APPLICATION NOTES

Genomics & Informatics Vol. 5(2) 83-87, June 2007

A DNA Microarray LIMS System for Integral Genomic Analysis of Multi-Platform Microarrays

Mi Kyung Cho[2], Jason Jongho Kang[1] and Hyun Seok Park[1,2]*

[1]Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea, [2]Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea

Abstract

The analysis of DNA microarray data is a rapidly evolving area of bioinformatics, and various types of microarray are emerging as some of the most exciting technologies for use in biological and clinical research. In recent years, microarray technology has been utilized in various applications such as the profiling of mRNAs, assessment of DNA copy number, genotyping, and detection of methylated sequences. However, the analysis of these heterogeneous microarray platform experiments does not need to be performed separately. Rather, these platforms can be co-analyzed in combination, for cross-validation. There are a number of separate laboratory information management systems (LIMS) that individually address some of the needs for each platform. However, to our knowledge there are no unified LIMS systems capable of organizing all of the information regarding multi-platform microarray experiments, while additionally integrating this information with tools to perform the analysis.

In order to address these requirements, we developed a web-based LIMS system that provides an integrated framework for storing and analyzing microarray information generated by the various platforms. This system enables an easy integration of modules that transform, analyze and/or visualize multi-platform microarray data.

Availability: The website for the system can be accessed at the following web address; http://dnachip.macrogen.co.kr/. The LIMS system was originally built to support the DNA Microarray System Development Center for Diagnostic and Prognostic Application to Genetic and Acquired Diseases, supported by the Ministry of Health & Welfare of Korea. Academic users who want to use the system described here should contact us via http://cafe.naver.com/dnachip.cafe/ to attain access privileges.

Keywords: dna chip, molecular diagnosis, data warehouse, LIMS

DNA Microarray System Development Center for Diagnostic and Prognostic Application to Genetic and Acquired Diseases

Our multi-platform microarray LIMS system was originally designed to support the DNA Microarray System Development Center for Diagnostic and Prognostic Application to Genetic and Acquired Diseases. The aim of the center, during the first phase of the project, was to validate eighteen prototype multi-platform microarrays for their potential clinical and diagnostic application. In the second phase of the project, which will be completed by the year 2010, studies are aimed at developing at least three diagnostic microarrays and/or commercially viable biomarkers. During this period, a vast amount of microarray data will be stored in our LIMS, in collaboration with five genome research centers of Korea; The Genome Research Center for Reproductive Medicine and Infertility, The Genome Research Center for Gastroenterology, The Skin Diseases Genome Research Center, The Genome Research Center for Lung and Breast/Ovarian Centers, The Genome Research Center for Cardiovascular Disease. It goes without saying that data-mining a logical set of microarray data of this magnitude will require a robust informatics system incorporating powerful analytical and visualization tools. To this end, we have designed an integral multi-platform microarray LIMS, which integrates biomaterial information, raw images, and extracted data. Our system is compatible with most types of array experiments and data formats.

Dealing with Multi-platform Microarrays Data

A variety of studies have emerged which measure total chromosomal copy number variations (CNV), single nucleotide polymorphism (SNP) array data, expression profiles or methylation data, at an increasingly high resolution. Comparative and multi-platform analyses need not be hampered by the use of heterogeneous platforms; rather, we may need to combine the results of heterogeneous platforms to generate better bioinformatics analyses. For

*Corresponding author: E-mail hspark@macrogen.com, neo@ewha.ac.kr
Tel +82-2-2113-7007, Fax +82-2-2113-7016
Accepted 2 April 2007

