List of Supplementary Materials:

Supplementary Online Methods (SOM)

Fig S1-S9

References (71-74)

Supplementary Online Methods

Datasets

We analyzed seven published gene expression data sets (**table S1**). For each dataset, we generated an expression matrix X (genes by samples). These include bulk RNA-Seq normalized by RPKM (GTEx, (*14*)) or RSEM (TCGA (*45*)), microarrays (ImmGen (*44*)), and single-cell RNA-Seq processed into UMI counts ((*48*)) or normalized by TPM ((*47,49*)). For each dataset, we capped expression at the 99.5th percentile value. Unless otherwise noted, the data were not further normalized beyond the normalization applied in the original publication.

Simulating compositional measurements

For each measurement vector (*i.e.* one row of matrix A), we randomly sampled g i.i.d. Gaussian variables, where g depended on the number of genes in a data set. These vectors were scaled to have a unit norm. We generated multiple measurement vectors sequentially, discarding any that had a modest correlation (>20%) with any already existing vector, to reduce linear dependency between the measurements. The matrix A is a vertical concatenation of m such measurement vectors, so that A has dimensions $m\ x\ g$.

In order to simulate m noisy compositional observations for each of s samples, we took the matrix product:

$$Y = A(X + noise)$$

with i.i.d. random Gaussian noise added to each element of X (*i.e.*, $noise$ is a random matrix of the same size as X). Thus, each element in a column of Y represents a linear combination of noisy expression levels in X according to the weights given in a row of A. The magnitude of the noisy components was set by a signal-to-noise ratio (e.g. $\frac{\|X\|}{\|noise\|} = 2$).

Sample-to-sample distances in low-dimensional embeddings

The Johnson-Lindenstrauss lemma provides bounds on the distortion of Euclidean distances for points mapped to a low-dimensional embedding. Specifically, the lemma states that there exists a Lipschitz mapping $f\colon \mathbb{R}^d \to \mathbb{R}^k$ for n points such that for any two points, u, v:

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq \|u-v\|^2(1+\epsilon)$$

where $k \geq O\left(\frac{\log n}{\epsilon^2}\right)$. Here, we do not search for the optimal embedding, but rather we project gene expression profiles onto a random embedding defined by the matrix A. We then calculate the Spearman rank and Pearson correlation coefficients between pairwise Euclidean distances for columns in Y and the corresponding distances for columns in X. The reported results (**table S2**) reflect the average of 50 random trials for each data set and a given number of measurements. For each trial, a new random measurement matrix was generated, and all pairwise distances were calculated for a maximum of 200 randomly chosen samples.

To determine if clusters of samples generated from the low dimensional embedding resembled clusters generated from the original data, we performed spectral clustering (n=10 clusters) on the low dimensional and the original data, and then calculated cluster similarity by the adjusted mutual information score. Briefly, the mutual information between two sets of clusters quantifies how much information the clustering in low dimensional space provides about the clusters in high dimensional space. If all of the members of any given low dimensional cluster were also clustered together in high dimension, and vice versa, then the mutual information is 1. If one set of clusters appears random relative to the other, then the mutual information is 0. The adjusted mutual information score accounts for a bias with increasing cluster numbers, where the mutual information will tend to increase. Both spectral clustering and similarity measures were implemented with scikit-learn (*71*) in Python 2.7.

High-dimensional clusters were also compared with clusters derived from high-dimensional gene expression, with the addition of noise (**figure S1**). The same clustering parameters were used, with the noisy input of $(X + noise)$ and an SNR=2. In datasets with clusters that were robust to noise (*i.e.* well separated), these clusters were highly similar to noiseless clusters (as quantified by mutual information).

Module activity by matrix factorization

Our model of gene expression posits that abundance levels can be approximated as a linear combination of gene modules. Very generally, any decomposition of multiple expression profiles in the following form fits with this interpretation:

$$X \approx UW$$

Each column of U is a vector of length g (with one entry per gene), and the coefficients in each column of W describe a linear combination of columns in U. We refer to the matrix U as the module dictionary – each column is one element of the dictionary – and the matrix W as the set of module activities. There are multiple ways to factorize any given expression matrix. We explored three different algorithms that apply different constraints to the dictionary, U, and activities, W.

In **Singular Value Decomposition (SVD)**, the columns of the module dictionary are orthonormal. This is an analytically convenient constraint, but there is no reason to suppose that it holds any particular relevance for gene expression.

In **Nonnegative Matrix Factorization (NMF)**, the entries of the module dictionary and the activities are nonnegative. With **sparse NMF**, additional sparsity constraints are enforced on the activity levels, such that there are relatively few nonzero entries in W (this is a soft constraint on sparsity that does not explicitly enforce k-sparsity).

We also developed a third algorithm here, **Sparse Module Activity Factorization (SMAF)**, which is a simple modification of sparse NMF to enforce *k*-sparsity in the module activity, to allow for negative module activity (repression), and to enforce (soft) sparsity in the module dictionary. Like some algorithms for sparse NMF, SMAF optimization proceeds through alternating updates to U and W:

$(1)\ initialize\ U, W$

$(2)\ update\ U\colon\ \min_{U} \|U\|_1\ such\ that\ \|X - UW\|_2^2 < \lambda_U, u_{i,j} \geq 0, and\ \|u_i\|_2 = 1\ for\ all\ i, j$

$(3)\ update\ W\colon\ \min_{W} \|X - UW\|_2^2\ such\ that\ \|w_i\|_0 < k\ for\ all\ i$

Step 2 is optimized by Lasso, and Step 3 is optimized by Orthogonal Matching Pursuit (OMP). Steps 2 and 3 can be iterated until convergence, or until a desired sparsity level is reached. The algorithm can be initialized with the output of SVD or sparse NMF. However, in practice, we found the best results with a random initialization, although this takes a larger number of iterations to converge. The parameter λ_U can be used to set a desired level of accuracy in Step 2.

Our implementations of SVD, sparse NMF and SMAF algorithms can be found in GitHub (https://github.com/brian-cleary/CS-SMAF), and make particular use of Sparse Modeling Software (SPAMS) for Python (*72*).

