JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 8, 2015

Pp. 715–728
DOI: 10.1089/cmb.2015.0085

Research Articles

Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data

SEAN SIMMONS,[1,2] JIAN PENG,[1,2,3] JADWIGA BIENKOWSKA,[2,4] and BONNIE BERGER[1,2]

ABSTRACT

Biology is being inundated by noisy, high-dimensional data to an extent never before experienced. Dimensionality reduction techniques such as principal component analysis (PCA) are common approaches for dealing with this onslaught. Though these unsupervised techniques can help uncover interesting structure in high-dimensional data they give little insight into the biological and technical considerations that might explain the uncovered structure. Here we introduce a hybrid approach—component selection using mutual information (CSUMI)—that uses a mutual information—based statistic to reinterpret the results of PCA in a biologically meaningful way. We apply CSUMI to RNA-seq data from GTEx. Our hybrid approach enables us to unveil the previously hidden relationship between principal components (PCs) and the underlying biological and technical sources of variation across samples. In particular, we look at how tissue type affects PCs beyond the first two, allowing us to devise a principled way of choosing which PCs to consider when exploring the data. We further apply our method to RNA-seq data taken from the brain and show that some of the most biologically informative PCs are higher-dimensional PCs; for instance, PC 5 can differentiate the basal ganglia from other tissues. We also use CSUMI to explore how technical artifacts affect the global structure of the data, validating previous results and demonstrating how our method can be viewed as a verification framework for detecting undiscovered biases in emerging technologies. Finally we compare CSUMI to two correlation-based approaches, showing ours outperforms both. A python implementation is available online on the CSUMI website.

Key words: dimensionality reduction, mutual information, RNA-Seq.

1. INTRODUCTION

THE INFLUX OF NOISY, HIGH-DIMENSIONAL DATA into the biological sciences has made the ability to tease out lower-dimensional structure from that data crucial. Recent years have provided numerous examples of how such low-dimensional structure can give insight into the underlying biology, serving both as a tool for

[1]Department of Mathematics, [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts.
[3]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois.
[4]Oncology Research Unit, Pfizer, San Diego, California.

understanding and visualization (Hofree et al., 2013; Bendall et al., 2014; Schmid et al., 2012; Shen-Orr et al., 2011). One of the main techniques applied in these results is known as unsupervised dimensionality reduction. As the name suggests, this technique involves taking high-dimensional data and projecting it down onto a much lower-dimensional subspace. Techniques such as principal component analysis (PCA) and nonnegative matrix factorization (NMF, see Devarajan, 2008) attempt to perform this reduction so that very little information is lost. Though such approaches have proven valuable, they do have limitations. In particular, it is left up to the user to unravel the biological meaning of the results on their own. Moreover, these methods can lead to users ignoring important information hidden in higher dimensions. How can one decide which of the reduced dimensions are biologically relevant or which can be traced to artifacts or noise?

We aim to overcome these limitations by introducing a statistical approach that relates the results of dimension reduction to the biology we are interested in. This allows a practitioner to better explore large-scale, higher-dimensional data in order to further uncover global structure—structure that is apparent when looking at the expression levels of all or many of the genes at once, but not in the expression levels of a single gene.

We focus on applying our hybrid approach to PCA, though in theory it can be applied to most dimensionality reduction techniques. PCA is one of the most commonly used dimensionality reduction techniques in the biological sciences. It has famously been used to help understand population structure (Menozzi et al., 1978), a method that has become standard in the fields of statistical and population genetics (Price et al., 2006). More recently, PCA has been used to explore structure in gene expression array data (Schmid et al., 2012). It was shown that the first few PCs are closely related to the tissue of origin and that projecting down onto the first two PCs provides us with an informative way to visualize this extremely high-dimensional data. A similar method has been used to study cancerous tissue and stem cells in standard gene expression data, an application that showed cancer tissues seem to be an intermediary between regular tissue and stem cells, supporting the cancer stem cell hypothesis (Palmer et al., 2012). These examples are just the tip of the iceberg when it comes to applications of PCA to biomedical data! As currently implemented, however, this approach is limited. Since the PCs are ordered according to how much variance they explain—the first PC explains the most variance, the second PC explains the second most, and so on—researchers often choose the first few PCs, completely ignoring information that might be hidden in other PCs. Previous methods to decide how many PCs to look at apply random matrix theory to see if the PCs are more informative than they should be at random, but give no information about what covariates might be contributing to this structure (Johnstone, 2001). We argue that researchers should alternatively be looking at the PCs related to their problem of interest, not simply those most related to the overall variation in a dataset.

We introduce component selection using mutual information, CSUMI, a method to choose the most relevant principal components for any given feature. This approach allows researchers to better utilize these later PCs when trying to understand and visualize their data, helping them decide which PCs are worth keeping and which can be thrown away—for example, it can show which PCs will be the most informative visually and statistically to project onto. Though there exists supervised dimension reduction techniques (LDA, partial least squares, etc.), to our knowledge our method is the first to use information from covariates to understand and better prioritize the results of unsupervised dimensionality reduction methods.

1.1. Related work

Our method is inspired by mutual information-based approaches for feature selection (Peng et al., 2005, among many others). Unlike feature selection, CSUMI is meant to be used as a data exploration and verification tool. By looking at high-scoring covariates, we are able to understand what covariates affect the data and in what way. This knowledge can be useful in figuring out which technical covariates need to be corrected for before analysis, among other tasks. Alternatively we can, for a given covariate, use CSUMI to decide which PCs are the most informative. This information can be useful in many settings, particularly data visualization.

There are many preexisting dimensionality reduction methods for taking covariates into account before projection, most notably linear discriminant analysis (LDA). These methods, however, are not directly comparable to CSUMI. In contrast, our approach is meant to be a tool to help better understand how the results of unsupervised dimension reduction methods relate to biological covariates, not as a dimension reduction tool in its own right.