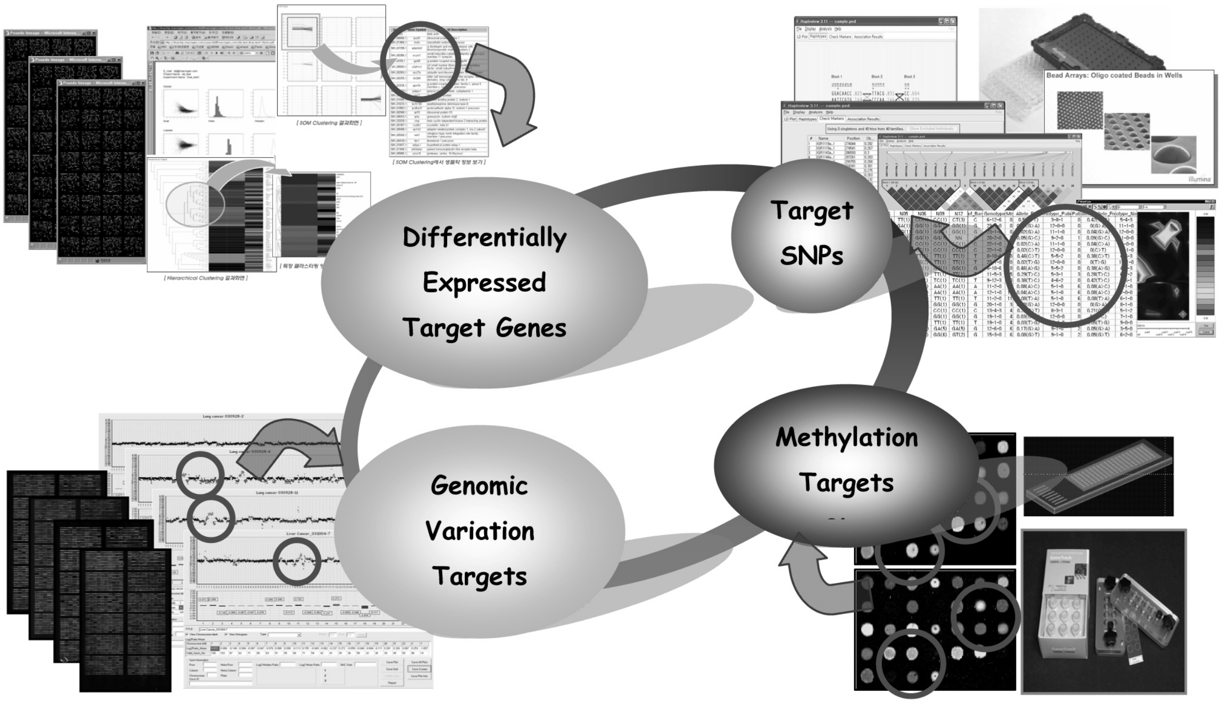


Fig. 1. Microarray technology has been utilized in various applications such as the profiling of mRNA, assessment of DNA copy number, genotyping, and detection of methylated sequences. The results of various platform experiments can be analyzed together for comprehensive analyses and cross validation.

example, we may need to develop a procedure that is capable of extracting both copy number and allele type information from SNP array data. This information could be used to derive allele-specific copy number across the whole genome, in order to visualize not only where amplifications and deletions occur, but also the haplotype of the region being amplified or deleted. A recent study shows that copy number data retrieved from either array comparative genome hybridization (CGH) or SNP arrays are comparable, and that the integration of genome-wide loss of heterozygousity (LOH), copy number and gene expression data is useful for the identification of gene specific targets (Judith *et al*., 2007). Other studies have shown an integrated genomic analysis using expression profiling and array-based CGH (Remco *et al*., 2007). It has also been reported that cross-platform classification of heterogeneous multi-platform microarray data sets yields discriminative gene expression signatures which can be detected and validated by a large number of microarray samples (Patrick *et al*., 2005; Mitchell *et al*., 2004). Artificial neural networks (ANNs) have also been successfully applied to two completely different microarray platforms (cDNA and oligonucleotide) (Bloom *et al*., 2004). In fact, predictive models generated by training heterogeneous microarray platforms can be better validated than those generated on a single data set, while displaying high predictive power and improved generalization performance. A comparison between array CGH, expression profiling and SNP arrays reveals that the overall concordance in detection was relatively high. Currently, our LIMS system is capable of dealing with various platforms including Illumina® GoldenGate Assay, Illumina Sentrix® HumanRef-8 Expression BeadChip, Array CGH: MacArray™ Karyo 4000, MacArray™ M-chip, GeneChip® Human Mapping 500K Array, CodeLink™ Human Whole Genome Bioarray, and Agilent Whole Human Genome Oligo Microarray Kit (44K).