With each algorithm we can specify the number of dictionary elements. For SVD and sparse NMF, we used a truncated decomposition, keeping the vectors corresponding to the largest singular values. We used the minimally sized dictionary with at least a 99% fit to the original data:

$$fit = 1 - \frac{\|X - UW\|_2^2}{\|X\|_2^2}$$

With SMAF, we set the desired fit according to constraints in steps 2 and 3. For very sparse dictionary elements, we would expect that in order to achieve the same fit as sparse NMF, more dictionary elements overall would be needed. We therefore set the SMAF dictionary size to be 4 times the size of the sparse NMF dictionary, without being larger than min $(1000, 1.5\ x\ \#samples)$.

To quantify the effective module sizes and activity levels for each matrix factorization, we calculated the Shannon diversity of the absolute values in each column of U or W:

$$effective\ diversity_l = e^{entropy(|u_l|)}$$

where $|u_l|$ denotes the absolute value of coefficients in column l.

Gene set enrichments

Gene set enrichments in each module were calculated using Homer (*73*). For each module, the module genes were determined by keeping the top genes, sorted by absolute value, up to 99% (or 50% for truncated) of the total module 'weight' ($mw_l = \sum_i^g u_{i,l}^2$). Significantly enriched gene sets in the Molecular Signatures Database (MSigDB) (*57*) were calculated using the list of all genes that participated in at least one module as a background set. Cutoff levels were based on an FDR q-value of 0.05, and only the top 5 gene sets per module were considered.

Compressed sensing for gene expression profiles

For each dataset we simulated 50 random trials of gene expression recovery using noisy composite measurements and a dictionary learned from training data. In a given trial, we randomly selected 5,000 genes (for both computational efficiency and to check for robustness to random subsets of genes), and then learned a module dictionary from a set of training samples (5% of all available samples, selected uniformly at random without replacement). The module dictionary, U, was given by SVD, sparse NMF, or SMAF. Observations in testing samples (*i.e.*, 95% of all available) were calculated as:

$$Y = A(X_{test} + noise)$$

Then, using the training dictionary we search for k-sparse module activities such that:

$$Y \approx AU\widehat{W} \ (eq. 1)$$

where $\widehat{W}$ is an unknown set of sparsely-populated module activity coefficients. After optimizing $\widehat{W}$ by OMP, and enforcing that each column has only k nonzero values, we recover the predicted gene abundances for each sample as $\hat{X} = U\widehat{W}$. The sparsity level,

k, was set to 15. Increasing the sparsity to $k = 5$ did not dramatically alter the results (data not shown).

To demonstrate the significance of sparsity and random Gaussian measurements in this context, we briefly review some key concepts in compressed sensing. First, recall that if the true activity levels are not sparse in the chosen dictionary (*e.g.,* with U_{SVD} we expect many nonzero coefficients in the module activity matrix), and if there are far fewer measurements than genes, then the problem is ill posed. This can be seen by rewriting $eq.$ 1 as:

$$Y \approx DW$$

and noting that if D has fewer rows than columns, then the problem will not have a unique solution (again, if W is dense). That is, for a given composite observation y_i we will not be able to determine the true activity levels because there will be many indistinguishable solutions, such that $y_i = D\widehat{w}_i$ and $y_i = D\widetilde{w}_i$ where $\widehat{w}_i \neq \widetilde{w}_i$. With the constraint that activity levels must be k-sparse, such solutions may be unique. The process might also fail if W is sparse, but the columns of D are linearly dependent. In compressed sensing it is common to control such linear dependency by bounding the spark of D. Briefly, the spark of a matrix is the smallest number of columns that are linearly dependent. If $spark(D) > \ 2k$, then, for a given k-sparse w_i, the set of compositional observations $y_i = Dw_i$ are uniquely determined, meaning that $y_i \neq Dw_j$ if $w_i \neq w_j$. In other words, with sufficient linear independency in D, two samples will not result in the same set of compositional observations, unless those samples have identical expression profiles. By choosing random Gaussian entries in the measurement matrix, the spark condition is satisfied with high probability when U is orthonormal.

Measures of correlation between predicted and observed data

We used several measures to compare predicted and observed expression levels.

We first computed “overall” correlations. For these measures we considered all genes from all samples together by flattening the predicted and observed expression matrices into vectors, and then computing the correlation between the two vectors. We computed both the Pearson and Spearman rank correlation. Pearson correlation can be more sensitive to accuracy in the dynamic range of predicted values, but can also be dominated by very large values. When the data are not normally distributed, the Spearman rank correlation is often a better overall indicator of performance.

We also considered the average Pearson correlation in gene- and sample-centric views. For the average gene correlation, we calculated the correlations across all samples for a given gene, and then averaged these correlations across all genes. Similarly, the average sample correlation was calculated from the correlation across all genes within a sample.

Blind compressed sensing with SMAF (BCS-SMAF)

Our BCS-SMAF algorithm follows the conceptual steps of Aghagolzadeh and Radha (*62*). A critical element of the algorithm is the use of variable Gaussian measurements; for each sample we generate m compositional observations using different measurement matrices, A_i. The algorithm proceeds as follows:

1. Get initial estimates of each sample as: $\hat{x}_i = A_i^T (A_i^T A_i)^{-1} y_i$.
2. Based on current estimates, calculate $SMAF(\hat{X})$: $\hat{X} \approx \tilde{U}\tilde{W}$.

3. Update our estimate of the dictionary with a standard dictionary learning (DL) algorithm, using $\widetilde{U}$ for initialization: $\widehat{U} = DL(\hat{X}, \widetilde{U})$.
4. Estimate module activities using OMP: for each column $\widehat{w}_i = OMP(y_i, A_i\widehat{U}, k)$.
5. Iterate steps 2-4.
6. Return the estimated signals: $\hat{X} = \widehat{U}\widehat{W}$.

Our code, available at https://github.com/brian-cleary/CS-SMAF, provides an implementation of BCS-SMAF following the steps above. In step 3, we use a dictionary learning algorithm provided in the SPAMS library for Python (*74*).

