

## Novel algorithm for finding robust cancer gene markers

### SUMMARY

Along with new advances in genomics, proteomics, and biotechnology is a growing trend to use biomarkers to assist clinicians in the diagnosis and treatment of a variety of diseases. However, challenges remain in the identification, selection, and validation of relevant and accurate biomarkers. A new and unique algorithm was developed to generate robust gene markers from tumor microarray data. This algorithm was used successfully to generate several robust biomarker sets that correlate with patient survival in breast cancer.

### APPLICATIONS

- Development of prognostic tests for patients with breast cancer.
- Identification and validation of biomarker panels specific to other cancer types or sub-types.
- Development of single- or multi-cancer prognostic test kits.

### CONCEPT

At present, it is very difficult to determine the most effective and safest course of treatment for individual cancer patients. Biomarkers are emerging as a tool that could help address this issue. Because several gene markers are generally needed to enable the biomarker analysis to exhibit sufficient specificity and sensitivity, the microarray assay is a technology of choice for finding multi-markers, as it allows the expression levels of thousands of genes to be measured simultaneously. Still, several challenges limit the identification of cancer biomarkers. Tumour heterogeneity impedes the identification of robust biomarkers. Together with the individual variability of the gene expression profiles, 'real' gene signatures are buried in the global gene profiles of tumours. Further, the number of datasets (tumour samples) for marker discovery is very small. To overcome these limitations, a new strategy to find cancer biomarkers was developed. In a first part, the strategy consists of identifying sets of genes with expressions correlating with survival status using genome-wide microarray profiling of cancer samples from patients with known clinical outcome; together, the samples and their associated data form the original training set. In a second part, 1 million random gene sets are generated from the above mentioned selected gene sets, and 34 random samples sets are generated from the original training set or from subsets. In a

third part, the optimal sets of biomarkers are identified by examining for each random gene set the correlation of the survival status of cancer patients in each random sample set, and by collecting and ranking the most successful random gene sets. Breast cancer was used as a starting point to assess and validate the strategy by finding robust biomarkers for poor prognosis. After applying this algorithm in two breast cancer datasets (training sets), nine marker sets defined as NRC1 to NRC9 were identified. Each marker set contains 30 genes. NRC1 to NRC6 are targeted to oestrogen receptor (ER) positive tumours, while NRC7 to NRC9 are targeted to ER negative tumours. The nine marker sets were validated using three independent breast cancer datasets.

### FEATURES AND BENEFITS

#### Validated marker sets for breast cancer with high prognostic performance

Due to the heterogeneous nature of cancer, current breast cancer biomarkers are generally of low predictive value when tested on populations (datasets) other than those used to identify them. The robustness of the selected biomarkers sets was validated by their ability to predict outcomes in three independent breast cancer datasets (in total, more than 650 samples). NRC1–NRC6 stratified ER+ patients into low-, intermediate-, and high-risk groups, while NRC7–NRC9 stratified ER- patients into low- and high-risk groups. The stratification was highly significant in all three datasets. Furthermore, high predicting rates were reached: the accuracies for ER+ and ER- low-risk groups were 90% and 92–100%, respectively (see Table).

#### Unique marker sets for breast cancer with a broad clinical spectrum

As opposed to other breast cancer prognostic markers that are only applicable to selected groups of patients, these marker sets can be used in all breast cancer patients (i.e., NRC1–NRC6 for ER+ patients and NRC7–NRC9 for ER- patients). Further, the gene lists of the marker sets have not been reported by others.

#### Algorithm with wide-scale applicability

Biomarkers offer promising prospects for the future as diagnostic and prognostic tools. Considering that most algorithms develop so far have not been designed to find markers for cancer subtypes, this novel algorithm offers a unique opportunity to provide the prognostic biomarkers needed for better stratification of patients, and therefore improved selection of therapeutic modalities, across a wide spectrum of cancer types.

#### Improved management of cancer patients

The ability to provide accurate predictions of cancer outcomes stands to have significant impact on health care. Biomarkers of prognosis would allow physicians to choose the most effective treatment for a particular patient by identifying, for example, individuals who could benefit from additional chemotherapy. In turn, this could improve survival duration and quality of life among patients and generate medical cost savings.

### PROTECTION STATUS

Breast prognostic marker sets and the method for finding them (NRC no. 12020).

Accuracies of non-recurrent breast tumors in the low-risk groups using the marker sets NRC1-9

Marker sets	Training set	Test sets		
		1	2	3
ER + samples Marker sets: NRC1--6	89%	94%	91%	87%
Samples	209	173	74	201
ER + samples Marker sets: NRC 7--9	93%	100%	92%	100%
Samples	77	46	43	31

### CONTACTS

#### Yves Quenneville

Tel.: (514) 496-8507

Business Development Officer

E-mail: yves.quenneville@cnrc-nrc.gc.ca

#### Daniel Desmarceaux

Tel.: (514) 496-5300

Business Development Officer

E-mail: daniel.desmarceaux@cnrc-nrc.gc.ca

#### Edwin Wang

Tel.: (514) 496-0914

Computational Chemistry and Biology Group

E-mail: edwin.wang@cnrc-nrc.gc.ca