Figure 2 shows one of our exemplary research results, which was produced in collaboration with The Genome Research Center for Gastroenterology. Using various algorithms and visualization tools that we have developed with our LIMS, nine common genes could be selected from 195 candidate SNPs, 144 clones from array CGH and 263 genes from expression profiles for distinguishing normal liver, chronic hepatitis, and hepatocellular carcinoma.

SNP chips, GoldenGateTM Assay Procedures

In an attempt to identify the normal liver, chronic hepatitis, and hepatocellular carcinoma risk conferred by SNP interactions, we studied 1,536 SNPs from a selection of 114 significant genes involved in major cancer-related liver disease. For the current study, we selected samples from

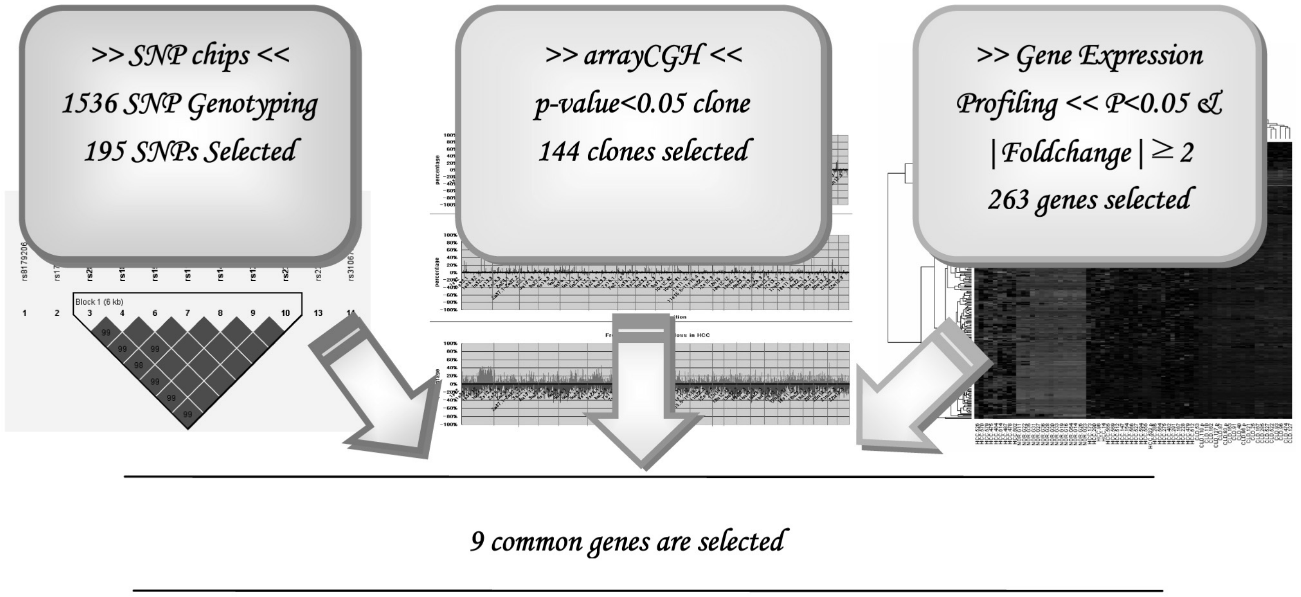


Fig. 2. An exemplary case of multi-platform analysis: nine common genes were selected from SNP, array CGH & expression profiles representing normal liver, chronic hepatitis, and hepatocellular carcinoma.

