Harvard University
Harvard University Biostatistics Working Paper Series

Year 2011 *Paper* 126

Statistical Properties of the Integrative Correlation Coefficient: a Measure of Cross-study Gene Reproducibility

Leslie Cope* Giovanni Parmigiani†

*Johns Hopkins University, lcope1@hmi.edu

†Dana Farber Cancer Institute and Harvard School of Public Health, gp@jimmy.harvard.edu

Statistical Properties of the Integrative Correlation Coefficient: a Measure of Cross-study Gene Reproducibility.*

Leslie Cope† Giovanni Parmigiani ‡

March 17, 2011

Abstract

The integrative correlation coefficient was developed to facilitate the validation of expression microarray results in public datasets, by identifying genes that seem to be reproducibly measured across studies and even across microarray platforms. In the current study, we describe a number of interesting and important mathematical and statistical properties of the integrative correlation coefficient, including a unique null distribution with the unusual property that the variance does not shrink as the sample size increases, discussing how these findings impact its use and interpretation, and what they have to say about any method for identifying reproducible genes in a meta-analysis.

1 Introduction

The integrative correlation coefficient (ICC) was developed by Parmigiani, Garret-Mayer, Anbazhagan and Gabrielson [24] to facilitate the validation of expression microarray results in public datasets, by identifying genes that seem to be reproducibly measured across studies and even across microarray platforms. It was discovered independently by JK Lee and colleagues [17] who used it to measure the cross-study reproducibility in microarrays, but did not apply it at the level of individual genes, and is very similar to the approach taken by HJ Lee et al. [16]

It is a bit difficult to crisply define *reproducibilty* in the multi-study context, but the general idea is that if we put the same samples on two different microarray platforms, say, we are interested in those genes whose expression values are well-correlated across platform. It being impossible to directly assess correlation when two independent sample sets are compared, the integrative correlation solution was to map out the relationships between genes, within each study, and select as reproducible those genes for which the local gene-space topography is the same in both studies. Thus the very simple algorithm for the integrative correlation of gene x is as follows:

1. within each study, calculate the correlation between genes x and y for every $y \neq x$
2. calculate the cross-study correlation of correlations over all y; a high value indicates that inter-gene relationships are the same within study
3. compare to a null distribution to select reproducible genes

The method does not use any phenotypic information, and so does not compromise any higher level analyses that might be done. And using correlations as the measure of similarity minimizes any study-specific differences in probe effect or scale.

*Work supported by NSF Grant: NSF034211

†The Sidney Kimmel Comprehensive Cancer Center at The Johns Hopkins University

‡Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute and The Department of Biostatistics, Harvard School of Public Health

	non-reproducible genes	reproducible genes
mis-annotated genes	82.2%	17.8%
double-probed genes	48.6%	51.4%
single-probed genes	17.9%	82.1%

Table 1: Each row, corresponding to one of 3 levels of reproducibility built into the simulation scheme, shows the proportion of genes that are deemed reproducible in a comparison to the null distribution.

Since its introduction, the method has been applied in a variety of situations to compare platforms[17], evaluate methodology[26], aid in meta-analysis of gene expression sets[11, 24, 32, 30], and even for comparisons across species[31]. It is also a key component of the recently developed Co-expression extrapolation (COXEN)[18, 19, 23, 27, 28] approach to the development of biomarkers for drug response starting with in-vitro response data in cell lines.

In the current study, we describe a number of interesting and important mathematical and statistical properties of the integrative correlation coefficient, including a unique null distribution with the unusual property that the variance does not shrink as the sample size increases, discussing how these findings impact its use and interpretation, and what they have to say about any method for identifing reproducible genes in a meta-analysis.

Several other, practical issues are not explicitly dealt with here, such as how to handle multiple probes per gene and how strictly the studies should be cross-annotated (Is it sufficient to map at the gene level or should specific transcripts be matched to avoid problems with alternative splice forms? Does an algorithmic annotation system like Unigene offer sufficiently reliable annotation or should the analysis be restricted to a more highly curated set of genes such as those included in Refseq? When should integrative correlation itself be used to match genes across studies, in the presence of multiple probes for each gene?). For discussions of these issues we refer the interested reader to these publications discussion methods, software and applications [17, 24, 5, 7, 32]. In the current study, it is enough to assume that we have answered these questions one way or another, and have a complete cross-study annotation.

