Supplementary Material

ColoGuidePro: A prognostic 7-gene expression signature for stage III colorectal cancer patients

Anita Sveen, Trude H. Ågesen, Arild Nesbakken, Gunn Iren Meling, Torleiv O. Rognum, Knut Liestøl, Rolf I. Skotheim, Ragnhild A. Lothe
Supplementary Methods

Microarray gene expression analysis

One μg total RNA from each of the 172 samples were individually processed according to the Affymetrix GeneChip® Whole Transcript (WT) Sense Target Labeling Assay manual (Affymetrix Inc, Santa Clara, CA). Fragmented and labeled sense strand DNA was hybridized onto the Affymetrix GeneChip Human Exon 1.0 ST Array for 16-18 hours. This array contains 1.4 million probe sets of which 289,961 belong to the “core” set of probe sets targeting well annotated full-length human mRNAs (1). Each gene is targeted by an average of 40 probes along the entire length of the coding sequence (corresponding to approximately 10 probe sets), ensuring that the calculated gene expression values are not biased by position of the probing sequences in the individual genes. Array washing, staining and scanning were performed according to the manufacturer’s protocol.

Preprocessing of gene expression data

Raw intensity data from scanned images of the microarrays were preprocessed by the Affymetrix GeneChip Command Console software (version 1.0). The expression intensities were stored as sample-wise cell intensity (CEL) files. Using the CEL-files as input, we applied the robust multi-array average (RMA) algorithm (2) implemented in the Affymetrix Expression Console 1.1 software for inter-chip quantile normalization across each of the two patient series, and summarized the perfect match probes at the gene-level using the Affymetrix HuEx-1_0-st-v2.r2 gene-core library files. The library files defined 22,011 transcript clusters, of which 17,617

Gene-level expression data for the validation series, analyzed on Affymetrix HG-U133 Plus2.0 arrays, were downloaded as CEL-files from GEO (GSE14333 and GSE17538). The raw data were preprocessed using Affymetrix Expression Console 1.0. The CEL-files were background corrected, quantile normalized, and summarized with the RMA algorithm applying the Affymetrix HG-U133_Plus_2.cdf library file. Using the Affymetrix HG-U133_Plus_2.na31.annot.csv annotation file, a total of 54,675 probe sets targeting 41,779 transcripts from annotated genes were identified.

**Filtering of the gene expression dataset from the learning series**

A genome-wide expression dataset of 17,617 genes from 95 stage II and III CRC samples, obtained from Affymetrix GeneChip Human Exon 1.0 ST arrays, was used as a learning series for survival modeling. Since the prognostic value of the resulting gene expression signature was to be tested also in an independent patient series analyzed on Affymetrix HG-U133 Plus2.0 arrays (the validation series), only genes identifiable by gene symbols on both types of arrays were retained (\( n = 15,851 \); annotation based on the human genome build GRCh37). This dataset was filtered to include only genes with expression variances higher than 0.2, and \( P \)-values from univariate Cox proportional hazards analyses below 0.5 (\( n = 3,098 \) genes). \( P \)-values (Wald test of predictive potential) were calculated in R 2.11.1 using the Bioconductor package Weighted Gene
Co-expression Network Analysis (WGCNA) (3). Codes were retrieved from The Comprehensive R Archive Network (CRAN) web pages (4).

Commands used to open the Bioconductor library and set up the working directory:

> library(WGCNA)
> setwd("A") #A = path to working folder

Commands used to set up the data for univariate Cox proportional hazards analyses:

> time<-c(B, ..., C) #B = time to recurrence or censoring for sample 1, C = time to recurrence or censoring for sample 95
> event<-c(D, ..., E) #D = 1 (recurrence) or 0 (censoring) for sample 1, E = 1 (recurrence) or 0 (censoring) for sample 95
> datExpr<-t(as.matrix(read.table("M.txt", header=TRUE, sep="\t", row.names=1, as.is=TRUE))) #M = tab-delimited gene expression signal matrix with sample headings and one column with probe set ids

Commands used to run univariate Cox proportional hazards analyses:

> c<-standardScreeningCensoredTime(time, event, datExpr, fastCalculation=FALSE)