1.2. Our contributions

In addition to the method itself as a novel contribution, we apply CSUMI to RNA-seq data taken from the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013). We show that the most informative PCs are not always the first few, but instead that later PCs play an important role as well. Not only does this insight allow us to better visualize the global structure in RNA-seq data, for the first time to our knowledge, but it allows us to measure how much certain biological and technical variables actually affect the data, information that could help us understand the underlying biological and technical drivers in a given dataset. We also apply our statistic to RNA-seq data from brain tissue. Our statistic shows that the first few PCs are not the most informative about which region of the brain a sample is taken from, but rather that later ones carry a large percentage of the information of interest. For instance, when we look at brain tissue using our method we see that PC 5 can differentiate the basal ganglia from other tissues. This finding is in sharp contrast to PCs 2, 3, and 4, which are not nearly as informative about tissue type (see section 3.7.1). We also consider other biological covariates (Supplementary Fig. 5).

We are also able to use our method to study technical variation in the RNA-Seq data. Our results agree with previous studies of how technical variation effects RNA-Seq data (SEQC/MAQC-III Consortium, 2014). This validation demonstrates our method's ability to help discover signal due to technical variation that needs to be corrected for before analyses. This application is particularly useful as new technologies emerge that have undiscovered biases. Thus, CSUMI can further be viewed as a verification platform for emerging technologies.

2. METHODS

2.1. Mutual information and PCA: CSUMI

We introduce a statistic, $MI_{\mathbb{C}}(C, j)$, which measures the percentage of the information contained in covariate C that is also contained in the jth PC. We compute this measure using mutual information, a standard tool in information theory (Wang et al., 2009; Reshed et al., 2011). Though our work is not the first to apply mutual information to transcriptomic data (Daub et al., 2004; Martinez and Reyes-Valdes, 2008), to our knowledge we are the first method to use it as a means of investigating dimensionality reduction. Given two random variables, X_1 and X_2, the mutual information (informally) measures the amount of information shared between the two variables. It is defined as:

$$MI(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \tag{1}$$

where $H(X)$ is the entropy of a given random variable, X. For discrete random variables this equals

$$H(X) = -\sum_{x \in image(X)} P(X = x) \log_2 (P(X = x)) \tag{2}$$

where $P(X = x)$ is the probability that X equals x.

Though the technique we introduce below works with any dimensionality reduction technique (such as nonnegative matrix factorization, etc.), we describe it in the context of principal component analysis. PCA is a technique used for, among other things, dimensionality reduction and visualization of high-dimensional data. One can view PCA as finding the directions with the highest degree of variance for a given dataset. More formally, assume we have a dataset $D = \{d_1, \ldots, d_n\}$, where $d_i \in \mathbb{R}^m$. Let V be the matrix with d_1 as row one, d_2 as row two, etc. Let U be the matrix formed from V by first making all the columns sum to 0 (by subtracting off the mean of the column from each entry in that column), then normalizing each column to have variance 1 (by dividing through by the standard deviation of that column). Then $W = U^T U$ is the covariance matrix of the dataset. Let $v_1, \ldots, v_m$ be the eigenvectors of W, ordered from largest to smallest eigenvalue. Then v_1 is known as the first principal component, v_2 the second principal component, and so on.

We can project the data onto the jth PC, where the projection of the ith data point onto the jth PC is $a_i = u_i \cdot v_j$, where u_i is the ith row of U. These $(a_i)_{i=1,\ldots,n}$ embody the information about the data contained in the jth PC.

Consider a covariate $(c_i)_{i=1,\ldots,n}$, where each c_i is drawn from the finite set $S = \{s_1, \ldots, s_r\}$ (for example the covariate could be the tissue type of a given sample). We want to determine if the jth PC contains information about our covariate. Note that we can view c_i as being produced by some random variable, let us call it C. Similarly we can view the a_i as being produced by a random variable, let us call it A. We would like to be able to measure $MI(C, A)$, since this would tell us exactly what we want to know; however, we do

not know the exact distribution of C or A. Since C is a random variable over a finite set we can approximate its distribution with the sample distribution we get from the c_i.

A is continuous, and thus we cannot simply approximate it with the sample distribution. Instead we discretize A. There are many ways to do this (alternatively, if we wanted to avoid discretization altogether we could try to estimate A by fitting some continuous distribution to the data). We use the following approach: pick a number k. It will be the number of discrete pieces we will use to approximate our continuous variable. Order all the $a_i's$ from smallest to largest. Let $r_1, r_2, r_3, \cdots, r_{k+1}$ be chosen so that $[r_i, r_{i+1})$ contains $\frac{n}{k}$ of the a_i's (if k does not divide n then each interval will contain either $\lfloor \frac{n}{k} \rfloor$ or $\lceil \frac{n}{k} \rceil$ elements; we will always let $r_1 = -\infty$ and $r_{k+1} = \infty$). Define a random variable B so that $B = i$ if $A \in [r_i, r_{i+1})$. Since B is a random variable over a finite set, we can approximate its distribution with its sample distribution. This allows us to estimate $MI(C, B)$, which will serve as our stand in for $MI(C, A)$. Note that $MI(C, B)$ can be affected by the amount of information in C (that is to say $H(C)$); to correct for this fact, we rescale to get the statistic

$$MI_{\mathbb{C}}(C, j) = \frac{MI(C, B)}{H(C)} \tag{3}$$

This measure can be viewed as the percentage of the information contained in covariate C that is also contained in the jth PC. This is the tool we will use to measure how well a given covariate corresponds to a given PC. To apply this approach when dealing with another dimension reduction technique, replace the a_i's with the coordinates we get after dimension reduction.

Intuitively we can think of $MI_{\mathbb{C}}(C, j)$ as estimating the percentage of information contained in our covariate that is explained by the jth PC. One could use numerous other measures—$MI(C, B)$, Kullback-Leibler divergence, conditional entropy, etc. It turns out that for most reasonable measures, including those mentioned above, the p-value we calculate will be exactly the same (this follows since all of them can be calculated knowing only $H(C)$, $H(B)$, and $H(B, C)$, where $H(B)$ and $H(C)$ are kept constant during each step of the Monte-Carlo procedure used to estimate the p-value, see below). The only case that the choice is likely to have an effect is when we fix the PC and are looking at multiple covariates.

If we are looking at multiple covariates and want to find which one best explains the PC, we can rank all of the covariates according to the p-value calculated from $MI_{\mathbb{C}}(C, j)$ (see section 2.2). The higher up on the list a covariate appears the more likely it is to give us information about the jth PC. Similarly, for a fixed covariate, we can score each PC to see how closely that PC relates to our covariate.