459 patients who presented with various stages of liver disease. All SNPs were genotyped using Illumina's GoldenGateTM assay. The association between the case-control status and each individual SNP, measured by the odds ratio and its corresponding 95% confidence interval, was estimated using unconditional logistic regression models. The robustness of the interactions, which were observed among the selected SNPs with stronger functional evidence, was assessed using a correction for multiple testing based on the false discovery rate (FDR) principle. Of the SNPs, 195 (67 genes) contributed to chronic hepatitis disease states and the risk of hepatocellular carcinoma individually.

arrayCGH

Bacterial artificial chromosome (BAC) clones were selected from Macrogen's proprietary BAC library (http://www.macrogen.com). Briefly, pECBAC1 (Friijter *et al*. 1997) was restricted using the HindIII endonuclease, and HindIII-digested pooled male DNA selected by size, was used to generate a BAC library. The clones were first selected using bioinformatic techniques to give an average genomic coverage of 2 Mb. All of the clones were two end-sequenced using Applied Biosystems 3700 sequencers, and their sequences were analyzed using BLAST and mapped according to their positions on the UCSC human genome database (http://www.genome. uscs.edu). Confirmation of locus specificity of the chosen clones was performed by removing multiple loci binding clones by individual examination under a standard FISH condition as previously described. These clones were prepared using a conventional alkaline lysis method to obtain BAC DNA. Each BAC clone was represented on an array as spots in triplicate and each array was pre-scanned using an Axon scanner to ensure proper spot morphology. The array used in this study consisted of 4,040 human BACs, which were spaced approximately at 2.3Mb across the whole genome. Chromosomal aberrations were categorized as a gain when the normalized $\log_2$ transformed fluorescent ratio was higher than 0.25 and as a loss when this ratio was below -0.25. These two threshold values were chosen empirically by selecting a 3X standard deviation value calculated from 30 normal male and normal female hybridization experiments. Macrogen's MAC viewer, aCGH analysis software, MS Excel VBA, and avadis3.3 Prophetic software was used for graphical illustration and image analysis of array CGH data.

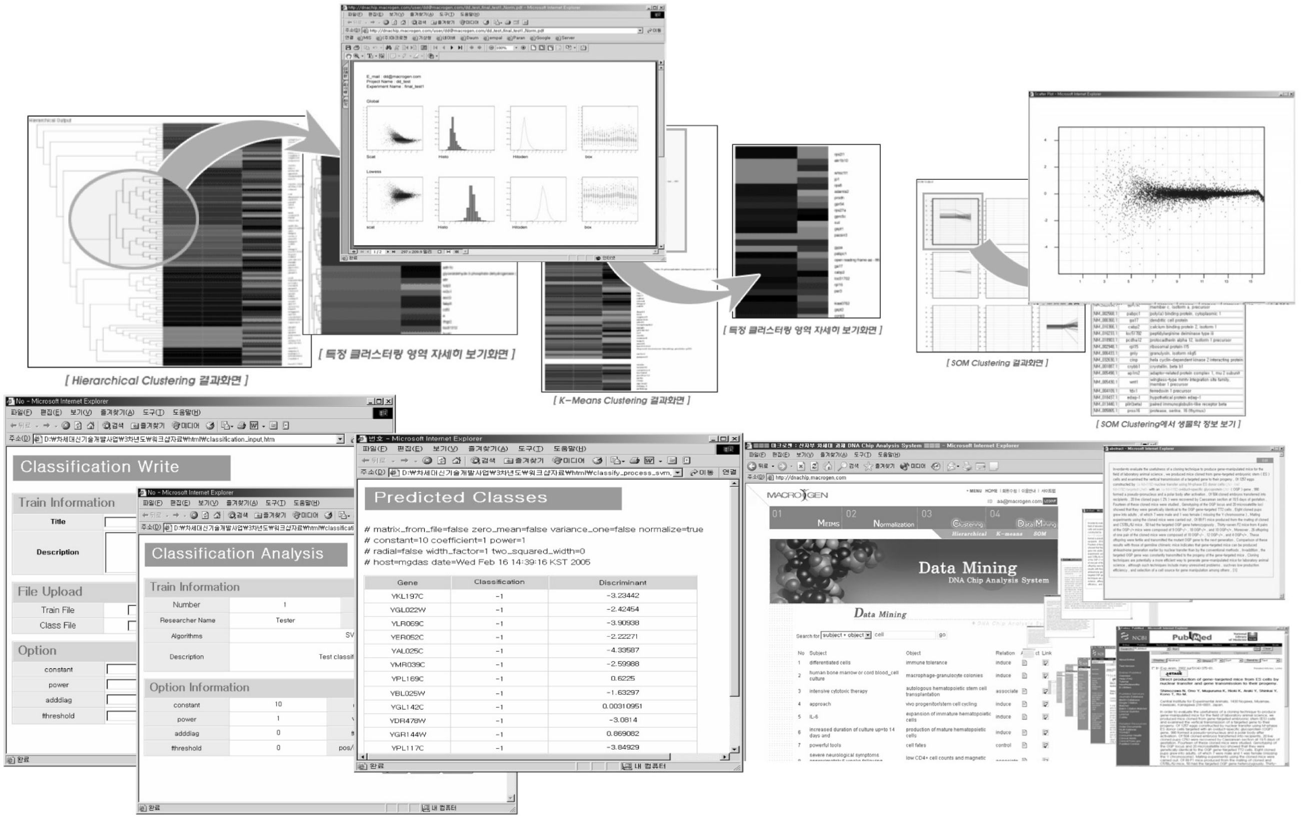
For this study, we selected samples from patients presenting with the various types of liver disease. For calling gains and losses in array CGH data, the Fisher's exact test in R software was used with a False Discovery Rate (FDR) <0.05. Chi-square analysis was applied to the generated tables to test for a significant difference in the distribution of loss or gain vs. no change between groups (normal liver, chronic hepatitis, and hepatocellular carcinoma). Using this method we detected 144 significant clones (83 genes).