2 Illustrative Example

To demonstrate the method, we present a semi-simulated example. To create a pair of studies with a variety of ICCs we started with a single, large breast cancer study [29] hybridized to the Affymetrix hgu133a expression array, splitting it into two independent sets of 100 samples each, measured on 3000 selected genes. The distribution of those genes is the key to the simulation. One third of the genes used for the simulation have a single probe on the array, which is used to represent the gene in both simulated studies, for a high level of reproducibility. One half of the genes are represented by 2 different probes, with the two studies simulated to use different probes. The remaining one sixth of the genes are simulated as annotation errors, by selecting probes for different genes in each simulated study. The null and observed distributions of integrative correlations are shown in Figure 1 and Table 1. To define *reproducible* genes in Table 1, we use the 99th percentile of null integrative correlations as a cutoff.

It is encouraging that 82% of the genes for which the same probe is used in both studies are found to be reproducible. Although it is not possible to determine what this number should be if all is well, we cannot expect 100% reproducibility. For example, the expression levels of some genes will not exhibit meaningful biological variation between samples, and so should not show any correlation. And since some genes were simulated to represent annotation errors, it is interesting to note that 18% of those poor genes are found to be reproducible using the 99th pecentile of null ICCs as a threshold. This should not come as a surprise however, since the genome is highly connected, the distribution of correlations between randomly selected pairs of genes should exceed any well-defined null distribution.

To illustrate the benefits of using integrative correlation coefficients to select reproducible genes, we extended the simulation by including 2 real, phenotypic groups in each of the simulated studies, and searched

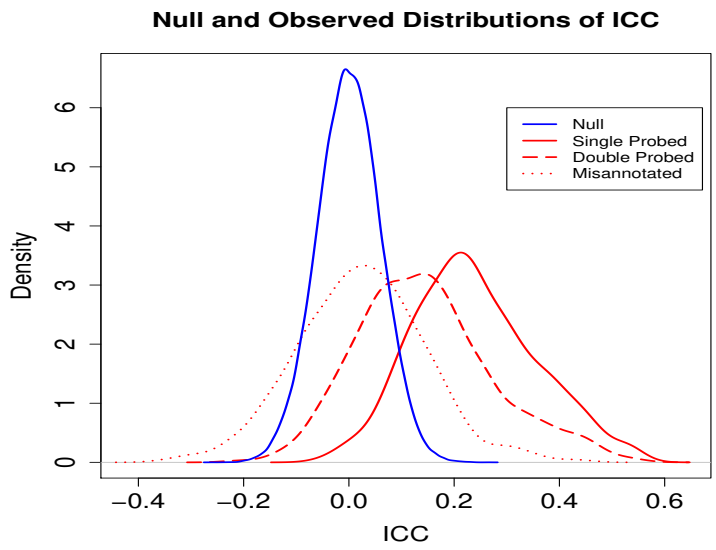

Figure 1: The observed integrative correlations are plotted in red by simulation group, and show the expected decline as the probes used in each study become more independent. Interestingly, even the *misannotated* genes show slightly higher integrative correlations than the null distribution. This is probably due to a high degree of co-expression throughout the genome, so that two randomly selected genes have a reasonable probability of being correlated.

for differentially expressed genes, comparing results across study by integrative correlation level. Two important classes of breast cancer are determined by the expression level of the estrogen receptor gene. Those cancers that express the gene, here called ER+ tumors, tend to be less agressive than the ER- tumors, and the two types are sufficiently different that we can be confident that the 3000 gene simulation will include a number that are actually differentially expressed in this phenotype. Accordingly, we used t-statistics to identify differentially expressed genes in each study. Figure 2 shows the cross-study correlation between t-statistics, overall, and after stratifying genes by integrative correlation. Although the phenoytype was not used in calculating the ICC, the correlation between t-statistics improves signifcantly when non-reproducible genes are filtered out. In this example, as above, the 99th percentile of null integrative correlations was used as the threshold for calling genes reproducible.

