Chapter 1
Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking

Franco A. Ubaudi, Paul J. Kennedy, Daniel R. Catchpoole, Dachuan Guo, and Simeon J. Simoff

Abstract Gene expression datasets used in biomedical data mining frequently have two characteristics: they have many thousand attributes but only relatively few sample points and the measurements are noisy. In other words, individual expression measurements may be untrustworthy. Gene Feature Ranking (GFR) is a feature selection methodology that addresses these domain specific characteristics by selecting features (i.e. genes) based on two criteria: (i) how well the gene can discriminate between classes of patient and (ii) the trustworthiness of the microarray data associated with the gene. An example from the pediatric cancer domain demonstrates the use of GFR and compares its performance with a feature selection method that does not explicitly address the trustworthiness of the underlying data.

1.1 Introduction

The ability to measure gene activity on a large scale, through the use of microarray technology [1, 2], provides enormous potential within the disciplines of cellular biology and medical science. Microarrays consist of large collections of cloned molecules of deoxyribonucleic acid (DNA) or DNA derivatives distributed and bound in an ordered fashion onto a solid support. Each individual DNA clone represents a particular gene. Several kinds of microarray technology are available to researchers to measure levels of gene expression: cDNA, Affymetrix and Illumina

Franco A. Ubaudi, Paul J. Kennedy
Faculty of IT, University of Technology, Sydney, e-mail: {faubaudi,paulk}@it.uts.edu.au

Daniel R. Catchpoole, Dachuan Guo
Tumour Bank, The Childrens Hospital at Westmead, e-mail: {DanielC,dachuang}@chw.edu.au

Simeon J. Simoff
University of Western Sydney, e-mail: S.Simoff@uws.edu.au

microarrays. Each of these platforms can have over 30,000 individual DNA features "spotted" in a cm^2 area. These anchored DNA molecules can then, through the process of "hybridization", capture complementary nucleic acid sequences isolated from a biological sample and applied to the microarray. The isolated nucleic acids are labeled with special fluorescent dyes which, when hybridized to their complementary spot on the microarray will fluoresce when excited with a laser–based scanner. The level of fluorescence is directly proportional to the amount of nucleic acid captured by the anchored DNA molecules. In the case of gene expression microarrays, the nucleic acid isolated is messenger ribonucleic acid (mRNA) which results when a stretch of DNA containing a gene is "transcribed" or is "expressed".

This chapter focuses on two color spotted cDNA microarrays, an approach which allows for the direct comparison of two samples, usually a control and a test sample. Data is derived from fluorescent images of hybridized microarray "chips" and usually comprises statistical measures of populations of pixel intensities of each gene feature. For example, mean, median and standard deviation of pixel intensities for two dyes and for the local background are generated for each spot on a cDNA microarray.

All microarray platforms involve assessing images of fluorescent features with the level of fluorescence giving a measure of expression. All platforms are beset by the same data analysis issues of feature selection, noise, high dimensionality, non–specific signal, background, and spatial effects. The approach we take in this chapter is non–platform specific, although we apply it to the "noisiest" of the platforms: glass slide spotted cDNA microarray. Comparing gene expression measurements between different technologies and between measurements on the same technology at different times is a challenge, to some extent addressed by normalization techniques [3]. A major issue in these data is the unreliable variance estimation, complicated by the intensity-dependent technology-specific variance [4]. There is also an agreement that different methods of microarray data normalization have a strong effect on the outcomes of the analysis steps [5]. Another issue is the small number of replicated microarrays because of cost and sample availability, resulting in unreliable variance estimation and thus unreliable statistical hypothesis tests. There has been a broad spectrum of proposed methods for determining the sample size in microarray data collection [6], with the majority being focused on how to deal with multiple testing problems and the calculation of significance level.