Composite measurements by hybridization and ligation-mediated amplification

Probe library generation

Each probe library has two groups of probes, upstream probes and downstream probes. Each of the probes will bind a short sequence on the target transcript, in this specific case the RNA (mRNA) molecule within a cell or sample. The two binding sites are juxtaposed so that once bound to target they can be ligated to yield a single ligation product, the abundance of which will then be measured. To ensure efficient ligation, we used T4 Polynucleotide Kinase (New England Biolabs) to phosphorylate the 5' end of downstream probe by incubation the probes under 37ºC with T4 PNK enzyme, T4 ligase buffer, and ATP solution for 2 hours, and heat-inactivated the enzyme at 95ºC for 10 minutes. The phosphorylated probes were then diluted to 5 uM and combined with upstream probes to form the probe pair mix for each individual gene at a final concentration of 2.5 uM per probe. The probe pairs for each target gene were then mixed as the designs specified by measurement matrix to yield the final detection probe library.

For large-scale library generation, Echo 550 Liquid Handler (Labcyte Inc.) was used to generate probe library mix from a source plate containing all gene-specific probe pairs.

Cell culture and RNA sample preparation

Human chronic myelogenous leukemia (CML) K-562 cells were cultured as recommended by the manufacturer (ATCC). Briefly, cells were maintained in Iscove's Modified Dulbecco's Medium (IMEM) supplemented with 10% FBS (HyClone), 2 mM GlutaMAX (Life Technologies), 100 U/ml penicillin, and 100 µg/ml streptomycin at 37ºC with 5% CO_2 incubation. Cells were seeded at a density of 1 million cells per mL for each subculture and a minimum of 2 million cells were used for each RNA extraction. Total RNA from cells was extracted using the RNease Mini Plus Kit (Qiagen) and normalized to 50ng per microliter concentration prior to downstream processing. For detection of background binding, genomic DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen).

To establish independent references for gene expression levels of all targets in the cell, we separately prepared a cDNA library from the same pool of extracted total RNA using qScript cDNA SuperMix (QuantaBio). The expression profile of each target gene used in our study was quantified individually by qRT-PCR, with gene specific primers and the PowerUp™ SYBR Green Master Mix (Thermo Fisher Scientific) in 384-well plates with at least four replicates per reaction in the LightCycler 480 Instrument (Roche Life Science). The final set of references was then calculated after second derivative maximum analysis as relative expression values.

Probe library hybridization and detection

A probe library corresponding to the measurement matrix was added to the extracted RNA sample to a ladder of different concentrations to detect the optimal level of probe mix under each experimental condition. For the final sets of experiment, 10ng of total RNA were used for each reaction, and a final concentration of 25pM or 0.25nM probe mix were added to the RNA sample to a final reaction volume of 20uL along with SplintR ligase buffer (New England Biolabs) (*66*) and RNase Inhibitor. To hybridize the probe library to the RNA sample, a slow ramping protocol was applied by incubating the mixture first at 75ºC for 5 minutes for denaturing of any possible RNA/DNA secondary structure, then slowly ramping down to 37ºC at a ramping rate of 0.1ºC per second with one minute incubation for every degree drop in temperature in a cycling manner over ~ 4 hours in Mastercycler Pro thermocycler (Eppendorf).

The hybridized mixture containing probe pairs bound to the RNA samples was then subjected to two different downstream processing workflows. For the first pipeline, hybridized samples were purified with poly-T conjugated magnetic beads (New England Biolabs) or the Poly-T mRNA purification Dynabeads (Thermo Fisher Scientific), according to the manufacturer's recommended protocol with slight modification to adapt to the small volume and large number of samples used in our experimental set-up. The purified samples were then ligated according to the protocol detailed below. For the second pipeline, the samples were processed in reverse order, where ligation reactions were first carried out followed by bead purification of ligated samples. For the ligation step, the hybridized library-RNA mix was ligated using SplintR ligase (New England Biolabs) or Taq DNA ligase (New England Biolabs) with additional RNase Inhibitor to generate the ligated products at 37ºC and 42ºC (for SplintR ligase) or 45ºC (for Taq ligase) for a period of 4 hours (Splint ligase) or 6 hours (Taq ligase). The enzymatic activity was then heat-inactivated at 65ºC for 20 minutes. For control experiments, negative control reactions without the addition of extracted RNA samples were also hybridized and ligated at the same time under the same experimental conditions.

All samples were then subjected to the same detection protocol at the same time to minimize variability between experiments. Briefly, 2uL of the ligation product were added to reaction mix for amplification using the library-amplification primer designed based on the measurement-specific adapter sequence of the probe library. To measure the abundance of ligated products in each condition we qRT-PCR with PowerUp™ SYBR Green Master Mix (Thermo Fisher Scientific) in 384-well plates with at least four replicates per reaction in the LightCycler 480 Instrument (Roche Life Science). Data analysis were carried out using second derivative maximum method followed by normalization of probe abundance value with reference to the total amount of probes detected in each experiment.