Analysis workflow for development of the prognostic gene expression signature

The filtered gene expression dataset from the learning series of stage II and III CRCs was used as input for the Bioconductor package penalized (5). This analysis tool contains an algorithm for
lasso penalized multivariate Cox proportional hazards modeling of survival data using gene expression signals, performing variable selection and shrinkage (6). The algorithm calculates a penalty parameter ($\lambda_1$) for the input variables based on leave-one out cross-validation of each of the 95 samples. The amount of genes to be included in the prognostic expression signature is a function of $\lambda_1$. The optimal $\lambda_1$-value corresponds to the best performing signature during cross-validation, *i.e.* resulting in the highest cross-validated partial likelihood. Applying higher $\lambda_1$-values will generally reduce the number of genes included in the prognostic expression signature, and the prediction will approach univariate estimation. At lower $\lambda_1$-values, the number of genes included in the signature increases, consequently increasing also the risk of over-fitting the signature to the specific prognostic associations within the learning data, and potentially reducing its predictive power. For the final step in the algorithm, the chosen $\lambda_1$-value was applied as a tuning parameter for lasso penalized modeling of the learning dataset. The analyses were done using version 2.11.1 of R. R codes were retrieved from the CRAN web pages (7).

Commands used to open the Bioconductor library penalized and set up the working directory:

```r
> library(penalized)
> setwd("A") #A = path to working folder
```

Commands used to read and set up the input data:

```r
> s<-Surv(time, event) #'time' and 'event' are vectors as described in the previous section
> exprData<-t(as.matrix(read.table("D.txt", header=TRUE, sep="\t", row.names=1, as.is=TRUE))) #D = tab-delimited gene
```
expression matrix with sample headings and one column with probe set id’s

Commands used to calculate the penalty parameter $\lambda_1$ for the gene expression data by leave-one-out cross-validation:

> opt <- optL1(s, penalized=exprData, fold=10)

Commands used to return the results from the cross-validation at the optimal $\lambda_1$-value:

> opt

Commands used to run and return the results from survival modeling using penalized gene expression signals:

> p<-penalized(s, penalized=exprData, lambda1=E) # $E$ = chosen $\lambda_1$
> show(p)
> coefficients(p)

The survival modeling was repeated 1,000 times. By cross-validation during these iterations, several models were found to accommodate optimal prediction in the learning series. The distribution of cross-validated partial likelihoods achieved an optimal value (approximately -252) for several values of $\lambda_1 > 13$, corresponding to gene expression signatures of various sizes ($\leq 12$ genes; Supplementary Figure 1). Across the iterations, seven different gene expression signatures were found to accommodate optimal survival prediction more than 50 times each (Figure 1). The most frequently reported signature ($n = 202$ times) included only one gene (OLFMA, $\lambda_1$ ranging from 30 to 44). Other optimal signatures included two, seven, eight, nine, eleven, or twelve genes ($\lambda_1$-values corresponding to no active predictors ($\lambda_1 \geq 44$) were also reported). All genes within the smaller signatures were included also in the larger signatures (no gene replacements). The $\lambda_1$-
values for the reported signatures ($\lambda_1 >13$) were all associated with the optimal cross-validated partial likelihood. For all these signatures, except the 1-gene signature, there were significant associations between patient survival and increasing numbers of genes expressed at levels associated with poor survival (genes were considered to be associated with poor patient survival at expression levels above the 80th percentile across the dataset for genes with univariate HR > 1, and below the 20th percentile for genes with univariate HR < 1). The HR (univariate Cox proportional hazards analyses for the 2-gene, 7-gene, 8-gene, 9-gene, 11-gene, and 12-gene signatures) ranged from 1.6 to 1.9 ($P < 0.04$). Further, dichotomization of patients to good and poor prognosis groups was tested for all possible stepwise increases in amounts of genes with associations to poor survival within each signature (Figure 1, heatmap in lower panel). For the 28 possible poor prognosis groups, 22 had significant associations with poor patient survival (univariate Cox proportional hazards analyses, HR ranging from 3.0 to 11.5; $P < 0.04$).