Common approaches to dealing with these data issues include visual identification of malformed spots for omission and normalization of gene expression measures [7]. Often arbitrary cut off measures are used to select genes for further assessment in attempts to reduce the data volume and to facilitate selection of "interesting genes". Other researchers (e.g. [8, 9, 10, 11, 12]) apply dimensionality reduction to microarray data with the assumption that many genes are redundant or irrelevant to modeling a particular biological problem. Golub [8] calls the redundancy of terms *additive linearity* to signify that many genes add nothing new or different. Feature set reduction of microarray data also helps to deal with the "curse of dimensionality" [13]. The view of John et al [14] is that use of feature selection, prior to model

construction, obviates the necessity of relying on learning algorithms to determine which features are relevant and has the additional benefit of avoiding overfitting.

Given the widespread use of feature selection for microarray data and the data quality issues, it is surprising that we were unable to find approaches in the literature specifically addressing data quality [15]. This is also true of more general feature selection literature [16]. Some researchers have endeavored to manage the issues of small sample size and microarray noise in their approaches. Unsupervised methods are used to evaluate the quality of microarray data in [17] and [18] but they are subjective and need manual configuration. Baldi and Hatfield [7] use variance of expression as prior knowledge when training a Bayesian network. Yu et al [19] apply an approach they call "quality" to filter genes, although it only uses expression data.

The feature analysis and ranking method proposed in this chapter addresses the issue of dealing with diverse quality across technologies and the small number of replicated measurements. Our approach explicitly uses quality measures of a spot (such as variance of pixel intensities) to compute a trustworthiness measure for each gene which complements its gene expression measure. Understanding of gene expression can then be based on the quality of the spot as well as its intensity. A "confidence" of the findings based on data quality can be incorporated into models built on the data. Also, training sets with few sample points, as are often found in gene expression may also mean that the assumption that test data has similar quality as training sets is no longer valid. The unsupervised learning we apply to all available data helps to gain an understanding of the quality of gene expression measurements.

1.2 Gene Feature Ranking

Gene Feature Ranking is a feature selection methodology for microarray data which selects features (genes) based on (i) how well the gene discriminates between classes of patient and (ii) the trustworthiness of the data. The motivation behind the first of these is straightforward. However, previous methods have not specifically addressed the issue of the quality of data in feature set reduction. Hence, our emphasis on assessing the trustworthiness of the data. A training subset of data is used to assess classification of patients by genes. All available data is used in an unsupervised learning process to assess the trustworthiness of gene measurements.

Gene Feature Ranking, shown schematically in Fig. 1.1, consists of two consecutive feature selection phases which generate ranked lists of features. Phase 1 ranks genes by the trustworthiness of their data. Genes ranked above a threshold are passed to the second phase. Phase 2 ranks the remaining genes using a traditional feature selection algorithm (such as Gain Ratio [20]). The goal of the approach is to maintain a balance between the trustworthiness of data and its discriminative ability. Following we describe how attributes and data sample points are used in GFR and then we describe the two phases in detail. Although data preprocessing methods which suppress genes that do not respect some measures may be seen as an alternative to GFR, such approaches have the limitation that they are generally ad

hoc and usually use expression measurements to assess the spot quality rather than "quality measures". Gene Feature Ranking, on the other hand, is a framework that specifically addresses the quality aspects of the data.

