Supplemental Methods

Much of the following text has been reproduced with few modifications from RAM’s doctoral dissertation [1].

Calculation of Variance Score

To estimate the distribution of probe intensity for each probe on chip $j$, data from the other chips is used. First, an artifact adjusted mean $\mu_j$ is calculated by using the expression values $x_i$ from other chips ($i \neq j$) in the dataset and artifact attenuation factor $\alpha_i$. As a default, the attenuation factor equals one for spots that have not been identified as artifactual, and zero for spots that have been identified as artifactual. After this artifact adjusted mean is calculated, an artifact adjusted deviation score $\sigma_j$ is calculated in a similar manner. These two statistics, $\mu_j$ and $\sigma_j$ thus form a description of the null distribution against which $x_j$ is to be compared. The final variance statistic $z_j$ for a location is then calculated as the difference between the observed intensity $x_j$ and the weighted mean $\mu_j$, divided by the weighted standard deviation $\sigma_j$.

$$\alpha_i = \begin{cases} 0 : & \text{artifact} \\ 1 : & \text{not artifact} \end{cases} \quad \mu_j = \frac{\sum_{i \neq j} x_i \cdot \alpha_i}{\sum \alpha_i}$$

$$z_j = \frac{x_j - \mu_j}{\sigma_j}$$

The result of this calculation is a variance statistic, $z$, for each spot on each chip in the study. A high magnitude $z$ indicates artifactual tendency, while a low magnitude $z$
suggests that the spot is to be trusted. Note that a high magnitude score could also be a result of biologically relevant gene expression, but those cases will be generally ignored during the artifact identification procedure described in the next section. A nonlinear scaling procedure is then applied to each $z$ to yield $h$, which is a score between 0 and 1 that has been adjusted for the number of chips in the dataset, and is similar, but not equivalent to 1 minus a p-value. This adjustment is made to provide consistent visualization and image processing

$$h = 1 - \exp \left( - \left( \frac{z_j}{n} \right)^2 \right)$$

In the above formula, $n$ is a scaling factor equal to the ratio of the $p=0.05$ critical point in the student’s $t$ distribution with degrees of freedom equal to one less than the number of chips used in the calculation of $z_j$, to the $p=0.05$ critical point in the Gaussian distribution. Since the number $n$ increases with decreasing degrees of freedom, this scaling factor helps ensure that successive rounds of caCORRECT do not progressively and indefinitely shave off more and more outliers until there is nothing left. Importantly, $h$ is set to zero if the number of non-artifact chips drops below 3. This statistic is used as a guideline for identifying regions of chips where artifacts are present.

**Gene expression model implementation**

If we assume that $X_p$ is rank 1, (that it can be perfectly approximated by the multiplication of a single column vector $\theta_p$ and single row vector $a_p$), then we can use a shortcut to estimate $\theta_p$ and $a_p$ directly from $X_p$ without using an SVD. In such a
situation the summation of the $b^{th}$ column of $X_p$ would give a clue to the $b^{th}$ element of $a_p$ as follows.

$$
\sum_{j=1}^{N}(x_{b,p,j}) = \sum_{j=1}^{N}(\theta_{p,j}a_{b,p}) = a_{b,p}\sum_{j=1}^{N}(\theta_{p,j})
$$

This $b^{th}$ column summation of $X_p$ is defined as $\varphi_{b,p}$, which can then be combined into a row vector, $\varphi_p$.

$$
\varphi_p = [\varphi_{1,p}, \ldots, \varphi_{b_p,p}] = [a_{1,p}, \ldots, a_{b_p,p}\sum_{j=1}^{N}(\theta_{p,j})]
$$

In this form, we can see that $\varphi_p$ is simply a scaled version of $a_p$. Because of the constraint on $a_p$ that geometric mean be set to 1, $a_p$ can be easily derived from $\varphi_p$ by normalizing $\varphi_p$ by its geometric mean as follows.

$$
a_p = \frac{\varphi_p}{\sqrt[\varphi_{b,p}]{\prod_{b=1}^{b_p}\varphi_{b,p}}}
$$

Similarly, summation of the $j^{th}$ row of $X_p$ would give a clue to $\theta_{p,j}$.

$$
\sum_{i=1}^{B_p}(x_{b,p,j}) = \sum_{i=1}^{B_p}(\theta_{p,j}a_{b,p}) = \theta_{p,j}\sum_{b=1}^{B_p}(a_{b,p})
$$

Since the summation of $a_p$ is calculable, so is each element of $\theta_p$.

$$
\theta_{p,j} = \frac{\sum_{b=1}^{B_p}(x_{b,p,j})}{\sum_{b=1}^{B_p}(a_{b,p})}
$$
Biomarker Selection

All gene expression calculations for RCC Affymetrix chips were obtained using the RMA algorithm, using RMAExpress [2-4]. RMAExpress is a Windows GUI for the Robust Multichip Average (RMA) method commonly used through Bioconductor. RMAExpress was chosen here for its speed, lack of selectable options, and ability to easily produce residual images for each chip. When using RMAExpress, all data were processed using both the quantile normalization and background correction features included in the software.

To determine the ranked gene list for each dataset we used one-dimensional linear Support Vector Machines (SVM.) SVM was chosen because of its ability to generalize well, and its reliable error estimation, which make it ideal for selecting ranked lists of candidate biomarkers. The SVM was implemented in parallel on a 74-node cluster for speed and convenience, and is accessed through a web-based GUI, called omniBiomarker (http://omnibiomarker.bme.gatech.edu). The SVM classifier used a Gaussian bolstering kernel with a bolstering radius of 1.4826 and 100,000 points of Monte Carlo integration for error estimation. For more information on our implementation of SVM, its application to microarray study, and sensitivity to quality control, we refer you previous work [5-8].

The classification error output by the omniBiomarker SVM was then used as a basis to rank each gene.

Two separate ranked lists of biomarkers were produced from the RCC Microarray data before and after preprocessing with caCORRECT, referred to as pre-QC and post-QC lists respectively. Pre-QC and post-QC biomarker lists were then compared to select out subsets of genes fulfilling the following criteria:
Set 1: The gene must be inside the top 100 of the pre-QC list, and not in the top 100 of the post-QC list.

Set 2: The gene must be inside the top 100 of the post-QC list, and not in the top 100 of the pre-QC list.

Set 3: The gene must be inside the top 100 of both the pre-QC list, and the post-QC list.

The genes in each of these sets were then reduced, based on availability of suitable PCR primers.

**Independent Biomarker Validation by PCR**

Fifteen of the genes meeting these criteria were then quantified with PCR on a cohort of independent RCC samples. 8 samples each of Clear Cell and Chromophobe RCC were used for the pilot study, with cDNA extractions from adjacent slices of each FFPE tissue block constituting duplicate samples.

This PCR work was then expanded to a panel of 96 genes as described earlier in the main methods section. This panel was tested by the microarray core facility at Emory University using even more independent tissue samples. In the main manuscript this 96-gene panel is referred to as batch II, while the previous 15-gene panel is referred to as batch I.

**References**


