Visualization and combined analysis of SNP and gene expression data with Rcnat

Sylvia Merk[1], Hans-Ulrich Klein[1], Claudia Haferlach[2], Martin Dugas[1]

1: Dep. of Medical Informatics & Biomathematics, University of Muenster, Germany
2: MLL Muenchner Leukaemielabor GmbH, Munich, Germany
sylvia.merk@ukmuenster.de

The analysis of single nucleotide polymorphisms (SNP) applying microarray technology involves the potential to understand mechanisms of disease pathogenesis on a molecular basis. Latest Affymetrix™ SNP-Arrays offer the possibility to examine up to 500.000 SNPs per sample. SNPs are widely used to assess copy number (CN) variants like amplifications or deletions as well as loss of heterozygosity (LOH) as molecular causes of disease, mainly in cancer research [1,2].
During the last years another microarray technology, the analysis of gene expression on the mRNA-level, was established successfully to investigate molecular causes of disease. For a better and comprehensive understanding of molecular pathogenesis it is highly desirable to accomplish an integrated analysis of SNP and gene expression data.

A first analysis of SNP-Data can be accomplished by applying the manufacturer's software „Copy Number Analysis Tool“ (CNAT) [3]. However, the capabilities of this software are limited. As an example there is no possibility to visualize data of more than one sample at a time, impeding an adequate comparison and interpretation of the given data.

There is very limited software available to enable enhanced visualization of SNP-data and to provide a combined analysis of SNP and gene expression data. For this reason we developed Rcnat.
Rcnat is implemented in the statistics language R. It was validated on a dataset comprising SNP- (Affymetrix Mapping 10K 2.0) and gene expression data (Affymetrix HG-U133A/B) derived from 33 patients suffering from AML with aberrant complex karyotype. Gene expression data (Affymetrix HG-U133A/B) from 100 AML-patients with normal karyotype served as control.

Import of SNP data is possible for *.cnt-files (exported from CNAT) as well as for tab-separated *.txt-files containing LOH-values or copy number-values. Gene expression data can be read in either as *.CEL-files or as tab-separated *.txt-files.
A major focus of Rcnat is visualization of data on a chromosome or sample basis as well as an integrated visualization of data from differing microarray sources. It facilitates the comparison of LOH, copy number and gene expression data of each sample at a glance enabling visual identification of chromosomal regions harbouring LOH, amplifications or deletions (Fig. 1B-C). Chromosomal regions showing copy number or gene expression alterations are identified by a segmentation algorithm. A graphical display of gene expression values is achieved by plotting the difference of the expression value of one particular probe set of the sample and the mean of the expression values of the same probe set measured for 100 control samples. An overview of color coded LOH or copy number values of all samples along a particular chromosome can be visualized as a heatmap where furthermore similar samples are clustered together (Fig. 1A). Additionally a graphic showing LOH-values of all included samples on each chromosome helps to discover „interesting“ chromosomes (not shown).
It is also possible to compare genes exhibiting alterations in SNP as well as in gene expression data. A SNP is considered as affected if a certain percentage of samples shows alterations in this particular SNP. To determine affection in gene expression data differentially

expressed genes between the two groups (33 aberrant karyotype and 100 normal karyotype) are determined by t-test and the resulting p-values are corrected for multiple testing. Genes are considered as differentially expressed if a false discovery rate <0.05 is achieved. The two gene lists thus obtained are subsequently compared to identify genes contained in both lists.
To assess influence of copy number alterations on gene expression, a method similar to the global test procedure proposed by Mansmann and Meister [4] is applied.

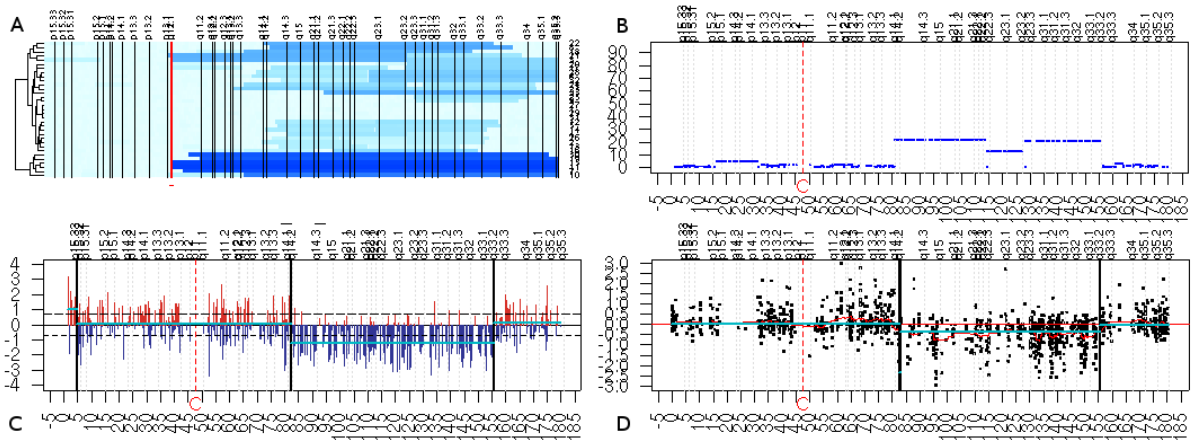

Figure 1:
A: Heatmap of color coded LOH-values along chromosome 5 including all 33 samples. Black vertical lines indicate cytobands; the red line represents the centromere. Rows (individual samples) are clustered according to similarity. Light blue indicates low, dark blue indicates high LOH-values.
B,C,D: LOH, CN and gene expression of one sample along chromosome 5. Black vertical lines in C and D indicate discovered segments. There is a large affected region apparent on the q arm.

In the present dataset we were able to identify noticeable alterations particularly on the q arm of chromosome 5 apparent in gene expression data as well as in LOH- and copy number data. In addition to the visual detection of changes, 48 genes were considered as affected by comparison of gene lists. Furthermore the alterations in SNP as well as in gene expression data could also be confirmed by the global test approach.
Analysis of data from newer array generations (Mapping 250K Array and HG-U133plus2) revealed that Rcnat is able to cope with larger data sets, too.

The combined analysis of SNP- and gene expression data can help to confirm results achieved through one of the mentioned techniques. It can also help to validate results obtained by other lab techniques like cytogenetics. Moreover, the investigation of diseases on this high resolution molecular level could assist in identifying new subgroups of known disease types observable only on a molecular basis.

[1] Hu N et al. Genome-wide loss of heterozygosity and copy number alteration in esophageal squamous cell carcinoma using the Affymetrix GeneChip Mapping 10K array. *BMC Genomics 2006, 7:299*
[2] Zhao X et al. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Research 2004, 64: 3060-3071*
[3] Huang J et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics 2004, 1: 287-299*
[4] Mansmann U and Meister R. Testing Differential Gene Expression in Functional Groups. *Methods Inf Med 2005,44: 449-453*