Gene expression profiling

According to the manufacturer's instructions, 750 ng of labeled cRNA from each sample was hybridized to each Sentrix HumanRef-8 Expression BeadChip for 16-18 h at 58°C (Illumina, Inc., San Diego, CA). Detection of the array signal was performed using Amersham fluorolink streptavidin-Cy3 (GE Healthcare Bio-Sciences, Little Chalfont, UK) by the method described in the BeadChip manual. Arrays were scanned with an Illumina Bead array Reader confocal scanner according to the manufacturer's instructions. Array data processing and analysis was performed using Illumina BeadStudio software. A whole genome expression study was performed to analyze differential gene expression profiles. The 63 samples were designated as normal liver, chronic hepatitis disease, or hepatocellular carcinoma. As a result of replication, we achieved a high accuracy of gene expression and a reduced error rate (correlation=0.98). A total of 19,091 probes were normalized by the quantile normalization method for use in the analysis. One-way analysis of variance (ANOVA) and Tukey's HSD test were applied to determine differentially expressed sets of genes across the three experimental groups. Statistical significance was adjusted by applying the Benjamini- Hochberg FDR multiple-testing correction method. We selected 1,950 probes with significant differential expression patterns for further analyses. By comparing the commonly affected genes with the three independently identified genes, we have identified 11 genes that showed a high correlation with liver disease progression and cancer formation.

System Architecture

The software used in this study was developed on the GNU/Unix operating system using the JAVA programming language (http://java.sun.com/). Data was stored in a related database (Oracle 9i) and communicated to the user via the Apache Webserver (http://www.apache.org/). When required, the user interface utilizes Perl software (http://www.perl.org/), and C++ software can be used for the more computationally intensive tasks on the server. The system follows MIAME guidelines, a set of minimal annotation rules for microarray-based experiments (Brazma *et al*., 2001). In our system, data can be visualized at several stages of analysis. Unmodified and transformed data sets can be plotted interactively as scatter plots, displayed in histograms, or viewed as tables (Fig. 3). Entire experiments can be displayed in various overview plots, and figures can be exported for publication. We have implemented the


[Hierarchical Clustering 결과화면]
[특정 클러스터링 영역 자세히 보기화면]
[K-Means Clustering 결과화면]
[특정 클러스터링 영역 자세히 보기화면]
[SOM Clustering 결과화면]
[SOM Clustering에서 생물학 정보 보기]
No - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
번호 - Microsoft Internet Explorer
MACROGEN
01 MEIMS
02 Normalization
03 Clustering
04 Data Mining
Hierarchical
K-means
SOM
Data Mining
DNA Chip Analysis System
완료
내 컴퓨터

주소(D) D:₩차세대신기술개발사업₩3차년도₩워크샵자료₩html₩classification_input.htm

Classification Write

Train Information

Title	
Description	

File Upload

Train File	
Class File	

Option

constant	
power	
adddiag	
fthreshold	

Classification Analysis

Train Information

Number	1
Researcher Name	Tester
Algorithms	SV
Description	Test classif

Option Information

constant	10	
power	1	
adddiag	0	
fthreshold	0	pos/

주소(D) D:₩차세대신기술개발사업₩3차년도₩워크샵자료₩html₩classify_process_svm.

