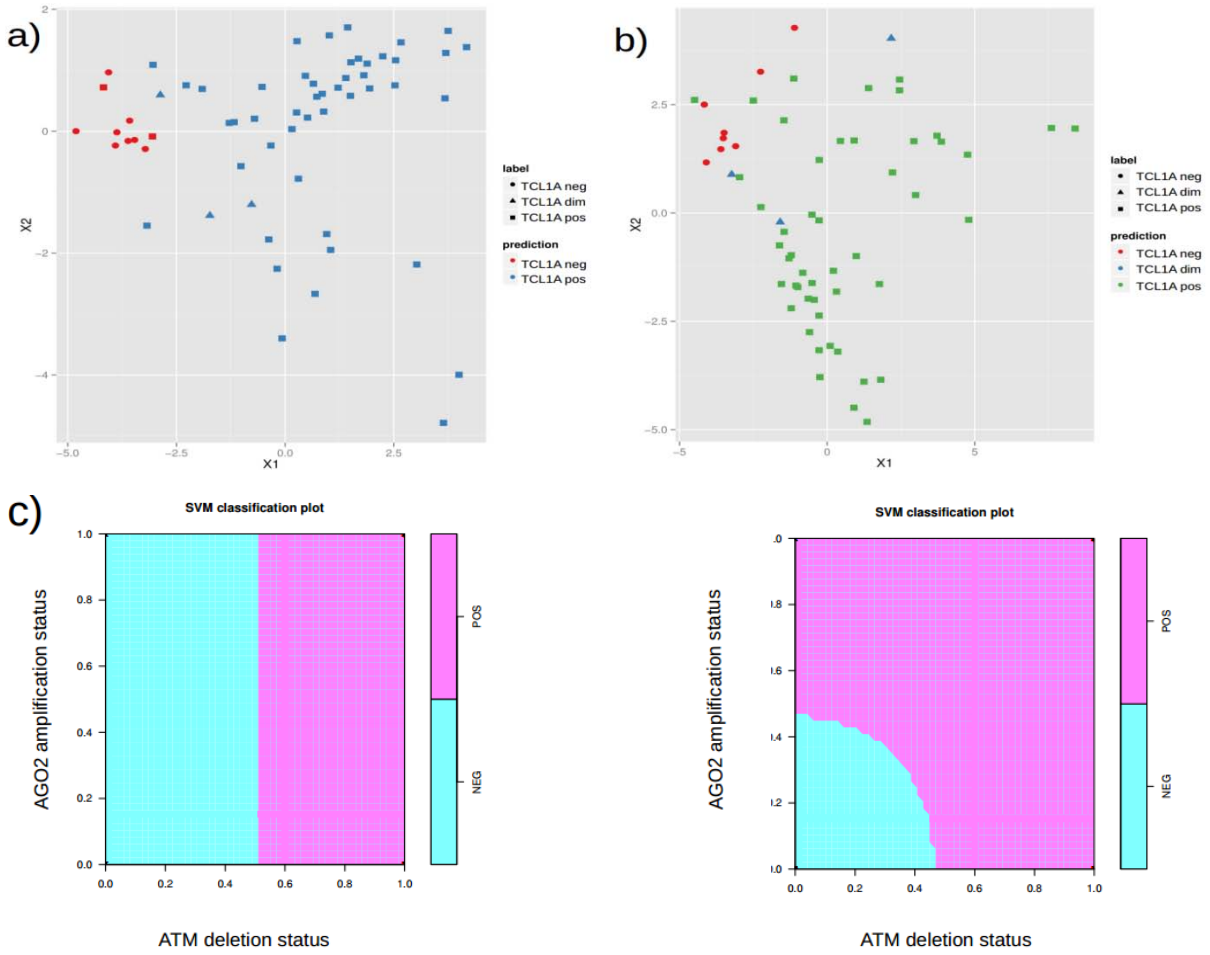