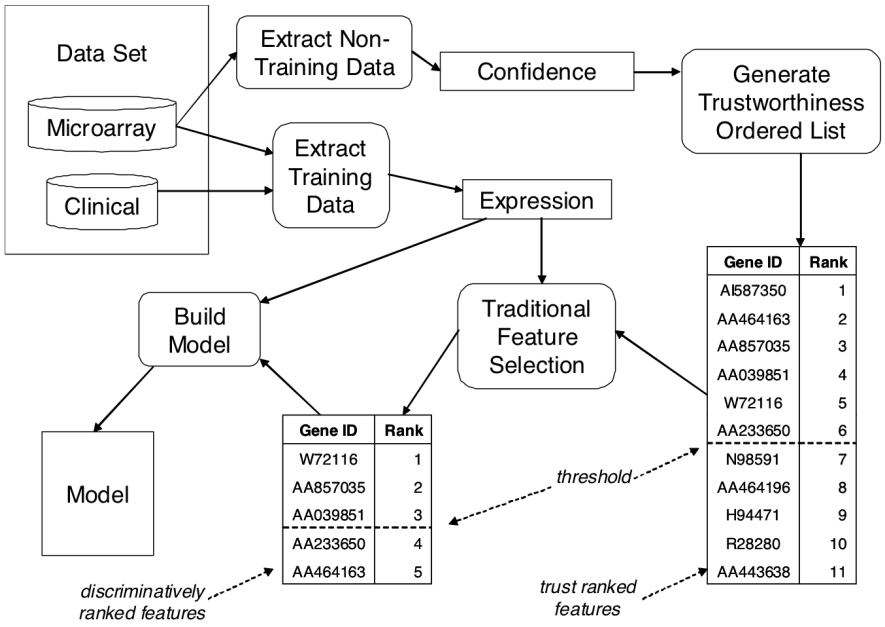

Fig. 1.1 Gene Feature Ranking applies two phases of feature selection to data. Phase 1, labelled "Generate Trustworthiness Ordered List", filters genes based on the trustworthiness of their data, as measured by the "confidence". Phase 2, labelled as "Traditional Feature Selection" filters the trusted genes based on how well they can discriminate classes.

1.2.1 Use of attributes and data samples in Gene Feature Ranking

Attributes associated with individual genes in the microarray data are grouped as either "expression" attributes or "spot quality" attributes, but not both. Expression attributes are considered to measure the level of expression of a gene. Spot quality attributes are statistical measures of likely accuracy of the expression for the gene. An example of an expression attribute is median pixel intensity of the spot and an example of a spot quality attribute is the standard deviation of pixel intensities for a spot. Expression attributes of genes are used to build the model and spot quality attributes are used to assess the trustworthiness (or confidence) of the gene.

Three subsets of available data are recognized in GFR. *Training Data Set* consists of "expression" attributes from the microarray gene expression data together

with associated clinical data. The purpose of this data is to build the model and it includes the class of patient. *Test Data Set* is a smaller set of "expression" attributes and clinical data used to evaluate the accuracy of the model built from the training data. *Non–Training Data Set* consists of "spot quality" attributes from all of the microarray gene expression data. This data is used in the first (unsupervised) phase of GFR. As this data is not used directly to build a model it does not contain information about the patient class or any clinical attributes.

1.2.2 Gene Feature Ranking: feature selection phase 1

Phase 1 of GFR determines the trustworthiness of genes by evaluating the confidence of each expression reading (labelled *confidence* in Fig. 1.1) in the non–training dataset. Trustworthiness of a gene is the median value of the N confidence values. The confidence value c for a gene in a specific microarray spot is

$$c = \frac{1}{\log\left(\frac{(\sigma_{control}+\sigma_{test})}{2} + \|\sigma_{control} - \sigma_{test}\|\right)}. \tag{1.1}$$

The spot quality attributes $\sigma_{control}$ and σ_{test} are the standard deviations of pixel intensity for the $control$ and $test$ channels respectively. High pixel deviation is associated with low confidence in the accuracy of expression measure. As cDNA microarrays have two intensity channels per spot, the definition of c in equation (1.1) comprises an average of the variations of the channels as well as a measure of their difference.

The genes are then ranked by trustworthiness with those below a threshold judged too noisy and then discarded. Gene Feature Ranking does not *a priori* prescribe the threshold. Two issues contribute to setting a threshold: (i) an understanding of the redundancy present in the attributes is needed to avoid setting the threshold too high; and (ii) analysis of the distribution of trustworthiness values for genes is needed to prevent setting the threshold too low. We set the threshold empirically, although an approach based on the calculated trustworthiness values would also be appropriate.

