Ruprecht-Karls-Universität Heidelberg

Fakultät für Klinische Medizin Mannheim

Dissertations-Kurzfassung

Gene expression profiling of prostate cancer using microarray data

Autor: Theodora Manoli

Institut: Zentrum für Medizinische Forschung der Fakultät für Klinische Medizin Mannheim der Universität Heidelberg sowie der Abteilung Theoretische Bioinformatik und der Abteilung Zelluläre und Molekulare Pathologie des Deutschen Krebsforschungszentrums

Doktorväter: Prof. Dr. N. Gretz, Prof. Dr. R. Eils und Prof. Dr. H.-J. Gröne

Prostate cancer is the second most common malignancy of men in the western world and one of the most intensively investigated tumors. Its slow progression and therefore the possibility for early diagnosis and therapy make it very challenging. DNA microarray technologies give us the chance to look through the transcriptional profile of thousands of genes in prostate cancer and precursor samples (e.g. PIN) at once. To accelerate and analyze effectively such huge datasets computational procedures are used. Using such methods we wish to find new interesting differentially expressed genes in prostate cancer, but we also focus on the recently developed pathway analyses. These analyses may be even more powerful than gene to gene analysis because they deal with a metabolic procedure as a whole. Moreover we show particular interest in the gene regulation through transcription factors and miRNAs. We also examine chromosomal locations with an accumulation of differentially expressed genes.

Discrepancies in results obtained by different statistical methods used in microarray analytics and in results coming from different microarray experiments have been cited and commented several times. Therefore it is interesting to examine how significant these discrepancies are.

In this thesis we analyzed one in-house dataset, consisting of normal, Prostate Intraepithelial Neoplasia (PIN) and tumor samples gained by Laser Capture Microdissection (LCM), but also three public microarray datasets of Welsh et al., Singh et al. and Ernst et al. consisting of normal and tumor samples. This data is used for the following purposes:

1. The in-house dataset to find new potential differentially expressed genes, including transcription factors and miRNA-targets, pathways or chromosomal aberrations in PIN and prostate cancer. This dataset is also used to compare annotation, normalization and statistical methods for differential expression and pathway analysis.
2. Three bigger public datasets to examine if data from different datasets and different statistical methods are better comparable on the level of common apparently regulated metabolic pathways than on the level of common differentially expressed genes. These datasets are also used to compare results of differential expression and pathway analysis using two different annotation systems - the Affymetrix CDF and the Custom CDF based system.

All data used came from experiments on Affymetrix microarrays. The in-house dataset had the HGU133A platform, while the other three public datasets were of the same HGU95A or HGU95Av2 platform. All data were imported using both standard Affymetrix CDF annotations and the new Custom CDF annotations. Two normalization methods, vsn and Mixed Model Normalization (MMN), and three statistical methods, SAM, Mixed Model Analysis (MMA) and PAM, to find differentially expressed genes were used and compared. To find apparently affected pathways three approaches were tested, Fisher's exact test, Kolmogorov Smirnov like test and the Global test.

The analysis of the in-house dataset of microdissected samples showed a clear distinction between the expression profiles of normal, PIN and cancerous prostate samples. The expression profiles of PIN samples had in general a higher similarity to normal samples than to cancerous samples. Down-regulation of genes was more dominant in cancer while up-regulation of genes was more dominant in

PIN. Looking for early malignant changes in PIN samples, we found three interesting groups of genes. First, three genes, AMACR, FABP5 and CHC1, were found to be up-regulated in both high-grade PIN and prostate cancer. Second, three genes, HOXC6, NDFIP1 and TARP, were found to be up-regulated already in low-grade PIN and prostate cancer. Third, two genes, KCTD14 and LTF, were found to be down-regulated in high-grade PIN and prostate cancer. The chromosomal analysis showed some changes in already cited aberrant regions like 3q21, 6q23, 8p21, 17p11 and 19p13. Some new potential aberrational regions are proposed to be relevant, e.g. the 17q21.