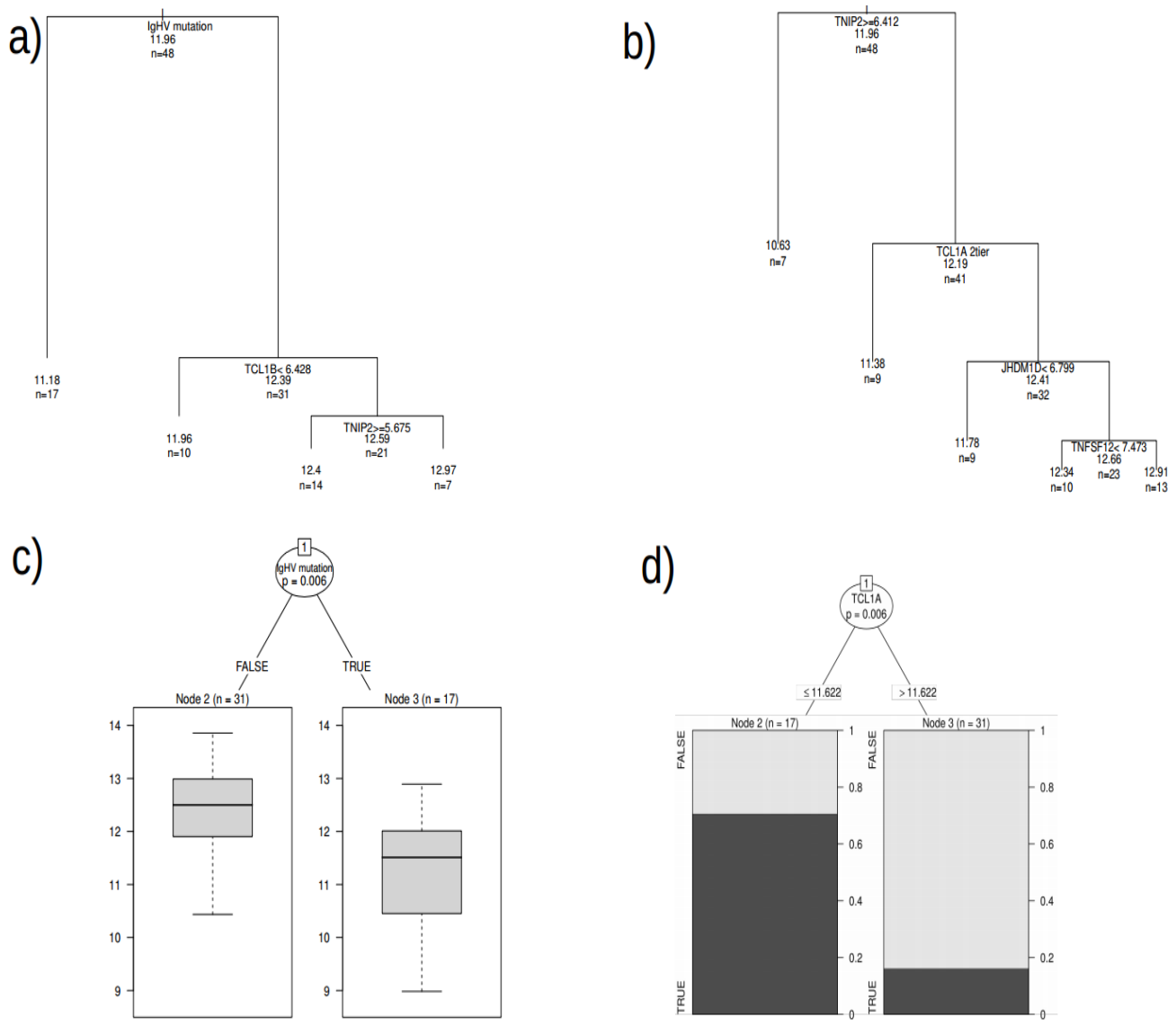