2.2. *Calculating* p-*values*

The above gives us a means of ranking how closely a given covariate relates to a given PC. We measure the significance of this relationship using a Monte Carlo–type approach. For each step in the Monte Carlo process, randomly permute the c_i's to get a new labeling $c_1', \cdots, c_n'$ of the data points. Let C' be the random variable that maps from each data point to the corresponding c_i'. We can then calculate $MI_{\mathbb{C}}(C', j)$. Repeat this process R times, and let r be the number of trials for which $MI_{\mathbb{C}}(C', j) \geq MI_{\mathbb{C}}(C, j)$. We can then approximate the p-value as $\frac{r+1}{R+1}$ (North et al., 2002); Ewens, 2003).

2.3. *Setting parameters*

To use our statistic we need to determine what we want k to be, where k is the number of intervals we use in our discretization step. Different values of k should work better for different cases, but in general small changes in k do not seem to change the results noticeably (Supplementary Table S2, available online at www.liebertpub.com/cmb). We decided to use a moderate value of k, namely $k = 6$, largely based off the fact that it is about equal to the number of values taken on by each of our covariates (which range from 2 to 13 possible values).

Besides setting k we also have to specify the number of trials we want to use when estimating our p-values. Using more trials will give us a better estimate of our p-value but will also take longer. In all our experiments we use 999 trials to estimate our p-values.

2.4. *Measuring clustering*

We want to measure how well samples with the same tissue type cluster together after dimension reduction. To do this we use a nearest neighbor–type approach—we estimate how much loss we expect by guessing a sample's tissue type based off its nearest neighbors after dimensionality reduction.

Consider a dataset $d_1, \cdots, d_n$ with corresponding tissue types $c_1, \cdots, c_n$. After performing dimensionality reduction we get a new data set $b_1, \cdots, b_n$. For $i = 1, \ldots, n$ we will generate a new label c_i' as follows: for each data point b_i we can find its three closest neighbors, b_{j1}, b_{j2}, b_{j3}. Then let c_i' be whichever label occurs the most in c_{j1}, c_{j2}, c_{j3} (if there is a tie then choose one randomly). We can then measure

$$Clust((b_i)_{i=l,\ldots,n}) = \frac{|\{i : c_i \neq c_i'\}|}{n} \tag{4}$$

The smaller this number, the better, as more of the tissue types tend to cluster together. Note that one could instead use more than three nearest neighbors, but our experimentation suggests changing this value does not affect our results qualitatively.

As another method to measure how well clustering is preserved we use k-means clustering. More specifically, we apply CSUMI with respect to the dataset $d_1, \ldots, d_n$ with covariates $c_1, \ldots, c_n$ as above, keeping only the p highest scoring PCs (where p is a parameter we can change). We then apply k-means clustering to this lower dimension set, where we take the number of clusters to be equal to the number of values c_i takes on (for example, if we were dealing with sex then we would have two clusters, corresponding to male and female; on the other hand, if we used tissue type as our covariate we would have seven clusters.). In order to see how well clusters are preserved, we measure the purity of the clusters by computing the mutual information between our covariate and the clusters we get using k-means (Liao et al., 2009).

3. RESULTS

3.1. Overview

Our approach, CSUMI, works by first performing PCA on RNA-seq data. For a covariate of interest we use our statistic, $MI_{\mathbb{C}}$, to compare that covariate to each PC in turn and explore the highest scoring PCs.

We demonstrate that our approach, CSUMI, informs practitioners which PCs give the most information about their biological question of interest. We begin by validating our approach on RNA-seq data from the seven most common tissues (Datasets) and show that the first six PCs are the most informative about tissue type. This finding tells us that, if we want to use PCs to better understand tissue type (for example if we want to perform clustering or something along those lines), then we should consider the first six PCs—without our measure we might think it suffices to consider only the first two or three PCs.

Next we consider the samples drawn from the brain (Datasets). Surprisingly, our method shows that it makes more sense to look at later PCs as opposed to looking at the first few. This finding can be very informative; in particular, if one is trying to visualize the transcriptome of the brain, then instead of projecting onto the first two PCs, practitioners can project onto later PCs. We also use our statistic to investigate how other covariates, both biological and technical, affect the global transcriptional landscape.

3.2. Datasets

Our RNA-seq data was obtained from the Genotype-Tissue Expression (GTEx) project in November 2013. This data is freely available online. For each sample we get a vector for which the ith entry equals the expression of the ith gene, where expression is given as number of reads per kilobase per million reads (RPKM). Note that, instead of looking at each gene we could have looked at the RPKM for each transcript—our experiments have shown that the overall picture we get by applying our method at the gene level is very similar, though not identical, to that at the transcript level.

For our first set of analyses we discarded tissue types represented by few samples, leaving only the seven most common tissue types: thyroid, whole blood, artery (tibial), lung, sun-exposed skin, muscle (skeletal), and heart (left ventricle). Note that samples from the artery are referred to as blood vessels in all our figures (this is due to the naming scheme of GTEx, which has two different labelings for each tissue type, one more specific). This gives us a total of 809 samples. For more details about this data see the GTEx website or reference (Lonsdale et al., 2013).

For our second set of analyses we looked at all the tissue samples from different brain regions. Focusing on the second PC, there is one clear outlier (Supplementary Fig. 1). This sample is labeled as being from a female brain, yet behaves much more like the testis (including when we look at Y chromosome expression, data not shown). It seems likely that this is a technical artifact (perhaps a mislabeled sample?), so we

exclude this sample, leaving us with 312 samples. The brain tissues are: caudate (basal ganglia), amygdala, cerebellum, hippocampus, frontal cortex, cortex, putamen (basal ganglia), hypothalamus, nucleus accumbens (basal ganglia), anterior cingulate cortex, cerebellar hemisphere, spinal cord, and substantia nigra.

GTEx also provides information about a number of covariates for each sample. There are five main biological covariates: age (broken down into 10-year increments), sex, Hardy score, and two different versions of tissue type (one more specific than the other—unless otherwise noted, when we refer to tissue type we mean the more specific one). The Hardy score is an integer between 0 and 4 that gives information about how a patient died—zero corresponds to a patient who died on a ventilator, while one through four correspond to the continuum from fast, violent deaths up through slower deaths. There are also many technical covariates, but we limit ourselves to those with at most 13 possible values (since our statistic is designed to work when the number of values is much smaller than the number of samples). This leaves us with four technical covariates. These are the site that enrolled the donor (coded with its BSS code), the time between death and final tissue stabilization, severity of autolysis, and technology used to isolate the DNA/RNA.