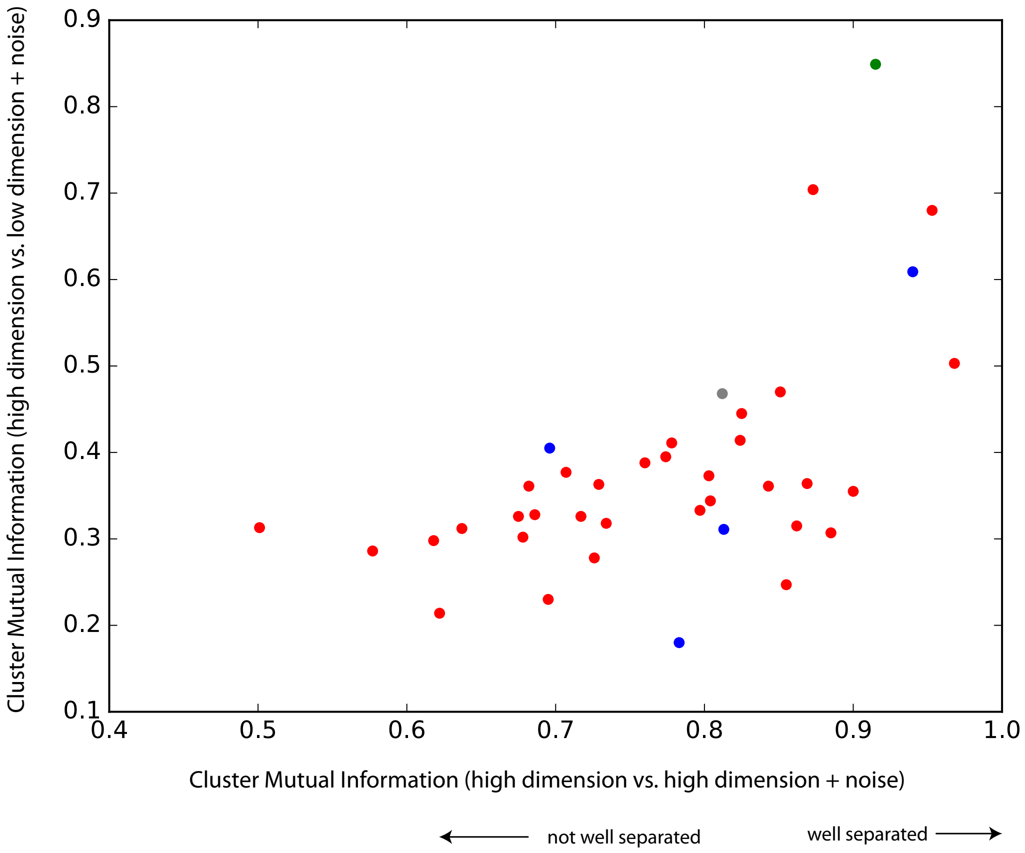

Figure S1: Clusters in well-separated datasets are preserved in low dimension

For each dataset (red: individual TCGA tumor types; green: GTEx; grey: TCGA "combined" dataset; blue: all other datasets) clusters were created from high-dimensional gene expression, gene expression plus random noise (SNR=2), and low-dimensional noisy projections (100 projections, SNR=2). Scatter plot shows the mutual information between high-dimensional clusters and high-dimensional plus noise clusters (x axis) compared to the mutual information between high-dimensional clusters and noisy low-dimensional clusters (y axis). Similar levels of mutual information were found with an alternative clustering method that does not specify the number of clusters (data not shown).

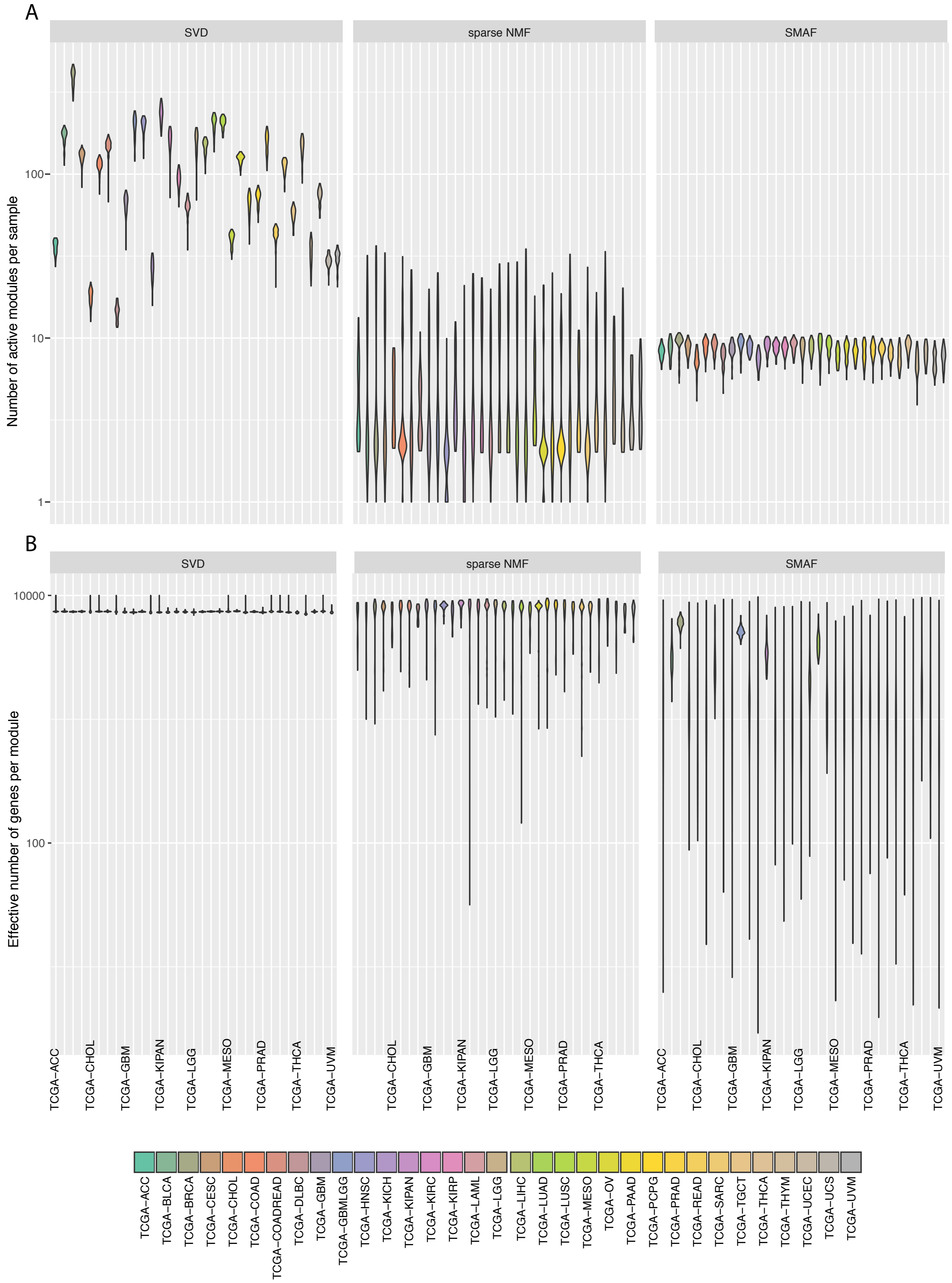

Figure S2: Modules across TCGA tumor types

For each of three algorithms, SVD (left), sNMF (middle), and SMAF (right), violin plots indicate the distribution of number of active modules per sample (top) and the effective number of genes per module (bottom) for each TCGA tumor type (columns).

