Re-examining expression profiling of pediatric acute myelogenous leukemia done by Ross *et. al* in 2002

Michael Schubert

28th November 2010

Introduction

The original experiment and analysis of acute myelogenous leukemia (AML) subtypes expression profiling using microarrays was done by Ross *et. al* in 2002 [1]. While he disease itself is relatively rare in pediatric patients, the prognosis is worse compared to its lymphoblastic counterpart (AML). This might be due to its heterogeneity, comprising various subtypes with distinct properties that do not all fit one treatment schema. Thus genes which were found to be up- or downregulated could be used for diagnosis or as therapeutic targets.

Materials and Methods

The Affymetrix microarray expression data from the original experiment[1] was used. See their publication for a detailed description of experimental parameters and sample preparation. The experiments were divided into three different sets: a training set on which most analyses were performed on, and two additional sets for classifier testing (cf. table 1).

In contrast to the original work, the open-source scripting language R and the Bioconductor package [2] were used for statistical analysis instead of Affymetrix's commercial software. First, the threestep method (affyPLM) was used for background correction (RMA.2), normalisation (quantile) and summarisation (median polish) on the raw data. The advantage over the expresso method (affy) is that it is faster while not significantly sacrificing accuracy. Using it, a set of 22283 genes was obtained.

Table 1: Pediatric AML subgroups and sample numbers; all samples named „other“ have been removed

Genetic subgroup	Sample name	Training no.	Test no.	Test no. in adult AML
CBFβ-MYH11	inv16	10	4	2
PML-RARα	t(15:17)	11	4	4
FAB-M7	Other-M7	7	3	0
AML1-ETO	t(8:21)	16	5	6
MLL chimeric fusion genes	MLL	18	5	6
Total number		62	21	18

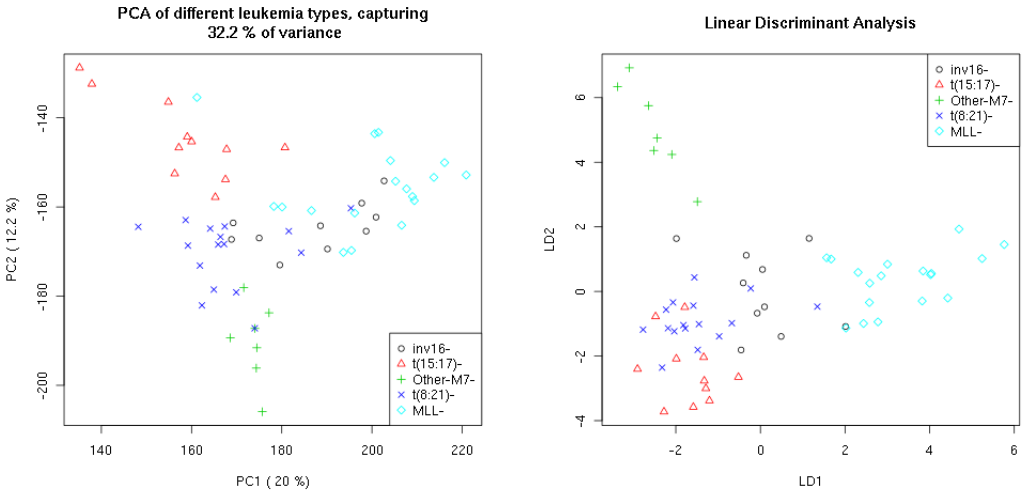
Significance analysis was directly performed on the whole training set, omitting a variation filtering step proposed in the original paper. This approach was chosen because multiple pre-filtering steps

[1] `http://www.stjuderesearch.org/site/data/AML1/`

(kOverA and a custom variation filter with different settings) were tried, but none was found to improve prediction accuracy while potentially losing significant data. A subset of the significantly up- or downregulated genes (limiting Benjamini-Hochberg [3] adjusted p-values to below 0.01) was used for PCA, hierarchical clustering, and SVM classification.

Results

Two types of exploratory analyses were carried out: principle component analysis and linear discriminant analysis. A kOverA filter (at least 6 genes with an expression value of above 150) was applied to the data in order to reduce CPU load, which should not have a great impact on the result because genes that no not vary in expression levels should not contribute to dataset variance. It can be seen (cf. figure 1) that the expression profiles of the different disease types already cluster relatively well, with the exception of inv16 (both methods) and an overlap between t(15:17) and t(8:21) in LDA. This suggests that the inv16 subtype is in itself quite heterogenous.



(a) Principle component analysis **(b)** Linear discriminant analysis

Figure 1: Exploratory analyses of a slightly reduced set of genes

Significance analysis was carried out using the eBayes method and Benjamini-Hochberg adjusted p-values for multiple testing. All possible combinations of two disease types were considered, where a summary of the significant differences found is shown in table 2. Note that the number of differently expressed genes for CBFβ-MYH11 (inv16) is lower compared to the other types, and the number between t(15:17) and t(8:21) is low, confirming the presumptions taken in the exploratory analyses.