2.1 A Reformulation

It is very easy to understand the ICC as a correlation of correlations, but reformulating it in matrix notation brings a number of important statistical and algorithmic features of the method into the spotlight.

To get notation and underlying assumptions out of the way first, suppose that S_a and S_b are two microarray studies, with sample sizes of n_a and n_b respectively, and a total of m common genes. The within-study correlation of two genes can be written as the inner-product of appropriately standardized expression values, so we can go ahead and standardize in advance. And so in study S_a for example, each gene is assumed to have a mean expression value of 0, and a variance of $1/n_a$. For a particular gene x, we will use x_a to denote the standardized expression values of the gene in study S_a and let A describe the $m-1 \times n_a$ matrix of standardized expression values for all other genes. Notation for study S_b is, of course, identical. It would perhaps make more sense to use A_x and B_x since these matrices depend on the choice of x, but we will sacrifice the subscripts for cleaner notation (and in anticipation of soon eliminating the dependence on x).

Finally, we note that mean-centering a vector is a linear operation, and so can be easily transcribed into

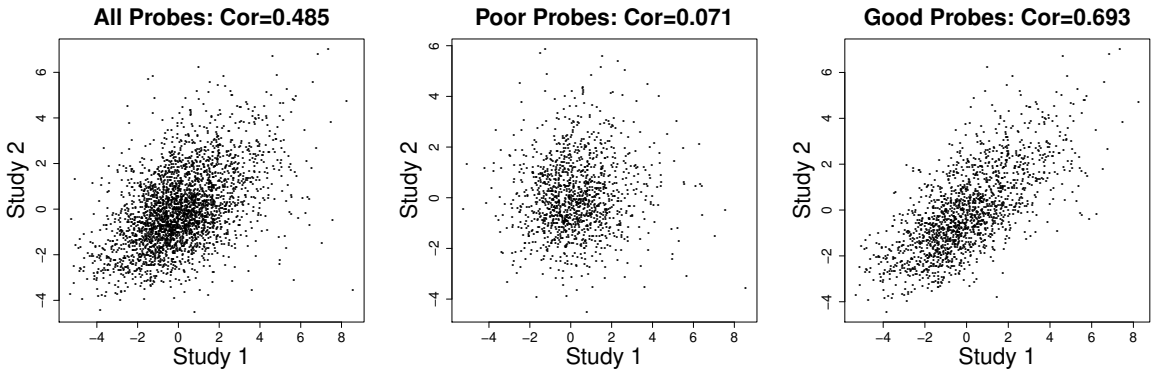

Figure 2: T-statistics comparing expression samples for ER- breast tumors and ER+ breast tumors were calculated for each probe in each study Probes are grouped according to ICC. Good probes have an ICC greater than the 99th percentile of the null distribution, while bad probes have ICC values below that threshold. Great probes are a subset of good probes having ICC greater than every cross-study sample correlation.

matrix notation. Thus, according to common usage I_m will denote the $m \times m$ identity matrix and cE_m the $m \times m$ matrix with every element equal to c. In this notation, if v and w are two random vectors of length m then $[I_m - 1/mE_m]v = v - \overline{v}$, and the quadratic form, $v^t[I_m - 1/mE_m]^2 w = cov(v, w)$. For notational simplicity, we will let $\mathcal{I}$ stand as shorthand for the squared matrix at the center of the covariance, $[I_m - 1/E_m]^2$.

Thus, we can write the integrative correlation coefficient for gene x in studies S_a and S_b as

$$\frac{x_a A^t \mathcal{I} B x_b^t}{\sqrt{x_a A^t \mathcal{I} A x_a^t}\sqrt{x_b B^t \mathcal{I} B x_b^t}} \tag{1}$$

and define *integrative covariance* and *integrative variance* as $x_a A^t \mathcal{I} B x_b^t/(m-1)$ and $x_a A^t \mathcal{I} A x_a^t/(m-1)$ respectively.