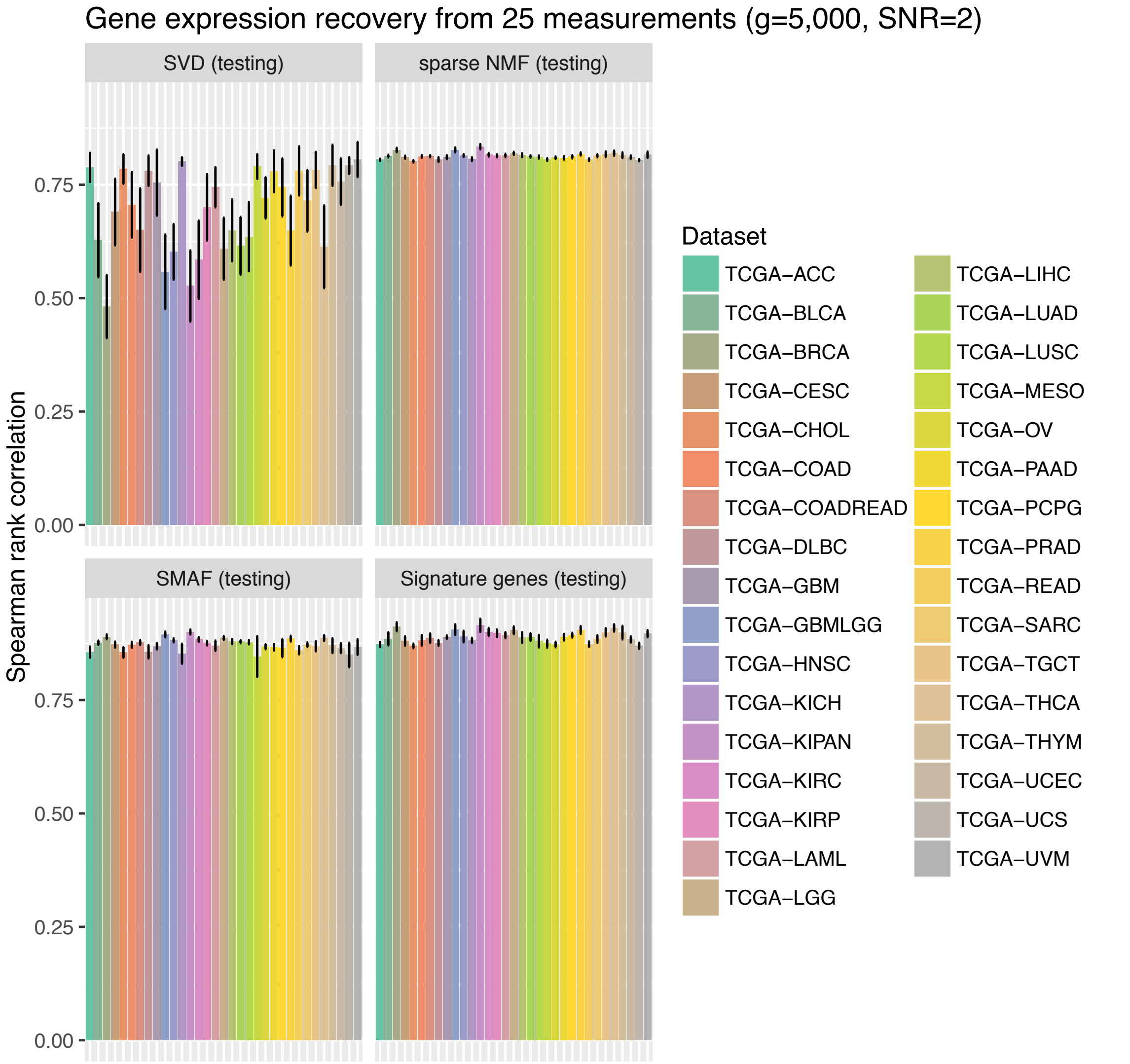

Figure S3: Compressed sensing across TCGA tumor types

For each of three algorithms (SVD, sNMF, and SMAF) module dictionaries were calculated in training data for each tumor type (individual columns). In testing data, 25 noisy composite measurements of 5,000 genes were simulated. These measurements were used to estimate module activity levels, and to predict 5,000 expression values. Compressed sensing was compared to measuring 25 signature genes, and using a model built on training data to predict the remaining unobserved genes (bottom right panel). Bar heights indicate the Spearman correlation between predicted and actual values. Error bars indicate one standard deviation across 50 random trials.

