"Evidence for a Molecular Signature of Metastasis in Primary Solid Tumors"

Supplement

Table of Contents

This document contains summary information regarding the files that can be downloaded from this supplemental information web site (Last updated 11/11/02).

1. **Manuscript** (Mets_Manuscript_Final.pdf) - An Adobe Acrobat .pdf format document containing the submitted version of the manuscript.

2. **Supplement** (Mets_Supplement_Revised_Final_110802_SR.pdf) - An Adobe Acrobat .pdf format document with detailed descriptions of the methods used for this paper's analysis. This document references and describes worksheets in the accompanying Microsoft Excel workbook (referred to as Web Spreadsheets) containing Supplemental Information (see below).

3. **Supplemental Information** (Mets_Supplement_Information_110802_Final_SR.xls) - A Microsoft Excel Workbook containing detailed information regarding the methods and results of the analysis for this manuscript. The contents of each of the individual pages of the workbook are described in the Supplement document. Summaries of the contents of the individual sheets are as follows:

    - **WEB SPREADSHEET A** - Mapping from the Affymetrix HU6800 / HU35KsubA microarray set to Affymetrix U95A microarray.

    - **WEB SPREADSHEET B** - Top 64 markers that are up-regulated and 64 that are down-regulated in the tumor versus metastases comparison from the Global Cancer Map.

    - **WEB SPREADSHEET C** - Colorgram corresponding to the marker selection analysis of WEB SPREADSHEET B.

    - **WEB SPREADSHEET D** - Summary of the prediction results for the tumor versus metastatic distinction (used to pick the number of features that best makes the tumor versus metastases comparison).

    - **WEB SPREADSHEET E** - Permutation test results for the tumor versus metastatic prediction.

    - **WEB SPREADSHEET F** - Mapping from the Affymetrix HU6800 / HU35KsubA microarray set to Affymetrix U95A microarray for the 128 markers from WEB SPREADSHEET B.

    - **WEB SPREADSHEET G** - Hierarchical clustering results for the 128-gene metastasis signature in the lung data.

    - **WEB SPREADSHEET H** - Colorgram for the Hierarchical clustering results from WEB SPREADSHEET F used to create Figure 2a in the manuscript.

- **WEB SPREADSHEET I** - Neighbor analysis for the discovered clusters in the lung dataset.
- **WEB SPREADSHEET J** - Hierarchical clustering results for the 17-gene metastasis signature in the lung data.
- **WEB SPREADSHEET K** - Results from hierarchical clustering of the lung dataset using 17-gene sets derived after 1000 random class label permutations.
- **WEB SPREADSHEET L** - Hierarchical clustering results in the lung data using all of the genes passing filtering.
- **WEB SPREADSHEET M** - Neighbor analysis used to define the refined 17-gene metastasis signature.
- **WEB SPREADSHEET N** - Histogram of the permutation results from WEB SPREADSHEET K.
- **WEB SPREADSHEET O** - Mapping of the 17-gene metastasis signature between the Affymetrix HU6800 / HU35KsubA microarray set and Rosetta oligonucleotide microarray set.
- **WEB SPREADSHEET P** - Hierarchical clustering results for the 17-gene metastasis signature in the Rosetta breast cancer dataset.
- **WEB SPREADSHEET Q** - Hierarchical clustering results for the 17-gene metastasis signature in the prostate cancer dataset.
- **WEB SPREADSHEET R** - Mapping between the Affymetrix HU6800 microarray and Affymetrix U95A microarray for genes from the 17-gene signature.
- **WEB SPREADSHEET S** - Hierarchical clustering results for the 17-gene metastasis signature in the medulloblastoma dataset.
- **WEB SPREADSHEET T** - Hierarchical clustering results for the 17-gene metastasis signature in the LBC lymphoma dataset.
- **WEB SPREADSHEET U** - Gene t-test correlation results for the 17-gene metastasis signature in each of the datasets.
- **WEB SPREADSHEET V** - Gene signal-to-noise calculations for the 17-gene metastasis signature in each of the data sets.
- **WEB SPREADSHEET W** - Analysis of the marker overlap between Rosetta's list of 70 prognostic markers and the 128-gene metastasis signature.
- **WEB SPREADSHEET X** - Summary of the outcome clustering results in each of the datasets.