Depending on the situation it is common to transform RNA-Seq data to a log or square root scale. Though our method could be just as easily applied to transformed data, for simplicities sake we decided to focus in the main article on the raw RPKM data, as opposed to the log and/or square root transformed version. Some analysis of log-transformed data is provided in the Supplementary Material. (Supplementary Table 1 and Supplementary Fig. 10.)

3.3. *Tissue type and the first few PCs*

We first investigate whether we can reproduce with RNA-seq data earlier results on gene expression data from GEO (Schmid et al., 2012); that is, when we project the data onto the first two PCs (Fig. 1a). As expected, the data segregates by tissue type. As an initial sanity check we also want to see if our method correctly identifies the relationship between tissue type and the first two PCs. In order to accomplish this goal we look at a number of different covariates. We rank these covariates using our $MI_{\mathbb{C}}$ score with respect to each of the first two PCs (Tables 1 and 2). We see that the tissue of origin gives the largest $MI_{\mathbb{C}}$ score in both cases (by a large margin), and has an estimated p-value less than 0.001 (note that, since we are using a Monte Carlo–type approach, we cannot better estimate the p-value without running an extremely large number of trials). It is worth noting that other covariates also achieve p-values under 0.001, something we will touch upon later.

Note that there are some samples that do not cluster with their tissue types, which could be for many reasons—random noise, underlying medical conditions of the patients, or possible problems with the data point (such as mislabeling, contamination, etc). Indeed, this work may better classify such samples.

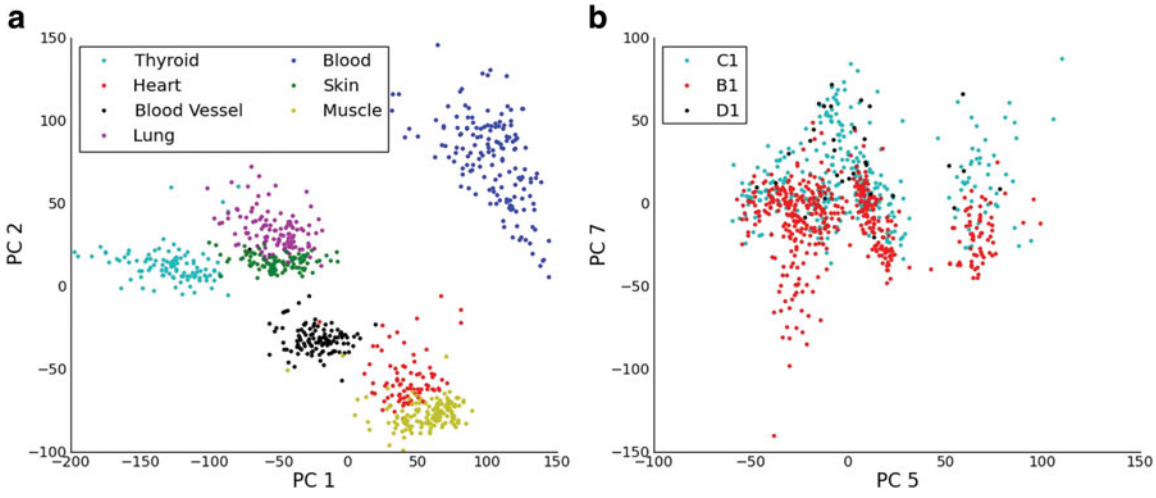

FIG. 1. **(a)** A plot of the RNA-seq data projected onto the first two PCs, where each spot corresponds to a sample and each color to a tissue type of origin. We see that, as expected, the samples from the same tissue cluster together. **(b)** A plot of the RNA-seq data projected onto the fifth and seventh PCs, now colored by enrollment center (i.e., BSS center codes C1, B1, and D1). We see that there is an obvious relation between the seventh PC and enrollment center. PCs, principal components.

Table 1. For Each Covariate We Calculate the $MI_{\mathbb{C}}$ Score for the First PC and the Corresponding *p*-Value

Covariate (C)	$MI_{\mathbb{C}}(C, 1)$	p-*Value*
Tissue type	**0.5674**	**≤0.001**
Nucleotide isolation technique	0.3574	≤0.001
Autolysis severity	0.3519	≤0.001
Tissue stabilization time	0.1181	≤0.001
Enrollment center	0.0039	0.899

Note that tissue type of origin (in bold) is the highest scoring among the covariates by far, just as we expected.

3.4. Tissue and later PCs

Clearly the tissue of origin is a large contributor to the variance in the first few PCs. How much information about tissue of origin is hidden in later PCs? To answer this question we apply our $MI_{\mathbb{C}}$ statistic to look at how much information about each of the first 100 PCs is embodied by tissue type (see Table 3 for the first nine PCs). The results show that tissue type has a large effect on even fairly distant PCs. For the first six PCs the $MI_{\mathbb{C}}$ score is quite high (above 0.3). The score then drops sharply, and from the 8th value to the 100th it fluctuates between 0.02 and 0.08 (which, even though it is quite a bit lower than for the first couple PCs, still has an estimated *p*-value of ≤0.001). The one exception to this is the 73rd PC, where the score is 0.014, which corresponds to an estimated *p*-value of about 0.033. The sharp divide that occurs between the sixth and eighth PC can be nicely seen by projecting the data onto the eighth and ninth PC (Supp. Fig. S2). Though our statistic tells us that there is a significant relationship between tissue type and PCs even this far out, we see that it lacks the obvious structure of earlier PCs, structure we see even when comparing the fifth and sixth PCs (Supp. Fig. S3). This sharp drop in the CSUMI score suggests that in many cases most of the most interesting information about tissue is contained in the first six dimensions (though clearly even higher dimensions contain some information as well).

The above behavior could have many different interpretations. One is that in some sense the ''dimension'' of the space of tissues is roughly six dimensional, and the affect on PCs beyond the sixth are from noise and nonlinearities related to this structure. This interpretation is especially enticing since it suggests that, if we are interested in the structure arising in the data due to the tissue type of our samples, then it may suffice to look at only the top six or seven PCs instead of looking at hundreds of them. This gives us a principled way of determining how many PCs we want to consider when exploring the data.

