ERCIM "ALAIN BENSOUSSAN" FELLOWSHIP PROGRAMME

Scientific Report

First name / Family name	MONALISA MANDAL
Nationality	INDIAN
Name of the *Host Organisation*	NORWEGIAN INSTITUTE OF SCIENCE AND TECHNOLOGY
First Name / family name of the *Scientific Coordinator*	PAL SATROM
Period of the fellowship	01/10/2016 to 30/09/2017

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Hepatocellular carcinoma (HCC) is one of the most deadly and common types of cancer having very limited therapeutic options available. Current treatment options include liver transplantation, tumor resection, and the receptor tyrosine kinase inhibitor, sorafenib. Specific treatment choices depend on tumor size and staging, and the general disease state of the patient's liver; e.g. highly cirrhotic livers are unsuited for resection. Methods for robust patient stratification are therefore important. Recent gene expression studies have uncovered several pathways that drive uncontrolled growth and aberrant survival [1]. Despite these advances, robust markers for diagnosis and prognosis of HCC patients are still lacking [2]. Recent findings show that microRNA (miRNA) expression can be more useful than messenger RNA (mRNA) based profiling for identifying tissue type of tumor origin [3]. Several studies have revealed that miRNAs are deregulated during liver cancer development and miRNAs are deregulated during liver cancer development and regulate

central oncogenic and anti-apoptotic liver cancer signaling pathways [4,5]. Furthermore, the expression levels of specific miRNAs have been identified to correlate with clinicopathological parameters and treatment responses in liver cancer patients.

With these developments in mind, we set out to develop a robust classifier for HCC based on miRNA expression in tumor tissue samples. Using Illumina high throughput sequencing and established small RNA library preparation protocols, we sequenced small RNAs from tumor and adjacent normal samples from 92 HCC patients in two batches (75 and 17 patients, libraries prepared with the TruSeq and NextFlex protocols, respectively). Reads mapping to annotated human miRNAs (miRBase release 21) were then used to create digital miRNA expression profiles, which were used in subsequent data analyses. Principal component analyses revealed a strong batch effect between the first and second set of samples (batches separated in principal component 1). As batch effects are common in genomics studies and a potential source for incomparable results, we focused our efforts on developing classification methods that were robust to such effects. Specifically, we investigated two specific strategies: first, a classification model based on the paired structure of the data; second, a method that used unsupervised analyses to identify and linear modeling to correct for differences between batches. Both strategies relied on support vector machines (SVMs) for tumor classification, and we used both cross-validation (within the 75 patient batch) and an independent test set (the 17 patient batch) to measure the performance of our classification strategies. For comparisons, we also trained classifiers by using the kNN and Naive Bayes algorithms.

For the pairwise strategy, we reasoned that differences between tumor and adjacent normal samples would be invariant to batch effects if all such pairs were sequenced at the same time. Indeed, the miRNAs with most significant expression differences between tumor and normal each had different average expression in the first and second batch, but the differences in average expression between tumor and normal were similar in both batches. An SVM trained on paired expression differences from the first batch achieved perfect classification (ROC-scores 1.0), with the caveat that true class labels were needed to construct the input data to the SVM. Nevertheless, this results shows that expression differences between tumor and adjacent normal tissue are highly reproducible between patients and sequencing batches.

For the batch correction strategy, we used the observation that normal samples clustered more tightly in a PCA plot than did tumor samples to automatically identify, estimate, and correct for batch differences for each miRNA. Specifically, first a PCA analysis on the test data was performed. Second, samples were clustered according to the first two principal components by using Hierarchical Clustering. Third, the largest cluster with the smallest intra-cluster distance of the resultant clusters was selected as the candidate normal samples. Fourth, the log fold change between the normal samples from batch 1 and the candidate normal samples from batch 2 were computed for each miRNA, and these values

were added to the respective miRNAs in each sample in the second batch. PCA analyses showed no batch effects in resulting data. The resulting classifier achieved ROC-scores of 0.93 and 0.97 in batches 1 (cross-validation) and 2 (test set), respectively. Thus, although intra-patient differences between tumor and normal tissue generalize between patients, as indicated by the perfect performance of the pairwise SVM, inter-patient differences in miRNA expression were too large to allow perfect tumor classification based on miRNA expression levels alone.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Classification of Paired Hepatocellular Carcinoma miRNA Expression Profiling Expression and Microarray.(Manuscript in Preparation)

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

Poster Presented in Basel Computational Biology Conference (BC2, 12-15.09.2017), Basel, Switzerland.

Robust classification of microRNA expression profiles from hepatocellular carcinomas

Mandal M, Mjelle R, Chawla K*, Tica V, Dima S, Popescu I, Sætrom P
*Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Norway

https://www.bc2.ch/2017/program/posters/accepted/

IV – RESEARCH EXCHANGE PROGRAMME (REP)

During my fellowship I visited Centrum Wiskunde & Informatica (CWI), Science Park 123, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. In CWI, my scientific coordinator was Dr. Alexander Schönhuth. I was there for ten days and we discussed on different projects. Our common interest was mainly in GWAS based studies. Although, on that time we were unable to have clear picture of our projects but we are in touch to have some good work in future.