Figure 6. a) Support vector machine (SVM) classifies samples of T-cell prolymphocytic leukemia (T-PLL) based on TCL1A protein status (positive, intermediate, negative; by flow-cytometry) predicted by *TCL1A* and *TCL1B* mRNA expression. As one can see in the top left two samples are misclassified by SVM as TCL1A-negative (red, but squared symbols). **b)** SVM of T-PLL samples of different TCL1A protein status ("dim" being intermediate) by numerous mRNA markers performs more robust classification. **c)** Example of a linear (**upper panel**) and a non-linear, radial / polynomial fit (**lower panel**) of a SVM. T-PLL samples which carry the *ATM* gene in mutated vs. unmutated constitution are classified by their status of *ATM* deletion and *AGO2* amplification. Results, as seen by approximate pattern in linear and more distinct pattern in non-linear classifier, elucidating that *ATM* unmutated samples are more likely to be biallelic for *ATM* and *AGO2*.

Decision trees (R libraries *rpart, tree* or *party;* Code 20) can also divide samples according to a class variable into further most informative binary portions of gene expression signatures (Figure 7a–b) or of other molecular features (i.e. mutational or cytogenetic strata in CLL) (Figure 7c–f; Suppl. Figure 5); measured by ANOVA for numerical or by entropy for categorical values. When looking for a cut-off for adverse prognosis, they can be further used in the form of regression trees [41]. Different parameters can be controlled in this approach, such as the maximum size of a tree or the number of portions / bins. It is recommended to keep these relatively low in the training set to avoid "overfitting" and thus enable re-evaluation in the test set. Random forests [42] (as an assembly of permutated decision trees) can be used to determine the chance of observing random tree branching (library *randomForest*) (Code 21). Both algorithms are also included in the *rattle* library, which offers a user-friendly GUI with interactive plots and a selection menu for class variable and co-variates as well as algorithm and parameter choices. For a more detailed review on current machine learning algorithms in GEP, we refer to [43].



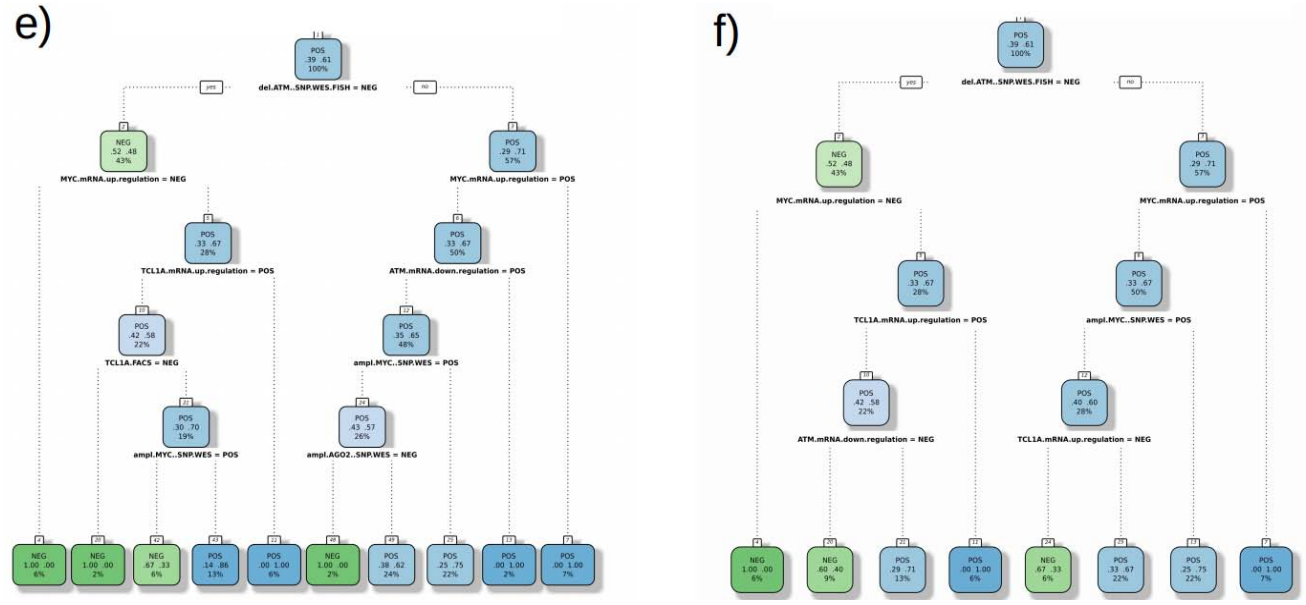


Figure 7. a) Example of a *rpart* decision tree. Chronic lymphocytic leukemia (CLL) samples are stratified according to TCL1A protein status. Model design then includes *IGHV* gene mutation status, mRNA markers linked to adverse prognosis (from algorithm described in **Figure 5c**), and further clinical features. *IGHV* gene mutations status, as seen in top of branch, is the most informative divider. When left out in **b)** an mRNA marker linked to adverse prognosis (with somewhat arbitrary cut-off for illustrative purposes) functions as the most informative divider. **c-e)** *ctree* offers more intuitive visualizations of decision trees. **c)** When stratifying CLL samples by *TCL1A* mRNA expression, *IGHV* mutations status is the most informative divider. **d)** This is confirmed when stratifying CLL samples by *IGHV* mutations status (switching the comparison) hence *TCL1A* mRNA expression is the most informative discriminator. **e)** T-cell prolymphocytic leukemia (T-PLL) samples stratified by *ATM* mutation status. Co-variates include *ATM* deletion, *miR-34B* deletion, *MYC* amplification, *AGO2* amplification, *MYC* mRNA upregulation, *ATM* mRNA downregulation, and *TCL1A* mRNA upregulation. *ATM* deletion status seems to be the most informative co-variate, however due to the excessive size of the tree (controlled by pruning and number of bins) there is a risk of "overfitting". **f)** Shown is a more feasible and smaller decision tree. Again, the most informative co-variate seems to be the status of *ATM* gene deletion. Followed by *AGO2* amplification status. This is further confirmed in random forests (permutated decision trees) in order to circumvent 'overfitting' (not shown).