1.2.3 Gene Feature Ranking: feature selection phase 2

Phase 2 of GFR ranks the remaining genes according to their discriminative ability using the training data set (as shown in Fig. 1.1). All genes passed through from phase 1 are considered to have high enough trust for use in phase 2. Genes ranked above a phase 2 threshold are selected to build a classifier. The feature selection method in phase 2 is not prescribed. Here, we use Gain Ratio. The result is a list of genes that are discriminative and trustworthy.

As in the first feature selection phase, choice of the selection threshold is the responsibility of the user. In this chapter, we choose several final feature subset

sizes to compare the achieved classification accuracy of models constructed using different feature selection methodologies. Two other possible approaches might use a measure of the minimum benefit provided for each feature with the following metrics: (i) the classification accuracy gained by the model; or (ii) a balance between trustworthiness and discriminative ability.

1.3 Application of Gene Feature Ranking to Acute Lymphoblastic Leukemia data

This section applies GFR to the subtyping of Acute Lymphoblastic Leukemia (ALL) patients. We describe the ALL data then build classifiers using GFR ranked genes and Gain Ratio only ranked genes to classify patients suffering from two subtypes of ALL: B lineage and T lineage.

Biomedical data was provided by the Children's Hospital at Westmead and comprises clinical and microarray datasets linked by sample identifier. Clinical attributes include sample ID, date of diagnosis, date of death and immunophenotype (the class label). The microarray dataset has several attributes for each gene for each array: f635Median, f635Mean, f635SD, f532Median, f532Mean and f532SD. The median and mean measurements are "expression" attributes and the standard deviation attribute is the "spot quality". These relate to median, mean and standard deviation of the pixel intensities for the fluorescent emission wavelength of the two dyes: 635 nm (red) and 532 nm (green). The 635 nm fluorescent dye was associated with leukemia samples and the 532 nm dye represented pooled normal (non–leukemic) samples. Consequently, this study compared on the one array ALL to normal control.

Data was preprocessed. Patients with missing data and genes not present on all microarrays and "non–biological" genes were omitted. Outlier expression measures that were close to the minimum or maximum measurable value were repaired to conform with other data. Microarray data was normalized by adjusting arrays to have the same average intensity. Preprocessing resulted in 120 arrays (47 patients) for B ALL and 44 arrays (14 patients) for T ALL with a total of 9485 genes.

Algorithms other than GFR are implemented in Weka [21]. Gain Ratio [20] was applied in GFR phase 2 to rank genes by class discrimination and alone to compare with GFR. The same parameter settings for Gain Ratio were used in both cases. AdaBoostM1 [22] was used for classification with parameters of a primary classifier of DecisionStump being boosted, 10 iterations of boosting and reweighting with a weight threshold of 100. AdaBoostM1 was chosen to handle the class imbalance. Test error of classifiers was estimated with ten–fold cross validation.

Figure 1.2 explores how trustworthiness differs between genes by graphing gene's trust by rank as calculated in GFR phase 1 with equation (1.1). Data attributes f532SD and f635SD were σ_{test} and $\sigma_{control}$ respectively. From this ranking we empirically set the phase 1 threshold to 7000.

Phase 2 of GFR then ranked the phase 1 selected genes with Gain Ratio using the training dataset. Gain Ratio balances feature trust and class discrimination.