To assess which stratification rule had the best predictive potential on independent samples, the same patient stratification according to the different gene signatures were repeated in the test series. Here, five of the dichotomizing stratification rules resulted in significant prognostic stratification (univariate Cox proportional hazards analyses, HR ranging from 2.9 to 5.8; $P < 0.04$). The best performing stratification rule across both series (by rank of $P$-values from univariate Cox proportional hazards analyses), assigned patients to a poor prognosis group when expressing three or more genes in the 7-gene signature at levels associated with poor prognosis. In the learning series, 23% of the patients were assigned to the poor prognosis group by this stratification rule, while 17% of patients were predicted to have poor prognosis in the test series. By the other four stratification rules having significant associations with patient survival also in
the test series, only \( \leq 6\% \) of the patients were predicted to have poor survival. Hence, for further assessment of prognostic predictive potential, the 7-gene signature was used.

As an inherent part of the lasso survival model, penalized multivariate regression coefficients were calculated for each gene in the 7-gene signature (Supplementary Table 1; penalty parameter \( \lambda_1 = 16 \)). Using these regression coefficients, and the corresponding gene expression values, a prognostic index (PI) was calculated for each patient:

\[
\text{PI} = [(\text{expression value})_{\text{Gene1}} \cdot (\text{regression coefficient})_{\text{Gene1}}] + [(\text{expression value})_{\text{Gene2}} \cdot (\text{regression coefficient})_{\text{Gene2}}] + \ldots
\]

For comparative purposes, the PIs were used to stratify patients according to prognosis, and high PIs indicated poor prognosis.

In the validation series, analyzed on Affymetrix HG-U133 Plus2.0 arrays, there were 12 probe sets targeting the seven genes in the optimal expression signature developed. For redundant probe sets, the sample-wise median expression value was used (Supplementary Table 1).

**Assessment of the 7-gene signature**

In all three patient series, there was a ‘dose-effect’ between amounts of genes in the 7-gene signature predicting poor survival and patient survival rates. That is, the sample-wise increase in amounts of genes with expression levels associated with poor survival (sample-wise amounts ranged from zero to four in the learning series, and zero to five in the test and validation series) was associated with increasingly poorer patient survival. In the learning series, the univariate HR was 1.8 [1.5, 2.3] \((P < 0.001, \text{ Wald test for predictive potential})\). The corresponding HR’s in the
test series and validation series were 1.5 [1.0, 2.1] and 1.6 [1.3, 2.1], respectively ($P = 0.04$ and $< 0.001$, respectively).

For comparison, sample-wise PI’s were calculated based on expression values and lasso-penalized multivariate regression coefficients for genes in the 7-gene signature. The PI’s ranged from -0.2 to 1.4, -0.06 to 1.50, and -0.5 to 0.9 in the learning series, test series and external validation series, respectively. Univariate Cox proportional hazards analyses for the PI’s showed significant survival stratification in all three series (HR = 17.7 [6.7, 46.7], 4.3 [1.2, 16.1], and 4.7 [1.6, 13.8], respectively; $P < 0.03$). To further compare these results to the stratification rule based on counting the sample-wise numbers of genes in the 7-gene signature with associations with poor prognosis, patients were dichotomized by assigning patients with PI > 80th percentile of PI’s across the individual series to poor prognosis groups (to simulate the assignment of approximately 20% of patients to poor prognosis groups when expressing three or more of the seven genes at levels associated with poor prognosis). The univariate HR’s for the learning series, test series and external validation series were 4.1 [2.2, 7.5], 2.1 [0.8, 5.5], and 2.6 [1.4, 4.7], respectively ($P \leq 0.001$, 0.1, and 0.003, respectively). Hence, the performance of the 7-gene signature was similar for stratification based on sample-wise amounts of genes being expressed at levels associated with poor survival and PI-calculations, indicating that the much simpler principle used for stratification in the former approach was a valid replacement for the commonly used regression model strategy in the latter approach.
**R code for calculating and plotting density distributions**

For plotting purposes (Supplementary Figure 2), density distributions of expression signals for genes in the prognostic signature were calculated for patients in the learning series using the Bioconductor package sm (8). The appropriate R code was retrieved from the CRAN web pages (9).