Transcription factor analysis showed that the transcription factor HOXC6 was up-regulated in both PIN and prostate cancer. It may therefore play a key-role in the initiation of carcinogenesis. Moreover, the transcription factor SIM2 was found to be up-regulated while the transcription factors ATBF1, ETV5, TP73L and TRIM29 were found to be down-regulated in prostate cancer. Three miRNAs, miR-1, miR-124 and miR-373, were found to potentially regulate a total number of nine differentially expressed genes in PIN or in tumor.

Pathway analysis showed that in both PIN and prostate cancer changes in amino acid metabolism and lipid metabolism play a dominant role. Carbohydrate metabolism seems to be significantly involved in prostate cancer, but not in PIN. In contrary changes in glycan biosynthesis and metabolism are more dominant in PIN than in cancer. Among the highly rated significantly regulated pathways in prostate cancer were pathways involved in the metabolism of cofactors and vitamins like 'etinol metabolism', 'biotin metabolism' and 'riboflavin metabolism', pathways of the group of biodegradation of xenobiotics, 'methane metabolism' and 'sulfur metabolism'. Moreover, two recently described multifunctional pathways were found to be changed in prostate cancer, the 'wnt-signalling' and the 'TGF-beta signalling pathway'. Among the highly rated significantly regulated pathways in PIN were 'Jak-STAT signalling pathway', 'proteasome activity', 'ubiquinone biosynthesis' and 'methane metabolism'. Interestingly changes in pathways belonging to the biodegradation of xenobiotics were also in PIN a dominant procedure. The pathway analyses of the datasets of Welsh et al., Singh et al. and Ernst et al. added some presumably changed pathways in prostate cancer to the ones mentioned above. 'Androgen and prostate cancer' showed the highest concordance between the different analyses, validating the meaning of our results. Some new interesting pathways were 'neuroactive ligand-receptor interaction', 'glutathione metabolism', 'hypoxia', 'prion disease', 'stress and toxicity', 'autoimmune and inflammation response' and 'insulin metabolism'.

Comparing two normalization methods, vsn and MMN, we concluded that both methods showed competitive results. Comparing three statistical methods for differential expression, SAM, MMA and PAM, we concluded that all three methods gave biological meaningful results but the overlaps between their results were generally small. The extent of overlap depended, however, on the dataset used. Differences were found not only between results of different statistical methods, but also between results coming from different datasets examining the same situation, in our case prostate cancer.

The Affymetrix annotation has some important drawbacks, and therefore was compared to a Custom CDF annotation system for their effects on results of differential expression analysis. Examining the HGU133A dataset, discrepancies of 15-41% were found, while the discrepancies for the HGU95A/HGU95Av2 datasets were only 8-28%. For the HGU95A/HGU95Av2 datasets, data imported with Custom CDFs gave bigger overlaps of significantly differentially expressed genes between SAM and MMA than data imported with Affymetrix CDFs.

Comparing the results of three pathway analysis methods, Fisher's exact test, Kolmogorov-Smirnov-like test and Global test, we observed that the overlap of significantly regulated pathways between different pathway analysis approaches was small. The Kolmogorov-Smirnov-like test behaves the most diverging. Fisher's exact test and Global test demonstrate the most concordant results when applied on lists of differentially expressed genes coming from different statistical methods and different datasets.
In this thesis, multiple microarray analyses using different statistical and pathway analysis methods have been applied on different datasets. This results in discriminative candidate genes and pathway regulation signatures. Results obtained by these analyses are likely to be more robust than those generated by a single analysis on a single dataset, and can be used for further implementation in the field of medical applications in prostate cancer therapy, e.g. gene therapy. Moreover, pathway analysis applied to gene lists that come from different statistical methods and different datasets yield interesting common results, diminishing the large discrepancies observed in direct comparisons of these gene lists.