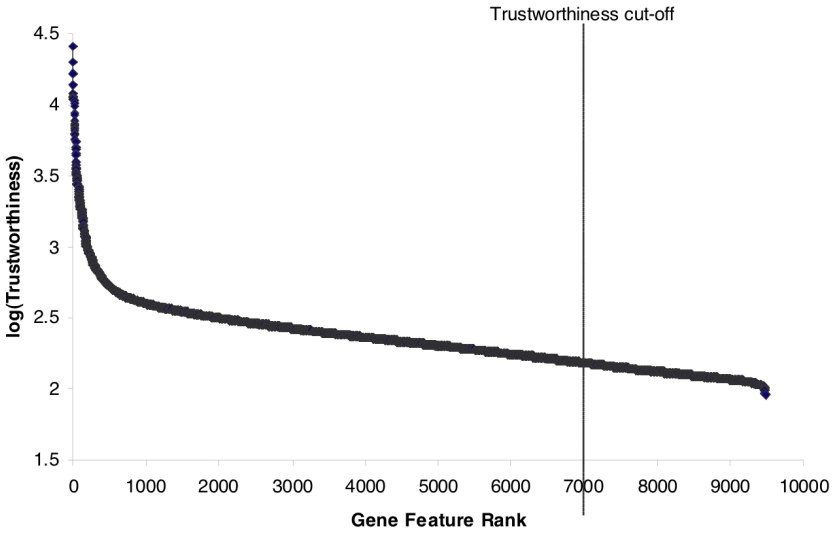

Fig. 1.2 Trustworthiness by gene ranked in phase 1 of GFR. Also shown is the trustworthiness threshold we selected, which is 7,000. All genes ranked after this threshold are removed from further consideration in GFR phase 2.

Genes from the training data were also ranked by Gain Ratio without applying the first phase of GFR. Comparison between the GFR ranked list and GainRatio–only ranked list show that the top ten ranked genes are the same in both lists. Four of the genes ranked within the top 20 by GainRatio are ranked below the trustworthiness threshold by GFR. Of the top 256 genes, 34 differ between GFR and Gain Ratio.

To further illustrate these changes between GFR and GainRatio–only selected genes, the raw data was converted to a basic expression ratio using

$$Expression = \log_2\left(\frac{\mathrm{F635}-\mathrm{B635}}{\mathrm{F532}-\mathrm{B532}}\right) \tag{1.2}$$

where B is the intensity fluorescence for each channel in the local background region around the feature (F). In eq. (1.2), if the feature is poor quality, faint, misshapen or has a particularly noisy background, often the $\mathrm{F}-\mathrm{B}$ value in one channel is negative yielding an incalculable logarithm value. The expression value for the 20 genes from each feature selection approach was identified for the 61 ALL patients, subjected to hierarchical clustering and displayed as a "heat map". Expression values which were "null" due to poor quality were represented as a grey square on the heat map. Figure 1.3 indicates that the four genes removed from the top 20 with GFR were all overly represented with null expression values across the patient cohort. Both sets of gene could distinguish B from T–lineage ALL, however GFR leads to the selection of more trustworthy genes.

Classifiers were built using the "immunotype" class attribute and expression attributes ("f635Median" and" f532-Median") for the top 16 ranked genes for GFR

and Gain Ratio. The motivation behind choosing this number of features was from considerations of being able to compare models for different classes.

Accuracy of these classifiers derived from 10–fold cross validation is reported in Table 1.1. Features selected with GFR result in a slightly more accurate classifier than with Gain Ratio. We also analyzed the impact of using different sized GFR feature subsets on classification accuracy (see Table 1.2). Classifiers were built using the same parameters as before, with the only difference being the number of features used (16, 1024 or all genes). The best performance arose from using the sixteen features chosen by GFR. Regardless of the number of features used or the feature selection method, all classifiers were accurate. This was expected due to the strong genetic relationship with leukemia cell types.

Table 1.1 Cross validation accuracy of classifiers built using the top 16 ranked features. Column 1 is classification accuracy. Column 2 is the Kappa statistic defined as the agreement beyond chance divided by the amount of possible agreement beyond chance [23] p.115. Column 3 are the total number of classification errors. Column 4 is the feature selection method applied.

Accuracy	Kappa	Total Errors	Feature selection methodology
99.39%	0.9844	1	GFR
98.17%	0.9537	3	Gain Ratio

Table 1.2 Accuracy of classifiers trained using different numbers of the top GFR ranked features. Column 1 is the number of features used. Column 2 is the classification accuracy. Column 3 is the Kappa statistic. Last four columns present counts of true and false positives and negatives.