Commands for opening the Bioconductor library and setting up the input data:

```r
> library(sm)
> expr<-c(A, …, B) # A = gene expression signal for patient 1, B = gene expression signal for patient 95
> C<-c(D, …, E) # D = ‘1’ for event (relapse or death) or ‘0’ for censoring for patient 1, E = ‘1’ for event (relapse or death) or ‘0’ for censoring for patient 95
```

This command was used to calculate and plot the density estimates of expression signals for the two assigned groups individually:

```r
> sm.density.compare(expr, group=C)
```
### Table S1. Genes in the prognostic expression signature

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Full name</th>
<th>Function</th>
<th>Cytogenetic band</th>
<th>Exon array ID</th>
<th>Penalized multivariate regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXCL9</td>
<td>Chemokine (C-X-C motif) ligand 9</td>
<td>Involvement in T cell trafficking and affects the growth, movement, or activation state of cells that participate in immune and inflammatory responses</td>
<td>4q21.1</td>
<td>2773947, 203915_at</td>
<td>0.79, -0.10</td>
</tr>
<tr>
<td>DMBT1</td>
<td>Deleted in malignant brain tumors 1</td>
<td>Candidate tumor suppressor gene for brain, lung, esophageal, gastric, and colorectal cancers</td>
<td>10q26.1</td>
<td>4037778, 208250_s_at</td>
<td>1.06, 0.04</td>
</tr>
<tr>
<td>NT5E</td>
<td>5'-nucleotidase, ecto</td>
<td>Hydrolyzes extracellular nucleotides into membrane permeable nucleosides</td>
<td>6q14.3</td>
<td>2915828, 1553994_at &amp; 1553995_at &amp; 203939_at &amp; 227486_at</td>
<td>1.43, 0.05</td>
</tr>
<tr>
<td>OLFM4</td>
<td>Olfactomedin 4</td>
<td>Indications as an antiapoptotic factor promoting tumor growth, and also facilitating cell adhesion</td>
<td>13q14.3</td>
<td>3490892, 212768_s_at</td>
<td>0.91, -0.08</td>
</tr>
<tr>
<td>SEMA3A</td>
<td>Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3A</td>
<td>This secreted protein can function as either a chemorepulsive agent, inhibiting axonal outgrowth, or as a chemoattractive agent, stimulating the growth of apical dendrites. Increased expression of this protein is associated with schizophrenia and is seen in a variety of human tumor cell lines</td>
<td>7p12.1</td>
<td>3059464, 206805_at &amp; 244163_at &amp; 244849_at</td>
<td>1.33, 0.10</td>
</tr>
<tr>
<td>UGT2B17</td>
<td>UDP glucuronosyltransferase 2 family, polypeptide B17</td>
<td>Enzyme catalyzing the transfer of glucuronic acid from uridine diphosphoglucuronic acid to a diverse array of substrates including steroid hormones and lipid-soluble drugs. Glucuronidation is an intermediate step in the metabolism of steroids</td>
<td>4q13.2</td>
<td>2772088, 207245_at</td>
<td>0.78, -0.02</td>
</tr>
<tr>
<td>WNT11</td>
<td>Wingless-type MMTV integration site family, member 11</td>
<td>The WNT gene family consists of structurally related genes which encode secreted signaling proteins. These proteins have been implicated in oncogenesis, e.g. in colorectal cancer, and in several developmental processes, including regulation of cell fate and patterning during embryogenesis</td>
<td>11q13.5</td>
<td>3382523, 206737_at</td>
<td>1.62, 0.10</td>
</tr>
</tbody>
</table>
aAffymetrix GeneChip Human Exon 1.0 ST array transcript cluster ID bAffymetrix HG-U133 Plus2.0 array probe set ID. For genes targeted by multiple probe sets, the median expression value was used cHazard ratios (HR) from univariate Cox proportional hazards analysis in the learning series of colorectal cancers dRegression coefficients from the lasso model at $\lambda_1 = 16$
Table S2. Cox proportional hazards analyses for patients in the learning series