It is worth pointing out that each of these first six PCs gives us different types of information about tissue type (Fig. 1a). Take for example the first two PCs. We see that the first PC does not differentiate between muscle and heart or between lung and skin, whereas the second PC does exactly that. By the same token PCs three through six also give different types of information about tissue type than the first two PCs (compare Fig. 1a to Supp. Fig. S3).

3.5. Tissue clustering

We have implied that projecting onto the first six eigenvalues should preserve most of the information about tissue clusters in the data. In order to show this more formally, we projected the data down onto the

Table 2. For Each Covariate We Calculate the $MI_{\mathbb{C}}$ Score for the Second PC and the Corresponding *p*-Value

Covariate (C)	$MI_{\mathbb{C}}(C, 2)$	p-*Value*
Tissue type	**0.5771**	**≤0.001**
Autolysis severity	0.3312	≤0.001
Nucleotide isolation technique	0.3057	≤0.001
Tissue stabilization time	0.1038	≤0.001
Enrollment center	0.0069	0.53

Note that tissue of origin (in bold) is again the highest scoring by far, just as we expected.

Table 3. For Each of the First Nine PCs, the $MI_{\mathbb{C}}$ Score Is Calculated Between the Tissue Type and that PC, As Well As the *p*-Value

PC	$MI_{\mathbb{C}}$ *score*	p-*Value*
1	0.5674	≤ 0.001
2	0.5771	≤ 0.001
3	0.4189	≤ 0.001
4	0.5108	≤ 0.001
5	0.5132	≤ 0.001
6	0.3036	≤ 0.001
7	0.1199	≤ 0.001
8	0.0396	≤ 0.001
9	0.0818	≤ 0.001

More formally, if T is the random variable that maps to tissue type, then this table looks at $MI_{\mathbb{C}}(T, i)$ for PCs $i = 1, \ldots, 9$. Note that all these relationships are very statistically significant, though there is a sharp drop in score between the sixth and eighth PCs.

first p components, where $p = 1, \ldots, n$, giving us a set of p-dimensional data points $b_1, \ldots, b_n$. We then measured how well these b_i preserve tissue clusters. In particular, we calculated $Clust((b_i)_{i=1,\cdots,n})$ (Fig. 2).

We see that, just as our method predicted, the clusters do not seem to improve much past $p = 6$, alternating between a *Clust* score of 0.0013 and 0.0017. Moreover, we see that using the first six PCs gives us a *Clust* score of 0.0016, which slightly improves on the *Clust* score of 0.0019 achieved by using all the PCs.

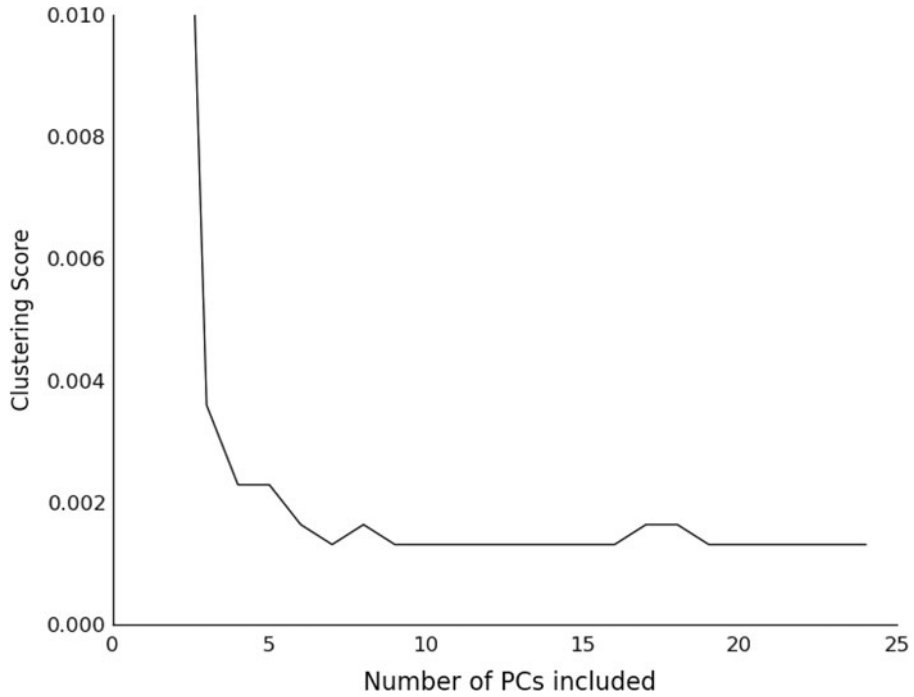

FIG. 2. Measure of how well tissues cluster together when projecting onto different numbers of principal components. The *x*-axis tells us how many PCs we are projecting down onto (note that we are projecting down onto the first few PCs as opposed to projecting down onto the highest scoring PCs), whereas the *y*-axis is our measure of clustering, *Clust*. Recall that a lower *Clust* score means that each tissue type tends to cluster together. We see that at first the *Clust* score decreases as we add more PCs, but after we add roughly six PCs there is no further improvement. This is consistent with our method, which implies that the first six PCs are much more informative about tissue type than later PCs are.

3.6. The effect of technical covariates

Though tissue type seems to be the main driver of variance in the first few PCs, our method also finds covariates that seem to influence later PCs. Moreover, many of these high scoring covariates are technical covariates—that is to say that they are not related to the underlying biology, but are instead related to the way data was collected. This knowledge is important to have before any analyses using this dataset, since otherwise it might be hard to tell what variance is due to biology and what variance is due to technical artifacts.

Consider, for example, the seventh PC (Fig. 1b; Table 4). Using our $MI_{\mathbb{C}}$ score we see that the highest ranked covariate is enrollment center—which is the BSS code of the collection site—with a score of about 0.17. We can also see enrollment center scores fairly highly when compared to the fifth PC (a score of 0.048, which has an estimated p-value less than 0.001). If we project the data onto the 7th and 5th eigenvalues and color by enrollment center, we see a definite effect (Fig. 1b). One might wonder if this is just due to confounding based on the relationship between tissue type and enrollment center. However, this is not the case since the same relationship is present when we plot only one type of tissue, namely the whole blood samples (Supp. Fig. S4, where we choose whole blood because it is the tissue we have the most samples from).