Although formally the data matrices A and B do not include x and so depend on the choice of gene, it makes little difference in practice whether x is deleted. The downside to leaving x in is that the correlation of x with itself is of course 1 in each dataset, so the second, correlation of correlations, will be slightly inflated, but unless the number of common genes, m, is very small, the effect is negligible. In contrast, the positive consequences can be substantial, and include, perhaps most dramatically, a notable reduction in computational complexity.

Here the key is that the central term in the integrative variance of x_a is the $n_a \times n_a$ sample covariance matrix $A^t \mathcal{I} A$, while the integrative covariance is centered around the analagous $n_a \times n_b$ matrix $A^t \mathcal{I} B$ in which the i, j-th entry is the covariance between sample i of study S_a and sample j of study S_b. When integrative correlation is calculated *by the book*, it is necessary to produce m very slightly different versions of these matrices, and the savings obtained by doing it once and for all is remarkable.

2.2 Interpretations

So, once again what does it mean for a gene to be *reproducible* across studies? Suppose that in study S_a, sample i expresses gene x most highly and imagine that we have some plausible rule (not dependent on x) for identifying those study S_b samples that are most like i. Then it should be a very good sign if x is also very highly expressed in those samples. This is exactly what the integrative correlation measures. The cross-study, sample covariance matrix $A^t \mathcal{I} B$ describes the similarity between samples, and pre- and

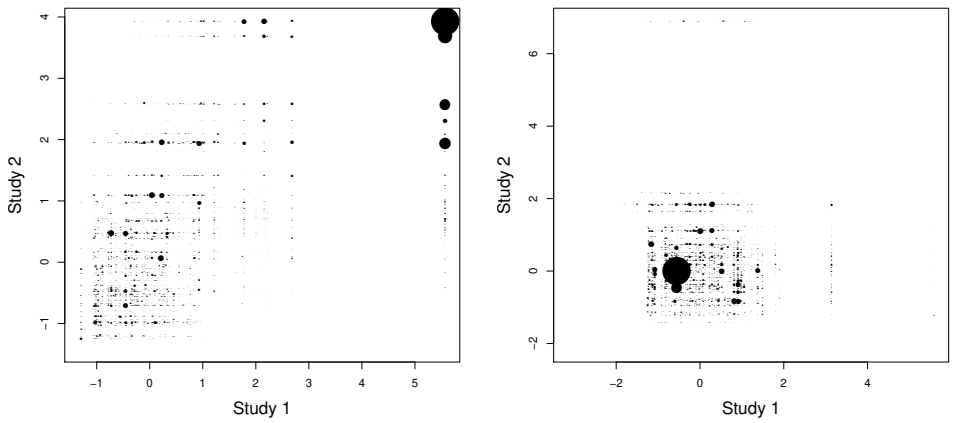

Figure 3: In each of these plots, $n_a \times n_b$ points are laid out in a grid, according to the expression values of the gene in each study. The plotting points vary in size, in proportion to the square of the covariance of each pair of samples, but the same distribution of sizes is seen in each figure. We chose to use the square because it makes the points with the lowest covariances disappear completely, showing the patterns of co-expression for the remaining genes to better effect. The gene shown on the left has an integrative correlation near 0.5, and the largest points tend to fall along the main diagonal. The gene plotted on the right has a correlation near zero. The amount of white space around the margins is the most prominent feature of this plot; this is due both to a lack of points in those regions, and because the samples having the most extreme expression values for this gene in each study are very poorly correlated, so the few points that do fall around the margins are extremely small.

post-multiplication by x_a and x_b^t scores the extent to which similar samples have similar expression for gene x.

Put another way, the integrative covariance is a weighted, cross-study covariance for gene x, in which each sample in study S_a is paired in turn with each sample in S_b, but the pairings are weighted by the sample similarity measures so that pairings between similar samples contribute more to the final sum. This idea is represented graphically in Figure 3 in which two genes with very different integrative correlations are shown. In each of these plots, $n_a \times n_b$ points are laid out in a grid, according to the expression values of the gene in each study. The plotting points vary in size, in proportion to the square of the covariance of each pair of samples. We chose to use the square because it makes the points with the lowest covariances disappear completely, showing the patterns of co-expression for the remaining sample pairs to better effect.