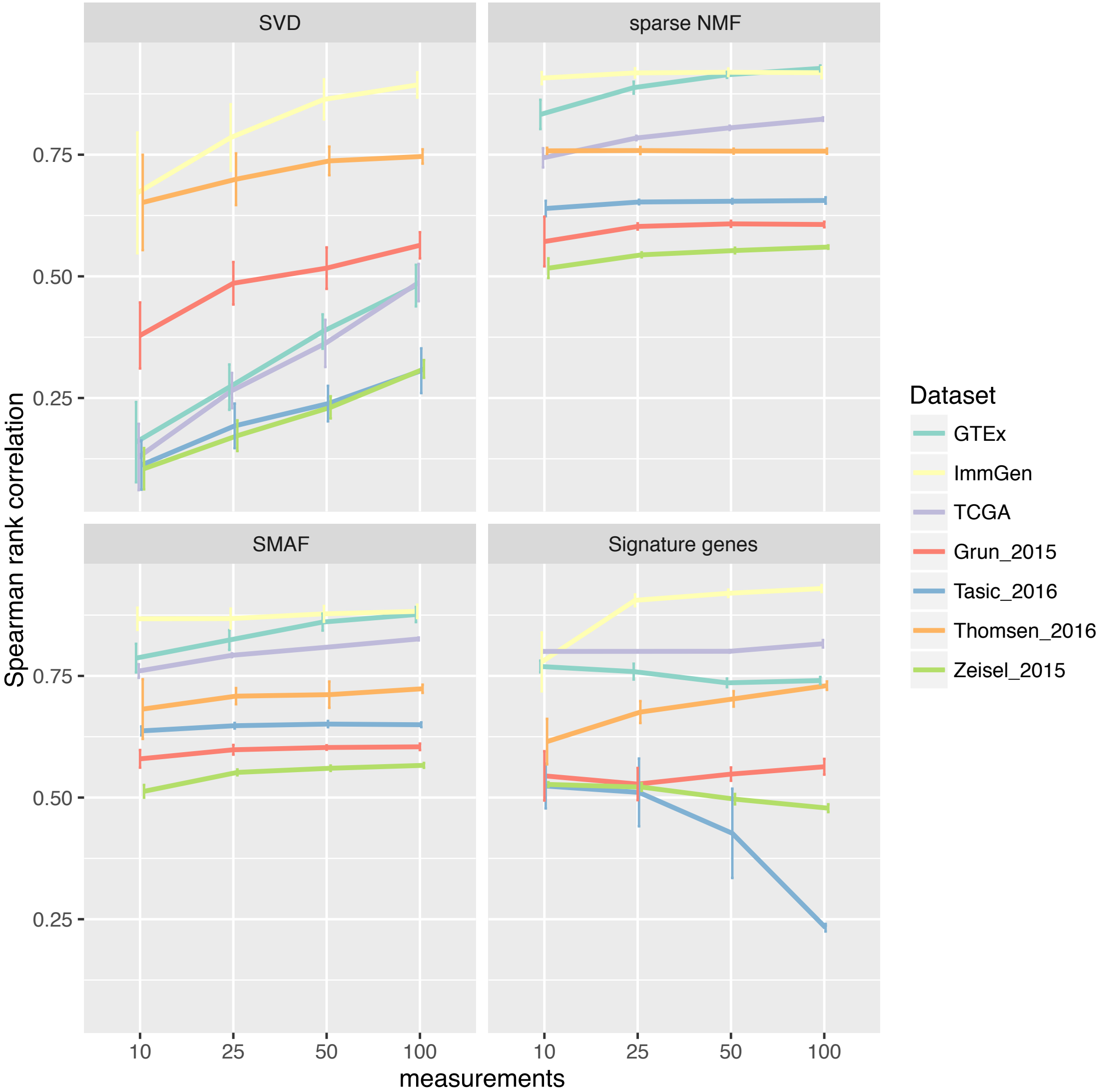

Figure S4: Compressed sensing with a different numbers of measurements

Compressed sensing and signature gene models were used to predict 5,000 gene expression values on the basis of 10-100 measurements (SNR=2, x axis). For each dataset (colors) and model (panels), the plotted data indicate the Spearman correlation (y axis) between predicted and actual values. For signature genes, different models (built on training data) and measurements (simulated in testing) are used for each number of measurements, while for compressed sensing the same module dictionary and measurement design (random composite) can be used in each experiment.

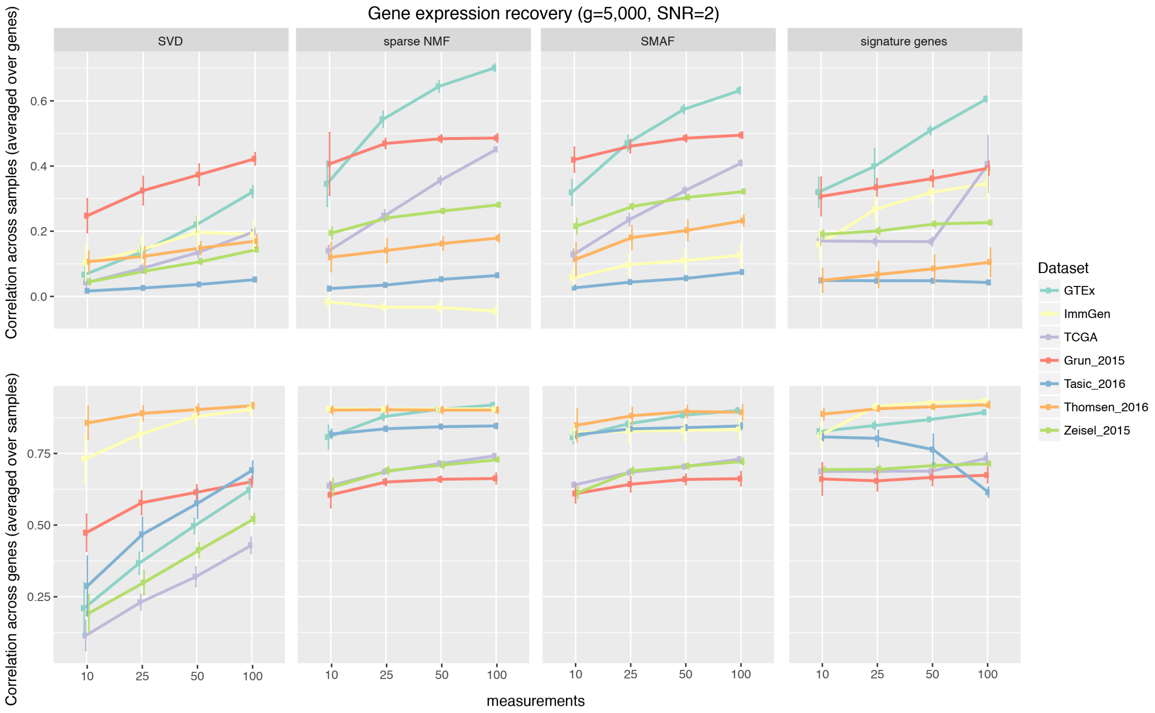

Figure S5: Gene expression recovery from different numbers of composite measurements as assessed by correlations averaged overs genes or over samples

In random trials, a variable number of measurements (x-axis) were used to predict 5,000 expression levels. Shown are the average Pearson correlation coefficients (y axis) across samples, averaged over all genes (top), or across genes, averaged over all samples (bottom) for each dataset (color code, legend). Error bars indicate one standard deviation across 50 random trials.