Luckily numerous techniques exist to help deal with technical artifacts (Baym et al., 2008); Parker et al., 2014). Our approach allows us to understand which technical artifacts are having a global effect on expression, something that could be useful when deciding which technical covariates need to be corrected for. It may also be possible to use this knowledge to better correct for such problems (perhaps, for example, by projecting out the seventh PC or including it as another covariate).

Moreover, even though the relationship with enrollment center seems to be a technical artifact, there may in fact be a biological reason for it—perhaps there is some biological difference between the patients in each of the centers that the authors are unaware of. Either way, we see that our technique allowed us to find a source of global variation that we need to correct for in the future, one that might have been hard to find without a statistic of the kind introduced here.

Note that the results in this section are not surprising since similar results have been found by the MAQC consortium, among others (SEQC/MAQC-III Consortium, 2014). This demonstrates that CSUMI is able to help us uncover the forces that are driving the variance in the dataset. CSUMI gives users a method for understanding which covariates need to be corrected for in a given dataset, something that could be especially important as new technologies with new biases are introduced.

3.7. What selected PCs tell us about the brain

3.7.1. Tissue type. Above we have looked at the seven tissues with the largest number of samples in the dataset. It is just as easy for us to apply our hybrid approach to some subset of tissues. In particular, we examine tissues that are drawn from the brain. We choose the brain because it is the most complex of the organs represented in GTEx, with many of its samples drawn from different regions of the brain.

Our method suggests that we should use the first and fifth PCs to visualize how tissue type relates to RNA-seq expression levels in the brain, a fact that would have been missed by the naive approach. The naive approach for using PCA is to project our brain samples onto the first two PCs. When we apply our method we see that the first PC is closely related to the brain region—it actually divides the cerebral tissues

Table 4. Each Covariate Shows the $MI_{\mathbb{C}}$ Score Relative to the Seventh PC and the Corresponding p-Value

Covariate (C)	$MI_{\mathbb{C}}(C, 7)$	p-*Value*
Enrollment center	**0.1750**	**≤ 0.001**
Tissue type	0.1199	≤ 0.001
Tissue stabilization time	0.1053	≤ 0.001
Autolysis severity	0.1003	≤ 0.001
Nucleotide isolation technique	0.0974	≤ 0.001

We see that enrollment center scores the highest, suggesting that the variance embodied in the seventh PC might largely be due to a technical artifact.

Table 5. Limiting Our Attention to the Brain, for Each of the First 10 PCs Calculate the $MI_{\mathbb{C}}$ Score Between the Tissue Type and that PC, As Well As the *p*-Value

PC	$MI_{\mathbb{C}}$ *score*	p-*Value*
1	**0.2716**	**≤ 0.001**
2	0.1666	≤ 0.001
3	0.0894	≤ 0.001
4	0.0461	0.239
5	**0.2415**	**≤ 0.001**
6	0.1431	≤ 0.001
7	0.0676	0.002
8	0.0542	0.046
9	0.0946	≤ 0.001
10	0.0719	≤ 0.001

More formally, if T is the tissue type, then this table looks at $MI_{\mathbb{C}}(T, i)$ for $I = 1, \ldots, 10$. Surprisingly, the two highest scoring PCs are the first and the fifth in bold, not the first and the second. It is also worth noting that the fourth PC gives extremely little information about tissue type, especially in comparison to the fifth.

(tissues labeled as cerebellum or cerebellar hemisphere) from the rest of the brain samples. Looking at the other PCs, however, we see that the fifth PC is much more strongly related to the brain region than the second PC (Table 5). In particular, we see that the fifth PC divides the tissues of the basal ganglia from the rest of the tissues of the brain (Fig. 3). When projecting onto the first and second PCs, on the other hand, the basal ganglia is not divided from the other tissues (Fig. 3).

Table 5 unveils another interesting observation, namely that the fourth PC seems to give almost no information about tissue type. All of the other top six PCs give us a p-value less than 0.001 when compared to tissue type, but the fourth PC does not. It seems, in fact, that the fourth PC is not related to any of the covariates we look at, and is instead related to another one. Using simple Pearson correlation we see that it is in fact related to the estimated fragment library size of a sample ($p < 10^{-12}$), something that is mildly surprising since the data has been normalized for a number of reads (this could be due to many things, including the effect of genes with low coverage). This shows one limitation built into our method, namely that it only works on categorical covariates. Understanding other types of covariates requires other methods.

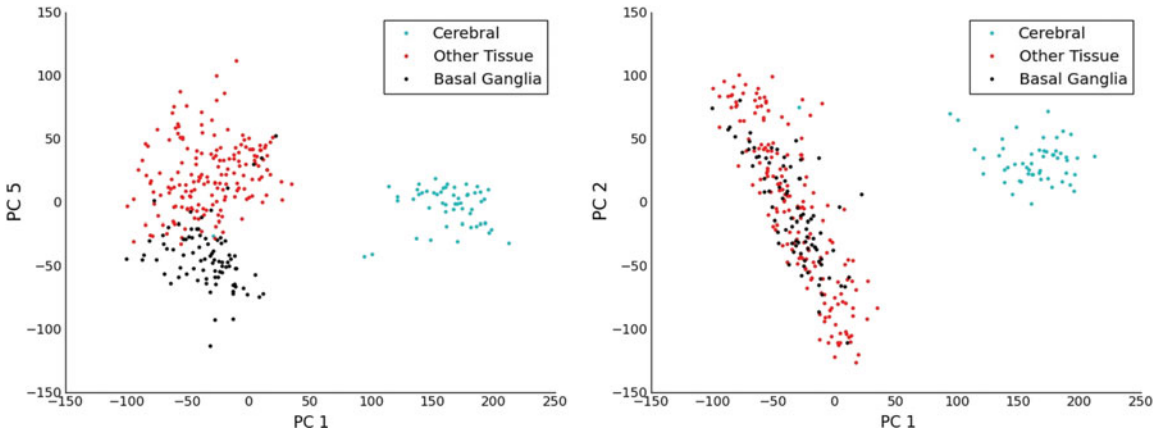

FIG. 3. Plot of the result we get by projecting the data from the brain onto the first and fifth PCs versus projecting onto the first and second PCs. We see that the first PC divides cerebral tissue (samples labeled with either brain-cerebellar hemisphere or brain-cerebellum) from the rest of the brain, and the fifth PC divides the basal ganglia from the rest of the brain. No such division is seen using the second PC.