<table>
<thead>
<tr>
<th>Clinical characteristic</th>
<th>Univariate Cox</th>
<th>Multivariate Cox</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR (95% CI)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>P-value&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Gender (male vs. female)</td>
<td>0.9 [0.5, 1.5]</td>
<td>0.6</td>
</tr>
<tr>
<td>Age at diagnosis (≥70 years vs. &lt;70)</td>
<td>1.8 [1.0, 3.2]</td>
<td>0.05</td>
</tr>
<tr>
<td>Tumor stage (III vs. II)</td>
<td>2.0 [1.1, 3.6]</td>
<td>0.02</td>
</tr>
<tr>
<td>Tumor location (right vs. left and rectum)</td>
<td>0.7 [0.3, 1.3]</td>
<td>0.2</td>
</tr>
<tr>
<td>MSI (MSI-high vs. MSI-low and MSS)</td>
<td>0.4 [0.1, 1.6]</td>
<td>0.2</td>
</tr>
<tr>
<td>Gene expression signature</td>
<td>4.0 [2.2, 7.2]</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

<sup>a</sup>HRs and corresponding 95% CIs from univariate or multivariate Cox proportional hazards analysis as indicated. Event = relapse or death from CRC within ten years. Censoring = no events within ten years.

<sup>b</sup>P-values from Wald test of predictive potential. Abbreviations: CI, confidence interval; HR, hazard ratio; MSI; microsatellite instability, MSS; microsatellite stability
Figure S1. Cross-validated partial likelihood and number of active predictors in the learning series as a function of the penalty parameter. The lasso survival modeling algorithm using multivariate and penalized Cox proportional hazards analysis calculated A) the cross-validated partial likelihoods and B) the number of active prognostic predictors as a function of the penalty parameter $\lambda_1$ in the learning series of colorectal cancers. The input data consisted of 3,098 genes with variances in expression levels higher than 0.2, and significance levels from univariate Cox proportional hazards analyses lower than 0.5, as well as information on 10-year relapse-free survival for the patients. The cross-validated partial likelihood represents the potential of the selected active predictors to predict patient survival, as calculated by leave-one out cross-validation, and indicating optimal prediction at high values. The distribution of cross-validated partial likelihoods achieved optimum (approximately -252) for several values of the penalty parameter $\lambda_1 > 13$, corresponding to selection of differently sized gene expression signatures with equal predictive potential (signatures ranging in size from 0 to 12 genes).
Figure S2. Expression in the learning series of the genes in the 7-gene prognostic signature. Density measures of relative amounts of samples (Text S1) are plotted versus gene expression (log-2 transformed). For each gene, the density is shown separately for samples taken from patients with relapse or death from CRC within the ten years of follow-up ($n = 48$; blue or yellow), and samples from patients with no relapse ($n = 47$; grey). Blue indicates genes with associations to poor patient survival at low expression levels ($HR < 1$), and yellow genes with associations to poor survival at high expression levels ($HR > 1$). The pale blue and yellow background areas mark expression levels exceeding the lower and upper 20th percentiles of expression values across the sample series, respectively, which were used as thresholds for designating associations with poor survival for each gene. HR, hazard ratio.
Figure S3. Survival curves for stage III patients in the validation series. Stage III patients in the external validation series who A) did not receive adjuvant chemotherapy and B) did receive adjuvant chemotherapy were significantly stratified according to survival by the 7-gene expression signature. Patients with no available information on post-operative treatment \(n = 16\) were not included in the analyses. Survival curves were calculated by Kaplan-Meier statistics and compared by log-rank tests. Relapse or death from CRC within the five years of follow up was regarded events, and patients with no events within the indicated period of follow-up were censored.
Figure S4. Survival curves for stage III patients in the validation series aged 75 years or older. Older stage III patients in the validation series were significantly separated according to survival by the 7-gene expression signature. Survival curves were calculated by Kaplan-Meier statistics and compared by log-rank tests. Relapse or death from CRC within the five years of follow up was regarded events, and patients with no events within the indicated period of follow-up were censored.
Supplementary References

1. Affymetrix Inc. GeneChip Exon Array Design. 2009. technical noteAvailable from URL:
   http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf

   Exploration, normalization, and summaries of high density oligonucleotide array probe
   level data. Biostatistics 2003;4:249-64.

3. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network

   http://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf

5. Goeman JJ. penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the
   Cox model. 2010. Available from URL: http://cran.r-
   project.org/web/packages/penalized/index.html

6. Goeman JJ. $L_1$ penalized estimation in the Cox proportional hazards model. Biom J
   2009;52:70-84.

7. Goeman JJ. Package 'penalized'. 2010. Available from URL: http://cran.r-
   project.org/web/packages/penalized/penalized.pdf