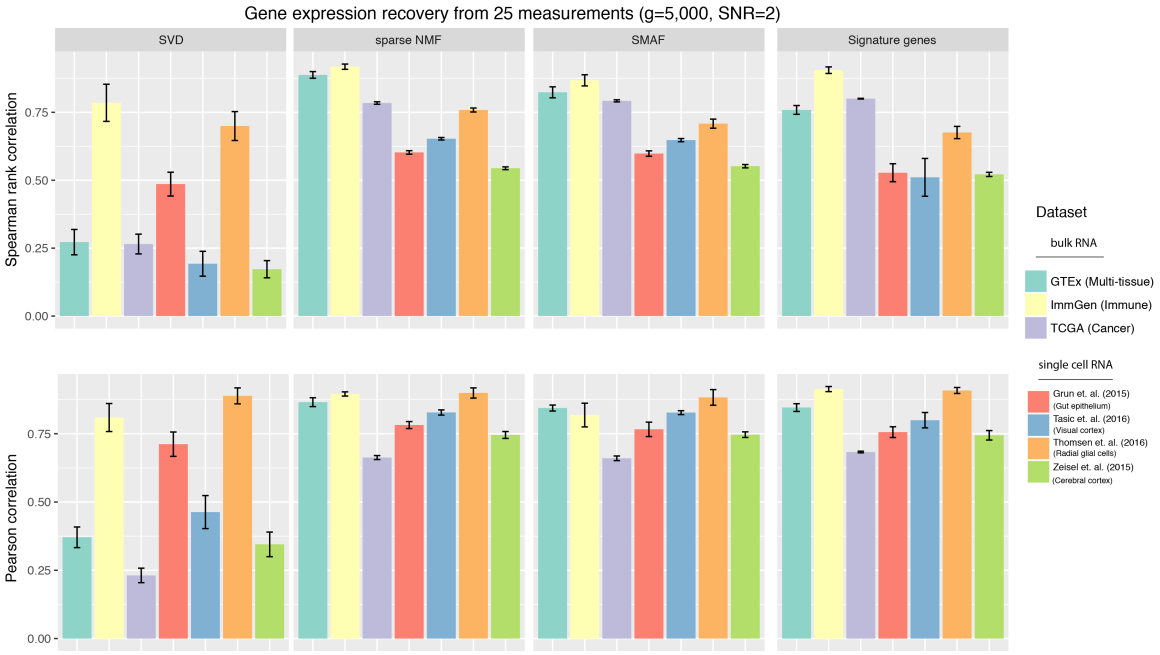

Figure S6: Gene expression recovery from 25 measurements as assessed by Pearson *vs.* Spearman correlation

In each dataset 25 random composite measurements were used to predict the abundance of 5,000 genes. Shown is the Spearman (top), or Pearson (bottom) correlation coefficient (y axis) between predicted and observed values. Error bars indicate one standard deviation across 50 random trials.

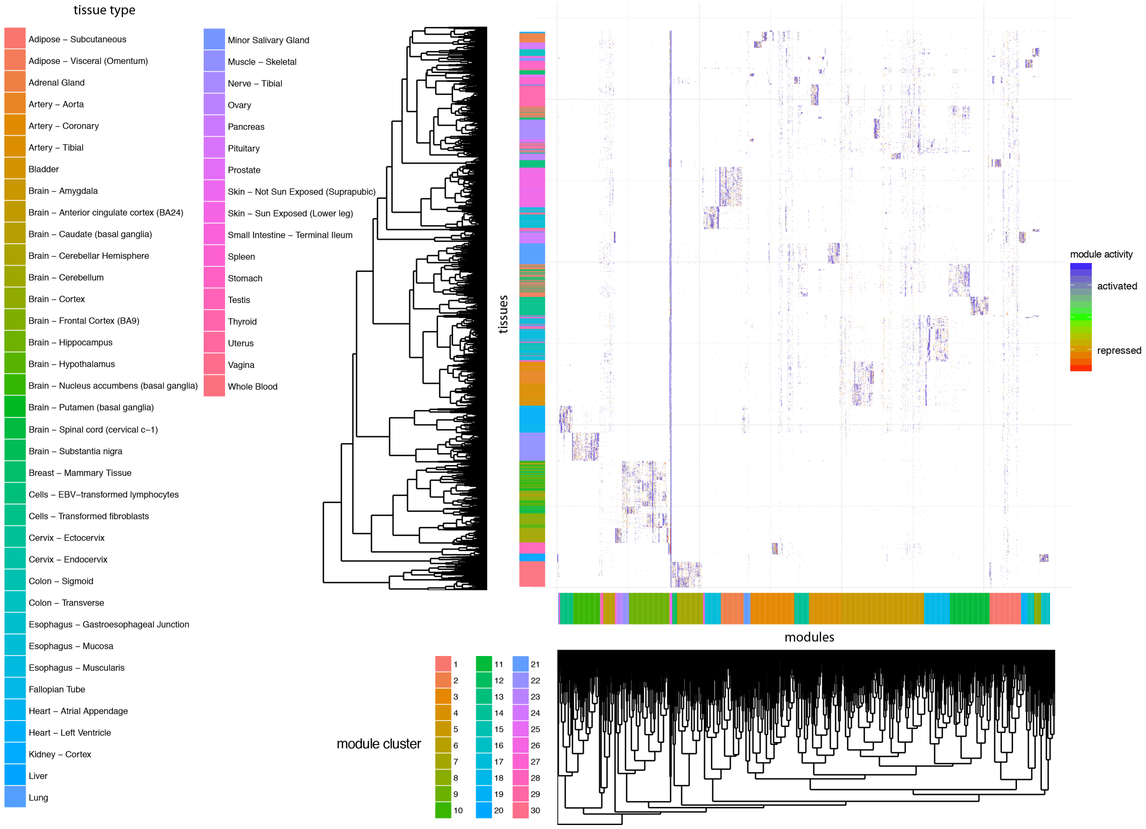

Figure S7: SMAF module activities in each GTEx sample

SMAF was used to calculate a module dictionary and module activity levels in each sample of the GTEx dataset. Heatmap shows the activity (blue: activated, red: repressed, white: not active; colorbar, right) of the modules (columns) in each sample (rows), where modules are hierarchically clustered by their similarity in the module dictionary (*i.e.* by cosine similarity of the module weights across genes), and samples are clustered by their similarity in high-dimensional gene expression levels.