12. Discussion

In this review we discuss procedures to optimize GEP analyses. We highlight the importance of advanced preprocessing, such as batch correction and admixture modeling, but also appraise the versatility and sophistication of analysis and classification algorithms. Many of the presented

methods, originally established for microarray data analysis, can also be applied to RNA-seq data (on the basis of read counts instead of fluorescence values). In addition to GEP, it is always desirable to aim for additional genetic information, including (somatic) copy-number alterations, structural variation, and genotyping of nucleotide variants for a most comprehensive genetic workup of the investigated cancer specimen. Epigenomic data, e.g. from methylome and ChIP-seq experiments may be added as a second layer. Besides setting up an own data repository in MySQL or RDF for managing internal data, one may also investigate the cBioPortal for Cancer Genomics [44]. TCGA (https://tcga-data.nci.nih.gov/tcga), ICGC (https://dcc.icgc.org), and other large curated data sets provide user-friendly search engines with multiple visualization options. Another helpful tool for combining gene expression data with available genomic knowledge in a network-based analysis is Expander [45]. Overall, this review and the attached source codes may provide guidance to both molecular biologists and bioinformaticians / biostatisticians to properly conduct GEP analyses from microarrays and to go beyond the application of standard analytic tools to optimally interpret the clinical and biological relevance of the obtained results.

Acknowledgements

M.H. (HE3553/3-1) and C.D.H. (SCHW1711/1-1) are funded by the German Research Foundation (DFG) as part of the collaborative research group on “Exploiting the DNA damage response in CLL” (KFO286). M.H. (HE3553/4-1), has been supported by the DFG as part of the collaborative research group on mature T-cell lymphomas “CONTROL-T” (FOR1961). Further support: German Cancer Aid (108029), CECAD, José Carreras Leukemia Foundation (R12/08) (all to M.H.); CLL Global Research Foundation (to M.H. and C.D.H.); Köln Fortune Program and Fritz Thyssen foundation (10.15.2.034MN) (both to M.H. and A.S.).

We gratefully acknowledge all contributing centers and investigators enrolling patients into the trials and registry of the German CLL Study Group (GCLLSG) and at the UT M.D. Anderson Cancer Center (MDACC), Houston/TX, USA; the GCLLSG and UT MDACC staff and the patients with their families for their invaluable contributions.

Contribution of Authors

Data analysis: G.C.; survival analyses: G.C., A.S., M.H.; experiments and conduction of GEP: A.S., C.D.H.; clinical data: M.H., C.D.H.; manuscript preparation: G.C., C.D.H., M.N., M.H.

Conflicts of Interest Disclosure

There were no competing interests interfering with the unbiased conduction of this study.

Patient Samples

Human tumor samples were obtained from patients under IRB-approved protocols following written informed consent according to the Declaration of Helsinki. Collection and use have been approved for research purposes by the ethics committee of the University Hospital of Cologne (#11-319) and UT M.D. Anderson Cancer Research Center. The cohorts were selected based on uniform front-line treatment as part of the TPLL1 [46] (NCT00278213) and TPLL2 (NCT01186640, *unpublished*) prospective clinical trials as well as FCR300 [47] or included in the nation-wide T-PLL and CLL registries of the German CLL Study Group (GCLLSG, IRB# 12-146).

References

1. International Human Genome Mapping Consortium. A physical map of the human genome. (2001) *Nature* 409: 934-941.
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. (2004) *Nature* 431: 931-945.
3. Ferreira PG, Jares P, Rico D, et al. (2014) Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res* 24: 212-226.
4. Ojesina AI, Lichtenstein L, Freeman SS, et al. (2014) Landscape of genomic alterations in cervical carcinomas. *Nature* 506: 371-375.
5. Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507:315-322.
6. Wang C, McKeithan TW, Gong Q, et al. (2015) IDH2R172 mutations define a unique subgroup of patients with angioimmunoblastic T-cell lymphoma. *Blood* 126: 1741-1752.
7. R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.
8. Gentleman RC, Carey VJ, Bates DM, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80. URL: http://www.bioconductor.org/.
9. Irizarry RA, Bolstad BM, Collin F, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.
10. Yang YH, Dudoit S, Luu P, et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15.
11. Huber W, von Heydebreck A, Sueltmann H, et al. (2002) Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* 18: S96-S104.