Features	Accuracy	Kappa	True positive	False negative	False positive	True negative
16	99.39%	0.9844	120	0	1	43
1,024	98.17%	0.9531	119	1	2	42
9,485	98.78%	0.9689	119	1	1	43

1.4 Conclusion

This chapter introduces Gene Feature Ranking, a feature selection approach for microarray data that takes into account the trustworthiness or quality of the measurements by first filtering out genes with low quality measurements before selecting features based on how well they discriminate between different classes of patient. We demonstrate GFR to classify the cancer subtype of patients suffering from ALL. Gene Feature Ranking is compared to Gain Ratio and we show that GFR outperforms Gain Ratio and selects more trustworthy genes for use in classifiers.

References

1. Hardiman, G.: Microarray technologies - an overview. Pharamacogenomics **3** (2002) 293–297
2. Schena, M.: Microarray Biochip Technology. BioTechniques Press, Westborough, MA (2000)
3. Bolstad, B., et al.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19** (2003) 185–193
4. Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S.B., Bassett, D.E.: Rosetta error model for gene expression analysis. Bioinformatics **22** (2006) 1111–1121
5. Seo, J., Gordish-Dressman, H., Hoffman, E.P.: An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics **22** (2006) 808–814
6. Tsai, C.A., et al.: Sample size for gene expression microarray experiments. Bioinformatics **21** (2005) 1502–1508
7. Baldi, P., Hatfield, G.W.: DNA Microarrays and Gene Expression: from experiments to data analysis and modeling. Cambridge University Press (2002)
8. Golub, T., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** (1999) 7
9. Mukherjee, S., Tamayo, P., Slonim, D.K., Verri, A., Golub, T.R., Mesirov, J.P., Poggio, T.: Support vector machine classification of microarray data. AI memo 182. CBCL paper 182. Technical report, MIT (2000) Can be retrieved from ftp://publications.ai.mit.edu.
10. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence **97** (1997) 245–271
11. Yang, J., Hanavar, V.: Feature subset selection using a genetic algorithm. Technical report, Iowa State University (1997+)
12. Efron, B., Tibshirani, R., Goss, V., Chu, G.: Microarrays and their use in a comparative experiment. Technical report, Stanford University (2000)
13. Bellman, R.E.: Adaptive Control Processes. Princeton University Press (1961)
14. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Eleventh International Conference (Machine Learning), Kaufmann Morgan (1994) 121–129
15. Saeys, Y., Inza, I., et al.: A review of feature selection tecnhiques in bioinformatics. Bioinformatics **23** (2007) 2507–2517
16. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Machine Learning Research (2003) 1157–1182
17. Wang, X., Ghosh, S., Guo, S.W.: Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Research **29** (2001) 8
18. Park, T., Yi, S.G., Lee, S., Lee, J.K.: Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. BioTechniques **38** (2005) 463–471
19. Yu, Y., Khan, J., et al.: Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homeoprotein six- 1 as key metastatic regulators. Nature Medicine **10** (2004) 175–181
20. Quinlan, J.R.: Induction of decision trees. Machine Learning **1** (1986) 81–106
21. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann, San Francisco (2005)
22. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, San Francisco, Morgan Kaufmann (1996) 148–156
23. Dawson, B., Trapp, R.G.: Basic & Clinical Biostatistics. Third edn. Health Professions. McGraw-Hill Higher Education, Singapore (2001)