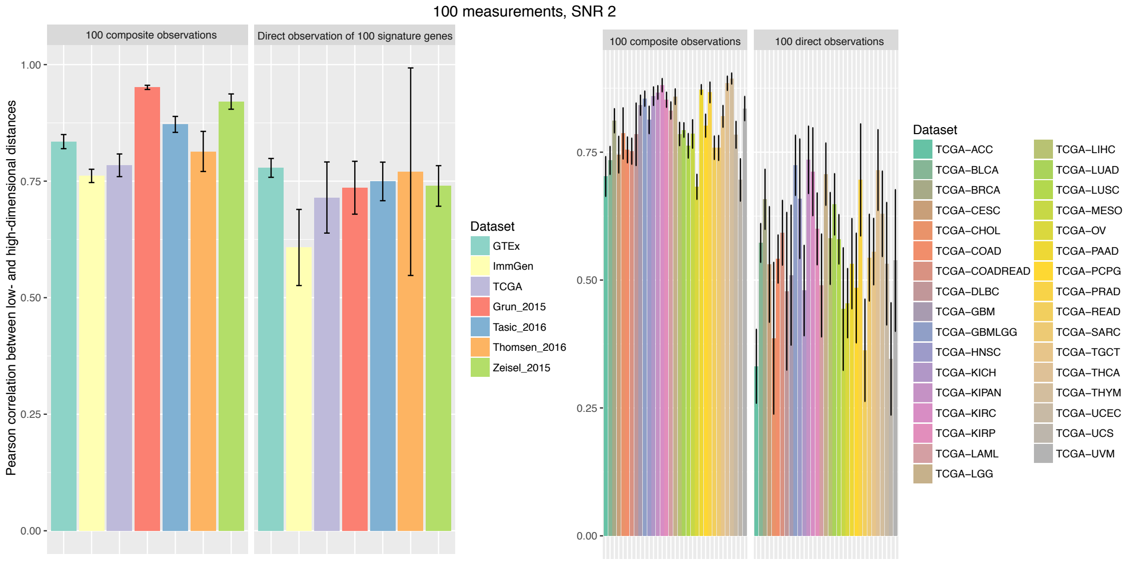

Figure S8: Calculating sample similarity by composite measurements *vs.* signature genes

For each dataset, 100 noisy composite observations and 100 signature measurements were simulated for each sample. Predicted pairwise distances were calculated based on these measurements, and compared with pairwise Euclidean distances calculated from the high-dimensional gene expression values. Shown are the Pearson correlation coefficients (y axis) between predicted and actual pairwise distances in each dataset. Error bars indicate one standard deviation across 50 random trials. Right: individual TCGA tumor types. Left: remaining datasets.

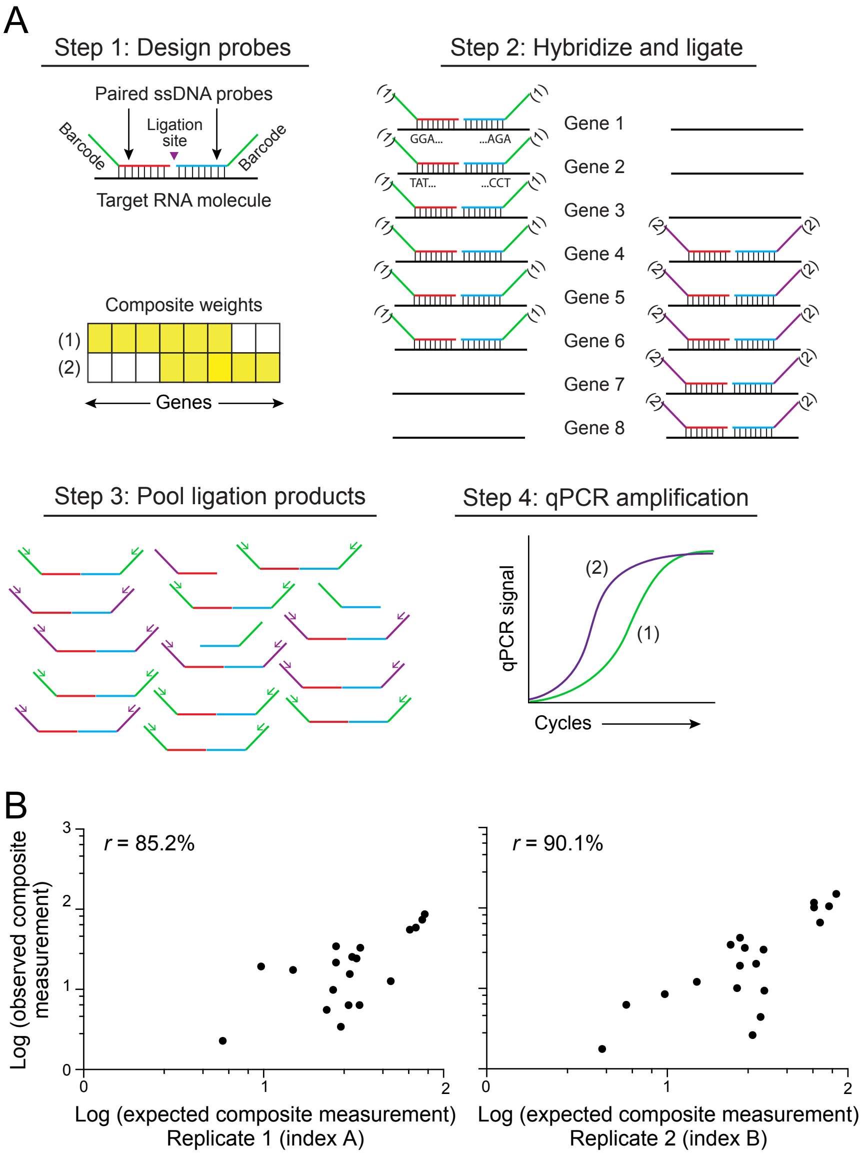

Figure S9: Composite measurements in practice

A. Overview of procedure. From top left: (Step 1) For each target gene, paired probes were designed to hybridize immediately adjacent to each other. Probes for all genes in a given measurement ((1) or (2)) share a common barcode. (Step 2) The probe library is hybridized to an RNA sample, and probe pairs are ligated using splintR ligase. (Step 3) Hybridized and ligated probe pairs are purified and then (Step 4) each measurement is performed by qPCR of the common barcodes (green and purple curves). **B**. Composite measurements of 23 genes. Each composite measurement was designed with random weights across 23 genes, and measurements performed according to the procedure in (A). Scatter plots show the expected composite measurement value (log-transformed) based on (computational) linear combination of the individual gene's qPCR values (x axis), compared to the observed composite measurements (log-transformed, y axis). The two panels indicate replicates performed with the same compositions, but using probe libraries with different measurement barcodes.