The above discussion leaves a few unanswered questions: Why are the cerebral tissues so different from the rest of the tissues in the brain? Why is the fifth PC so much more informative than the second (and even more informative than the third and fourth)? We do not attempt to answer these questions here, and instead leave them as open questions for researchers with more of a biological background.

3.7.2. Tissue clustering. We next investigate how our method affects clustering by tissue type in the brain. Naively we might expect projecting down onto the first four PCs to better preserve our tissue clusters than projecting down onto the four-dimensional space spanned by the first, second, third, and fifth PCs. Our method, on the other hand, would predict the opposite, since the fifth PC scores higher than the fourth PC (Table 5). We use our clustering measure to investigate this discrepancy. If $b_1, \ldots, b_n$ is the data set we get by projecting down onto the first four components and $b'_1, \ldots, b'_n$ is the data set we get by projecting onto the four-dimensional space spanned by the first, second, third, and fifth components, then we see $Clust((b_i)_{i=1,\ldots,n}) = 0.0713$ while $Clust((b'_i)_{i=1,\ldots,n}) = 0.0608$. Since lower scores mean clusters are better preserved, we see it is better not to use the first four components, just as our statistic suggested.

3.7.3. K-means clustering. We apply our *k*-means clustering measure (Methods) to the brain samples using different covariates. First we consider tissue type (Fig. 4). For a varying threshold p (aka the number of dimensions we project down to), we calculated the MI between tissue type and the *k*-means clustering. We see that for most values of *p* our method achieves higher scores than the naive approach. When we perform the same procedure using age (Supp. Fig. 6) and Hardy score (Supp. Fig. 7) as covariates, we also see that our method outperforms the naive approach.

3.7.4. Comparing to correlation. How does our method compare to correlation-based approaches? In particular we compared CSUMI to approaches using Pearson and Spearman correlation. To use correlation-based approaches we need real values (or ordered values) for our covariates. Thus first we randomly label different tissue types with integers 1 through 13. We then compare how well our approach

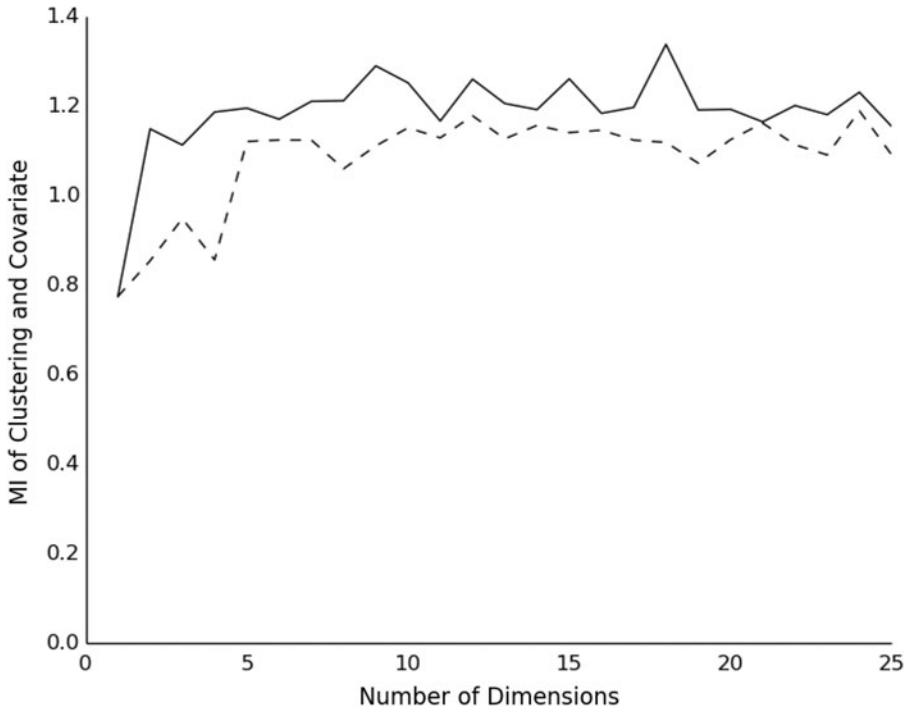

FIG. 4. Clustering by CSUMI (solid line) is much more informative than the one using the naive approach (dashed line). For a varying number of dimensions, *p*, we take the data and project it onto a *p*-dimensional space using either CSUMI or by picking the *p* largest PCs, respectively. We then perform *k*-means clustering, compute the mutual information between the tissue type and the clustering, and plot the result. CSUMI, component selection using mutual information.