This view of the ICC has a few important implications for the usage and interpretation of the measure. Suppose that the expression level of gene x is by itself completely deterministic of some phenotype P, and that gene x is measured perfectly in both studies, and closely linked to P as expected. By any reasonable definition, the effect of expression in x should be called reproducible, but the integrative correlation coefficient will not be significant for this gene unless the phenotype is for some reason broadly associated with the expression of many other genes as well. This is a highly artifical example, but illustrates the point that genes are found to be reproducible to the extent that their expression patterns recapitulate broad transcriptional patterns in the data. A related problem can arise in the presence of batch effects, which may dominate the co-expression patterns of large numbers of genes that otherwise show very little variation, enormously inflating their intregrative correlations. Below, in Section 3.3 we discuss a modification designed to help with these situations, illustrating with an example of data compromised by batch effects.

3 Statistical properties of the integrative correlation coefficient

In this section, we describe several telling characteristics of the integrative correlation coefficient. The most straightforward results are mentioned briefly for completeness, and are not proved here. Along the way we point out where open questions remain.

Theorem 1 *The integrative correlation is invariant to re-ordering of data columns, or data rows (as long as rows are re-ordered in tandem in both studies, to retain gene cross-annotations).*

Proof omitted

Theorem 2 *After sample covariance values are calculated, the vectors x_a and x_b can be re-scaled without changing integrative correlation.*

Proof Any scaling factor appears in both the numerator and the denominator of 1 and is cancelled out.

3.1 Upper bounds on integrative correlation coefficients

A slightly different view of the ICC puts it in context in relationship to canonical correlation theory and multiple linear regression. The terms $x_a A^t$ and Bx_b^t represent linear combinations of the columns of A and B respectively with coefficients defined by the expression values of gene x, and the integrative correlation for x is simply the Pearson correlation of these two linear combinations. Accordingly we can look to canonical correlation theory, which is concerned with characterizing maximally correlated, linear combinations of sets of random variables, for the largest integrative correlation that can be obtained between two studies. Similarly, multiple linear regression can be used to calculate particular, gene-specific upper bounds

Canonical correlation is typically applied where the same samples have been measured on two different but related sets of variables, with the goal of finding the combinations of the variables in each set that maximize the correlation of the two. The solution is obtained via an eigenvalue decomposition of appropriately selected matrices, where the largest eigenvalue is the square of the maximum correlation that can be obtained, and a pair of corresponding eigenvectors describe the linear combinations that achieve that maximum. Subsequent eigenvalues and eigenvectors describe the maximal correlations that can be obtained over nested, orthogonal subspaces. To apply it here, it is necessary to reverse the traditional roles of the samples and variables since we are correlating linear combinations of samples over genes.

Theorem 3 *Integrative correlation coefficients are bounded from above by the largest canonical correlation, calculated across studies, over common genes.*

Proof Having cast the ICC as a correlation between linear combinations of two sets of random vectors, the fundamental result is immediately established, in that the largest canonical correlation is the maximum possible correlation between linear combinations of samples from each set.

There is, however, one issue that needs to be addressed. Since x_a and x_b are centered in advance, the coefficients for ICC-defined linear combinations necessarily sum to 0, while the calculated canonical correlation is not subject to the same restriction. Thus it is possible that the upper bound calculated in this way is unachievable.

As it turns out, however, the canonical correlation does provide a tight upper bound, it does not matter whether $x - a$ and x_b are centered since each row of A and B already is. Indeed, suppose that x_a is not centered, having mean=$\mu \neq 0$, and define $x'_a = x_a - (\mu_a ... \mu_a)$. Then

$$\begin{aligned} x_a A^t &= (x' + (\mu_a ... \mu_a))A^t && (2)\\ &= x'A^t + (\mu_a ... \mu_a)A^t && (3)\\ &= x'A^t + \mu_a 0 && (4)\\ &= x'A^t, && (5) \end{aligned}$$

since each row of A has mean 0. ∎

In the same way, multiple linear regression offers gene-specific upper bounds on the integrative correlations, though these are necessarily conditional on one set of study-specific expression values.