- **WEB SPREADSHEET Y** - Table summarizing the error rates and *P*-values from the outcome clustering results in each of the datasets.

4. **Datasets**:

- **Dataset A** - Global Cancer Map Tumor vs. Met. (DatasetA_Tum_vs_Met.res) - Data for the Tumor vs. Metastases subset of samples from the Global Cancer Map (this data was previously published in 2001 by Ramaswamy et al. - see referenced web site) contained in Whitehead .res file format. Formed from 64 primary adenocarcinomas and 12 metastatic adenocarcinomas (from lung, breast, prostate, colon, ovary, and uterus), from unmatched patients prior to any treatment. The clinical stage of these primary tumors and outcome is unknown. This file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the primary tumor and metastases samples used for defining the metastases signature in this study. These average difference values were generated by Affymetrix's GeneMicroarray software (MAS4). Associated with each average difference expression number is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneMicroarray software) based on matched and mismatched probes for each gene. The file format is organized such that columns contain data for samples and rows contain data for genes.

- **Dataset B** - Lung Outcome (DatasetB_Lung_outcome.res) - Stage I and II primary lung adenocarcinomas with greater than 4 years of clinical follow-up after surgical resection and a clinical endpoint of overall survival. This data was previously published in 2001 by Bhattacharjee at al. This .res file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the defined subset of lung cancer adenocarcinomas used for discovering and validating the metastasis signature in this study. These average difference values were generated by Affymetrix's GeneMicroarray software (MAS4). Associated with each average difference expression number is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneMicroarray software) based upon the matched and mismatched probes for each gene. The file is organized such that columns contain data for samples and rows contain data for genes.

- **Dataset C** - Rosetta Breast Outcome (DatasetC_Rosetta_breast_outcome.res) - This dataset contains data from 78 stage I primary breast adenocarcinomas with greater than 5 years of clinical follow-up after lumpectomy and a clinical endpoint of time to metastasis. This data was previously published in 2002 by Van't Veer et al. The .res file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the defined subset of breast cancer adenocarcinomas used for testing the discovered metastases signature in this study. The expression values in this .res file come directly from the published data files for the Rosetta oligonucleotide arrays with all measurements getting a P assignment. The file is organized such that columns contain data for samples and rows contain data for genes.

- **Dataset D** - Prostate Outcome (DatasetD_prostate_outcome.res) - This dataset contains data from 21 stage I primary prostate adenocarcinomas with greater than 4 years clinical follow-up after radical prostatectomy and a clinical endpoint of time to PSA relapse after radical prostatectomy. This study was previously published in 2002 by Singh et al. The .res file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the defined subset of prostate cancer adenocarcinomas used for testing the discovered metastases signature in this study. These average difference values were generated by Affymetrix's GeneMicroarray software (MAS4). Associated with each average difference expression number is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneMicroarray software) based upon the matched and mismatched probes for each gene. The file is organized such that columns contain data for samples and rows contain data for genes.

- **Dataset E** - Medulloblastoma Outcome (DatasetE_medulloblastoma_outcome.res) - This dataset contains data from 60 medulloblastomas with greater than 5 years clinical follow-up after multi-modality treatment with a clinical endpoint of overall survival. This study was previously published in 2002 by Pomeroy et al. The .res file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the defined subset of medulloblastoma sample data used for testing the discovered metastases signature in this study. These average difference values were generated by Affymetrix's GeneMicroarray software (MAS4). Associated with each average difference expression number is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneMicroarray software) based upon the matched and mismatched probes for each gene. The file is organized such that columns contain data for samples and rows contain data for genes.

- **Dataset F** - LBC Lymphoma Outcome (DatasetF_lymphoma_outcome.res) - This dataset contains data from 58 large B-cell lymphomas with greater than 5 years of clinical follow-up after combination CHOP chemotherapy with a clinical endpoint of overall survival. This study was previously published in 2002 by Shipp et al. The .res file contains expression values in Affymetrix's scaled average difference units (as described in the supplemental information document) for the large B-cell lymphomas data and is used for testing the discovered metastases signature in this study. These average difference values were generated by Affymetrix's GeneMicroarray software (MAS4). Associated with each average difference expression number is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneMicroarray software) based upon the matched and mismatched probes for each gene. The file is organized such that columns contain data for samples and rows contain data for genes.