Predicted Classes

```
# matrix_from_file=false zero_mean=false variance_one=false normalize=true
# constant=10 coefficient=1 power=1
# radial=false width_factor=1 two_squared_width=0
# host=mgdas date=Wed Feb 16 14:39:16 KST 2005
```

Gene	Classification	Discriminant
YKL197C	-1	-3.23442
YGL022W	-1	-2.42454
YLR069C	-1	-3.90938
YER052C	-1	-2.22271
YAL025C	-1	-4.33587
YMR039C	-1	-2.59988
YPL169C	-1	0.6225
YBL025W	-1	-1.63297
YGL142C	-1	0.00310951
YDR478W	-1	-3.0814
YGR144W	-1	0.869082
YPL117C	-1	-3.84929

주소(D) http://dnachip.macrogen.com

ID aa@macrogen.com

Data Mining

Search for subject + object cell go

No	Subject	Object	Relation
1	differentiated cells	immune tolerance	induce
2	human bone marrow or cord blood_cell culture	macrophage-granulocyte colonies	induce
3	intensive cytotoxic therapy	autologous hematopoietic stem cell transplantation	associate
4	approach	vivo progenitor/stem cell cycling	induce
5	IL-6	expansion of immature hematopoietic cells	induce
6	increased duration of culture up=to 14 days and	production of mature hematopoietic cells	induce
7	powerful tools	cell fates	control
8	severe neurological symptoms	low CD4+ cell counts and magnetic	

Fig. 3. A snapshot of multi-platform microarray LIMS interfaces.

within-slide intensity-dependent LOWESS (locally weighted scatter plot smoothing) module (Cleveland *et al*., 1988), and the MDS (normalization and multidimensional scaling) analysis module (Kruskal *et al*., 1978) to compute the distance metric between samples.

Acknowledgement

This work was partially supported by a grant from the Ministry of Information and Communication of Korea under the grant number of ab05, and by the Ministry of Health & Welfare of Korea under grant number of A050558. We would like to thank Macrogen Inc., DNA Link Inc., Genomic Tree Inc., and Digital Genomics Corp for various communications.

References

Benjamini,Y. and Hochberg,Y.(1995).Controlling the False Discovery Rate, A Practical and Powerful Approach to Multiple Testing. *Journal of Royal statistical Soceity Series B*. 57, 289-300.

Bloom, G., Yang, I.V., Boulware, D., Kwong, K.Y., Coppola, D., Eschrich, S., Quackenbush, J., and Yeatman, T.J. (2004). Multiplatform, multi-site, microarray-based human tumour classification, *Am. J. Pathol.* 164, 9-16.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum Information about A Microarray Experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* 29, 365-371.

Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted Regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* 83, 596-610.

Dijkman, R., Doorn, R., Szuhai, J., Willemze, R., Vermeer, M.H., and Tensen, C. (2007). Gene-expression profiling and array-based CGH classify CD4+CD56+ hematodermic neoplasm and cutaneous myelomonocytic leukemia as distinct disease entities. *Blood* 109, 1720-1727.

Judith, N., Kloth, J.N., Oosting, J., Wezel, T., Szuhai, K., Knijnenburg, J., Gorter, A., Kenter, G.G., Fleuren, G.J., and Jordanova, E.S. (2007). Combined array-comparative genomic hybridization and single-nucleotide polymorphismloss of heterozygosity analysis reveals complex genetic alterations in cervical cancer. *BMC Genomics* 8, 53.

Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling,* Sage.

Mitchell, S.A., Brown, K.M., Henry, M.M., Mintz, M., Catchpoole, D., LaFleur, B., and Stephan, D.A. (2004). Inter-platform comparability of microarrays in acute lymphoblastic leukemia, *BMC Genomics* 5, 71.