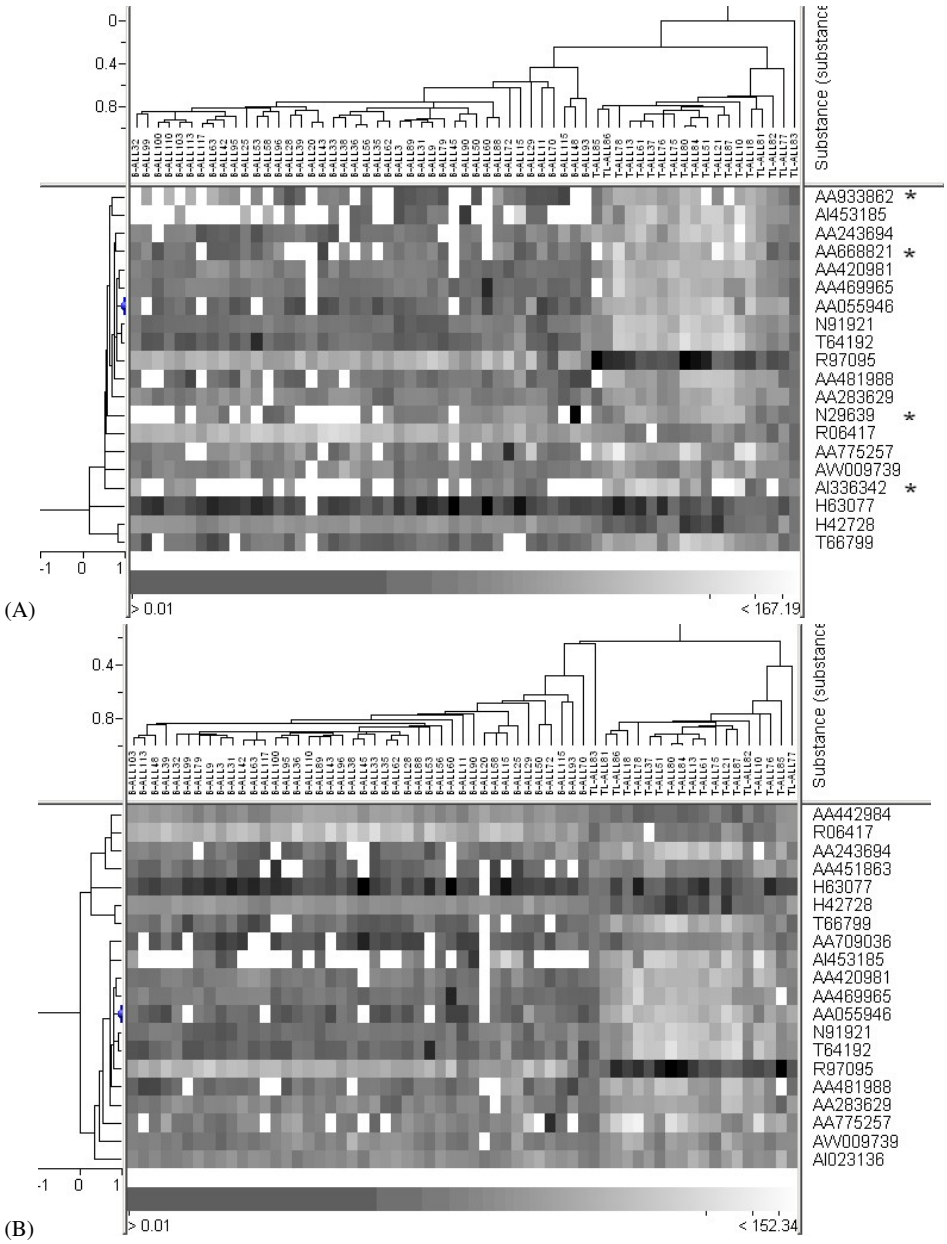

Fig. 1.3 Hierarchical cluster analysis of the top 20 genes selected following (A) Gain Ratio only and (B) GFR feature selection. The dendrogram at the top of each plot identifies the relationship of each patient on the basis of the expression of the 20 genes; B–lineage (on left and marked B–) and T–lineage (on right and marked T– or TL–). Each box on the heat map represents a gray–scale annotation of the expression ratio for each gene (row) in each ALL patient (column), with the gray scale at the base of each plot. Increasing light gray boxes represents increasing expression in ALL patients compared to control samples, whilst dark gray represents decreasing expression. White boxes represent "null" expression value indicative of poor quality features. Genes marked with * were removed by GFR phase 1.