Table 2: Pair-wise comparison of the number of significant differences (adj. $p < 0.01$) in gene expression between subtypes, leading to a total of 7408 distinct genes

Group1/Group 2	CBFβ-MYH11	AML1-ETO	MLL	PML-RARα	FAB-M7
CBFβ-MYH11	-	389	554	1471	1407
AML1-ETO		-	2064	1341	1978
MLL			-	3272	2317
PML-RARα				-	3284
FAB-M7					-

Using the result from the significance analysis, the 10 top-scoring genes from each combination were taken and combined into a set of 43 unique genes, which were analysed for their disease-class discriminating potential. The hierarchically clustered expression profile of these is shown in figure 2, being able to divide the different disease types into distinct clusters with short paths within one and long paths between types. Also, the list of found genes is nearly identical to the one in the original paper, which leads to the conclusion that the reduced set does still very well reflect expression variance.

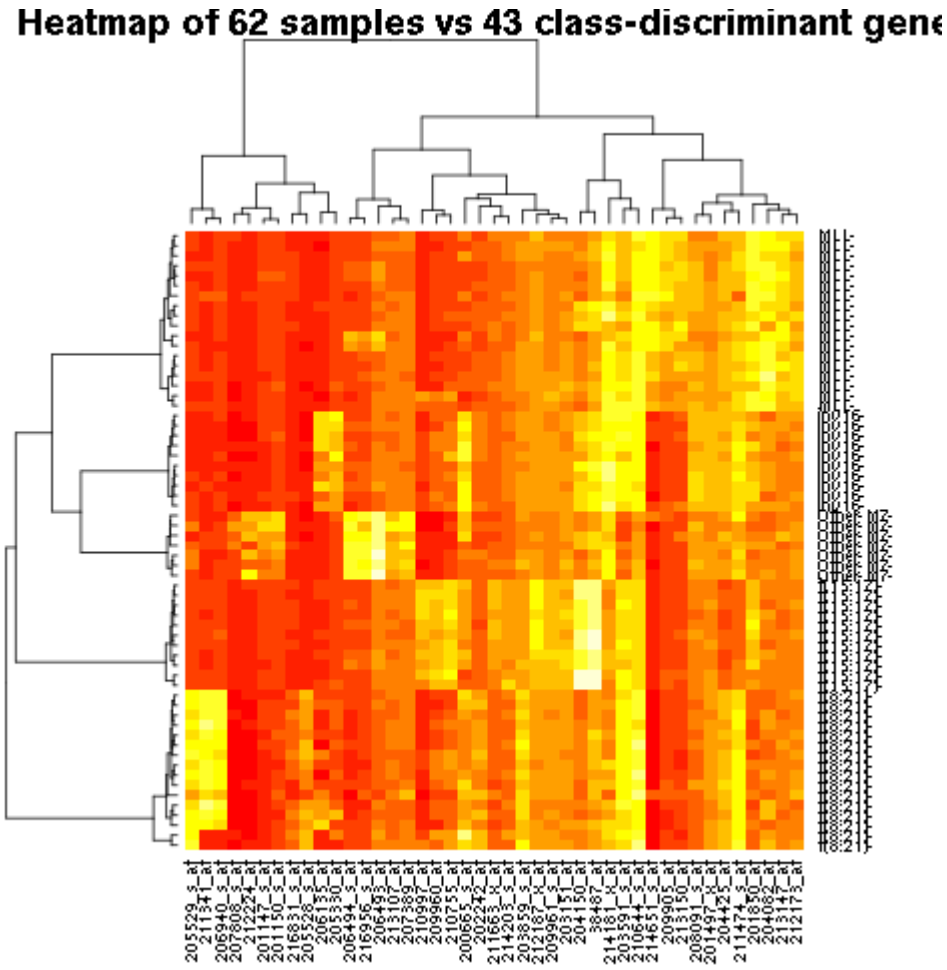


Figure 2: Heatmap of the expression levels of 43 class-discriminating genes (log2 values)

The same downsized gene set was used for training and testing SVM classification. The training set consisted of 62 samples (with 43 features each). When testing the classification on the training set, all samples were assigned the correct labels. 10-fold cross validation was also done to prove no overfitting occurred, which produced the same results in each run. Classification was equally successful on the testing set consisting of 21 samples, and the adult AML cases with 18 samples (being checked manually[2] because of missing experiment labels). It has thus been more successful than the SAM/ANN approach taken in the original paper, which classified 2 MLL cases incorrectly in both training and test set and some more in the adult samples. The SVM accuracy percentage scores are shown in table 3.

[2] `http://www.stjuderesearch.org/site/data/AML1/SectionI/table_s3`

Table 3: Diagnostic accuracies [%] for AML using 43 class-discriminant genes

	Training, n=62	Test, n=21	Adult AML Test, n=18
CBFβ-MYH11	100	100	100
AML1-ETO	100	100	100
MLL	100	100	100
PML-RARα	100	100	100
FAB-M7	100	100	none in set

References

[1] Ross M. E., Mahfouz R., Onciu M., Liu H. C., Zhou X., Song G., Shurtleff S. A., Pounds S., Cheng C., Ma J., Ribeiro R. C., Rubnitz J. E., Girtman K., Williams W. K., Raimondi S. C., Liang D. C., Shih L. Y., Pui C. H., and Downing J. R. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, 104(12):3679–3687, 2004.

[2] Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[3] Benjamini Y. and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.