12. Durinck S, Moreau Y, Kasprzyk A, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21: 3439-3440.
13. Bae J, Leo CP, Hsu SY, et al. (2000) MCL-1S, a splicing variant of the antiapoptotic BCL-2 family member MCL-1, encode s a proapoptotic protein possessing only the BH3 domain. J *Biol Chem* 275: 25255-25261.
14. Noble WS (2009) How does multiple testing correction work? *Nat Biotechnol* 27: 1135-1137.
15. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Statistical Sci* 18: 71.
16. Pollard KS, Dudoit S, van der Laan MJ (2005) Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer.
17. Kitchen RR, Sabine VS, Sims AH, et al. (2010) Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. *BMC Genomics* 11: 134.
18. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118-127.
19. Yoshihara K, Shahmoradgoli M, Martínez E, et al. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 4: 2612.
20. Gaujoux R, Seoighe C (2013) CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* 29: 2211-2212.
21. Barrett T, Wilhite SE, Ledoux P, et al. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41: D991-995.
22. Kolesnikov N, Hastings E, Keays M, et al. (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43: D1113-1116.
23. Petryszak R, Keays M, Tang YA, et al. (2016) Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* 44: D746-752.
24. Malone J, Holloway E, Adamusiak T, et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26: 1112-1118.
25. Taminau J, Meganck S, Lazar C, et al. (2012) Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC Bioinformatics* 13: 335.
26. Dennis Jr G, Sherman BT, Hosack DA, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4: P3.
27. Kamburov A, Stelzl U, Lehrach H, et al. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41; D793-800.
28. Subramanian A, Tamayo P, Mootha VK, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U. S. A.* 102: 15545-15550.

29. Szklarczyk D, Franceschini A, Wyder S, et al. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: D447-452.
30. Gentile M, Cutrona G, Neri A, et al. (2009) Predictive value of beta2-microglobulin (beta2-m) levels in chronic lymphocytic leukemia since Binet A stages. *Haematologica* 94: 887-888.
31. Wiestner A, Rosenwald A, Barry TS, et al. (2003) ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood* 101: 4944-4951.
32. Clark TG, Bradburn MJ, Love SB, et al. (2003) Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 89:232-238.
33. Bradburn MJ, Clark TG, Love SB, et al. (2003) Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer* 89: 431-436.
34. Peto R, Pike MC, Armitage P, et al. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 35: 1-39
35. Kaplan EL, MeierP (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53: 457-481
36. Cox DR. (1972) Regression models and life tables (with discussion). *JR Statist Soc B*: 34187-34220.
37. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98: 5116-5121.
38. Chen K (2001) Generalized case-cohort sampling. *Statistical Methodol* 63: 791-809.
39. Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley.
40. Vapnik VN (1995) The Nature of Statistical Learning Theory. Berlin: Springer
41. Breiman L, Friedman J, Stone CJ, et al. (1984) Classification and Regression Trees. Chapman and Hall/CRC.
42. Ho TK (1995) Random Decision Forests Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August pp. 278-282.
43. Hahne F, Huber W, Gentleman R, et al. (2008) Bioconductor Case Studies. Springer Press.
44. Gao J, Aksoy BA, Dogrusoz U, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6: l1.
45. Ulitsky I, Maron-Katz A, Shavit S, et al. (2010) Expander: from expression microarrays to networks and functions. *Nature Protocols* 5: 303-322.
46. Hopfinger G, Busch R, Pflug N, et al. (2013) Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia. *Cancer* 119: 2258–2267.
47. Keating MJ, O'Brien S, Albitar M, et al. (2005) Early results of a chemoimmunotherapy regimen of fludarabine, cyclophosphamide, and rituximab as initial therapy for chronic lymphocytic leukemia. *J Clin Oncol* 23: 4079-4088.

48. Sotiriou C, Wirapati P, Loi S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98: 262-272.
49. Spinner S, Crispatzu G, Yi JH, et al. (2016) Re-activation of mitochondrial apoptosis inhibits T-cell lymphoma survival and treatment resistance. Leukemia. Mar 8.
50. Bachy E, Urb M, Chandra S, et al. (2016) CD1d-restricted peripheral T cell lymphoma in mice and humans. *J Exp Med* 213: 841-857.
51. Lilienthal N, Lohmann G, Crispatzu G, et al. (2016) A Novel Recombinant Anti-CD22 Immunokinase Delivers Proapoptotic Activity of Death-Associated Protein Kinase (DAPK) and Mediates Cytotoxicity in Neoplastic B Cells. *Mol Cancer Ther* 29.
52. Dürig J, Bug S, Klein-Hitpass L, et al. (2007) Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia* 21: 2153-2163.

AIMS Press

© 2016 Giuliano Crispatzu et al., licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)