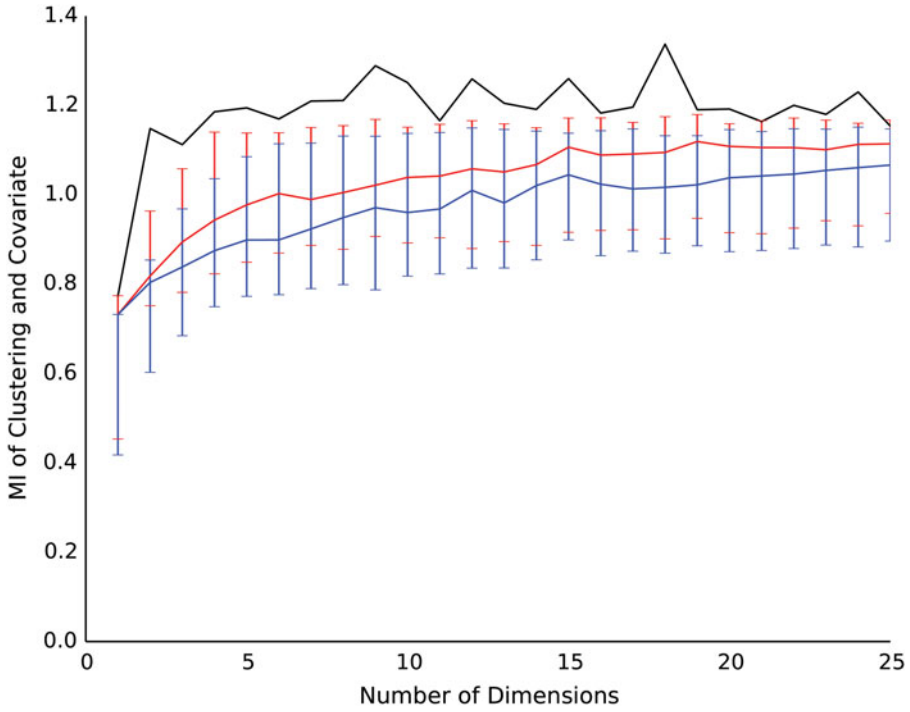

FIG. 5. Clustering by CSUMI (black line) is more informative than the one using the Spearman (red) or Pearson (blue) correlation. For a varying number of dimensions, p, we take the data and project it onto a p-dimensional space using either CSUMI or onto the p-dimensional space spanned by the principal component analyses with highest Spearman or Pearson correlation, respectively (the score for each correlation is based off the median taken over 100 random orderings of the tissue types, where the error bars range of the 25th to 75th percentiles). We then perform k-means clustering, compute the mutual information between the tissue type and the clustering, and plot the result. We see that CSUMI greatly outperforms either correlation-based approach.

preserves clusters compared to correlation. Second, we calculate the p-value derived from Pearson or Spearman correlation and project down onto the PCs with the lowest p-values. Since this method involves a random ordering of the tissues we repeat this procedure 100 times and record median scores. In Figure 5, we show that CSUMI outperforms both types of correlation. In fact, by comparing Figures 4 and 5 we see that simply ranking PCs by eigenvalue does better. Even if we look at the 75th or 95th percentile among the correlation scores instead of the median (Supp. Figs. 8 and 9) we see that CSUMI does at least as well as either correlation measure.

4. DISCUSSION

We introduce a hybrid method to help us better understand the PCs derived from RNA-seq data. This mutual information-based statistic gives us clues about what kind of biological information is unveiled by PCA. We also see that this statistic can reveal important information about the effect of technical variation on a data set—information that might help investigators decide how reliable biological conclusions are likely to be.

It is our hope that the above approach will help put the use of PCA for exploring biological on firmer ground. We provide a tool that allows us to uncover potentially important sources of variance hidden in later eigenvalues and allows us to measure how significant these findings are.

As with all other tools that allow users to explore a data set, ours comes with the caveat that users must be careful not to let the results of the analysis negatively bias later statistical analysis—in any application one must be careful to use multiple test correction or other approaches to avoid false findings.

It is also worth considering the relation between our statistical approach and clustering, something that we touched on here. Previous work (Yeung and Ruzzo, 2001) has argued against the common practice of

using PCA for dimension reduction before clustering biological data since only looking at the first few PCs can lead us to throwing away a lot of information about clusters in the data. Our results suggest an approach using our technique might avoid this drawback, allowing better clustering of high-dimensional biological data.

Our results show that, despite conventional wisdom, researchers can get better results out of PCA by using later PCs instead of using the first few PCs. Our technique provides a means of selecting which of these PCs are most useful for a given task. We believe our approach will be useful to the many biological researchers struggling to find structure in their data.

ACKNOWLEDGMENTS

We would like to thank Zak Kohane, Nathan Palmer, and Patrick Schmidt for helpful comments and inspiration. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplement to the University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941), the University of Chicago (MH090951 & MH090937), the University of North Carolina-Chapel Hill (MH090936), and to Harvard University (MH090948). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant no. 1122374 and NIH grant GM081871.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no conflicts of interest exist.

REFERENCES

Baym, M., Bakal, C., Perrimon, N., and Berger, B. 2008. High-resolution modeling of cellular signaling networks. *RECOMB*. 4955, 257–271.

Bendall, S., et al. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*. 157, 714–725.

Daub, C., Steuer, R., Selbig, J., and Kloskam, S. 2004. Estimating mutual information using B-spline functions an improved similarity measure for analysing gene expression data. *BMC Bioinform*. 5, 118.

Devarajan, K. 2008. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol*. 4, e1000029.

Ewens, W. 2003. On estimating *p*-values by Monte Carlo methods. *Am. J. Hum. Genet.* 71, 496–498.

Hofree, M., Shen, J., Carter, H., et al. 2013. Network-based stratification of tumor mutations. *Nat. Methods.* 10, 1108–1115.

Johnstone, I. 2001. On the distribution of the largest eigenvalues in principal component analysis. *Ann. Stat.* 29, 295–327.

Liao, C., et al. 2009. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics.* 25, 253–258.

Lonsdale, J., et al. 2013. The genotype-tissue expression GTEx project. *Nat. Genet.* 45, 580–585.

Martinez, O., and Reyes-Valdes, M. 2008. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *PNAS*. 105, 9709–9714.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. 1978. Synthetic maps of human gene frequency in Europeans. *Science*. 201, 786–792.

North, B., et al. 2002. A note on the calculation of emperical *p*-values from Monte Carlo procedure. *Am. J. Hum. Genet.* 71, 439–441.

Palmer, N., et al. 2012. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. *Genome Biol.* 13, R71.

Parker, N., et al. 2014. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics.* 30, 2757–2763.

Peng, H., Long, F., and Ding, C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238.

Price, A., et al. 2006. Principal components analysis for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909.

Reshef, D., et al. 2011. Detecting novel associations in large data sets. *Science*. 334, 1518–1524.

Schmid, P., et al. 2012. Making sense out of massive data by going beyond differential expression. *PNAS*. 109, 5549–5906.

SEQC/MAQC-III Consortium. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32, 903–914.

Shen-Orr, S., et al. 2011. Cell type-specific gene expression differences in complex tissues. *Nat. Methods.* 7, 287–289.

Yeung, K., and Ruzzo, W. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics*. 17, 763–774.

Wang, K., et al. 2009. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. Biotechnol.* 27, 829–837.

Address correspondence to:
Dr. Bonnie Berger
Department of Mathematics
Massachusetts Institute of Technology
77 Mass Avenue, 2-373
Cambridge, MA 02139

E-mail: bab@mit.edu