Theorem 4 *Define* $y = x_a A^t \mathcal{I}^{1/2}$ *and find* $\hat{\beta}_x$ *as the least squares solution to the linear equation*

$$y = \mathcal{I}^{1/2} B \beta_x^t + \epsilon.$$

Then the standardized vector $(\hat{\beta}_x - \text{mean}(\hat{\beta}_x))/\text{var}(\hat{\beta}_x)$ *is the linear combination of study* $\mathcal{S}_b$ *samples that would maximize the ICC, and the square root of the multiple* R^2 *is an upper bound on the conditional integrative correlation coefficient for the gene.*

Proof In 1966, Chow[4] showed that the least squares solution to a multiple linear regression problem maximizes the square of the correlation between the response variable Y and linear combinations of the predictors X. As in Theorem 3, the limitation in ICC to centered random variables is not a restriction. ∎

3.2 The null distribution

When the integrative correlation coefficient was originally described[24], a null distribution was generated by calculating a full set of ICCs after individually permuting each row of each data matrix. This might be done several times and the results concatenated, depending on how thoroughly one wishes to characterize the distribution. After such a permutation, however, all samples are independent and so the elements of the cross-study sample covariance matrix $A^t \mathcal{I} B$ all estimate zero, while the within-study covariance matrices $A^t \mathcal{I} A$ and $B^t \mathcal{I} B$ is diagonal. The resulting null distribution shows too little variation, estimating ICCs for genes measured in unrelated studies, rather than approximating what we would see if a single unreproducible gene were evaluated in the current studies.

The very simple solution is to replace x and y in Equation 1 with permuted versions of the same while leaving the covariance matrices untouched, thus modelling the effects of comparing unrelated expression values within the actual context of the two studies under consideration. Alternatively, x and y can be drawn by sampling independently from the standard normal, or any other distribution, simulations, not included here, show that results are very similar no matter how the random variates are drawn.

3.2.1 Asymptotic null distribution

The large sample distributional properties of the null distribution are one of the most interesting aspects of the ICC method. It turns out that with a few reasonable assumptions on the distributions of the sample expression profiles, calculating the Pearson correlation between random linear combinations of the samples from each study is asymptotically equivalent to calculating the Pearson correlation between randomly selected, individual samples from each study.

The very simple form of the resulting asymptotic distribution is derived in Theorem 5 below, and as we go on to demonstrate, offers a very accurate approximation to the true null even with moderate sample sizes.

Theorem 5 *Assume that*

1. *the univariate distribution of the row-standardized expression values in each column is the same for all column/samples*
2. *sample expression profiles are selected from a population in which pairwise sample correlations have a mean of zero, and a variance of* σ^2
3. *to generate the null distribution, independent gene expression values are drawn as i.i.d.* $N(0, n_l^{-1/2})$ *random variables for each sample, in each study.*

then as $n_a, n_b \to \infty$ *the null distribution converges to* $N(0, \sigma^2)$.

Note that it is not necessary that all samples be drawn from the same multivariate distribution, nor that they be independent. These would be unrealistic requirements for microarray studies, and independence would in fact invalidate the ICC.

Proof We will argue in three steps that the denominator converges in probability to 1, while the numerator converges in distribution to a Gaussian distribution with the given mean and variance.

1. Since the entire collection of expression values has been standardized gene by gene to have a mean of 0 and a variance of 1, then under asssumption 1 each sample must have that same distribution and sample covariances are in fact sample correlations. At the same time expressions can be simplified by eliminating the matrix $\mathcal{I}$ since column centering is not necessary. Thus, $x_a A^t \mathcal{I} A x_a^t/(m-1) = x_a \Sigma_a x_a^t$, which has an expected value of $tr(\Sigma)/n_a$ and a variance of $2tr(\Sigma^2)/n_a^4$ [12]. Because Σ is in fact a sample correlation matrix, $tr(\Sigma)/n_a \to 1$ and as $n_a \to \infty$ and $0 \leq \lim_{n_a \to \infty} 2tr(\Sigma^2)/n_a^4 \leq \lim_{n_a \to \infty}(n_a + n_a^2)/n_a^4 = 0$

2. Under assumption 1, $x_a A^t \mathcal{I} B x_b^t/(m-1) = x_a \Sigma_{ab} x_b^t$ which can in turn be rewritten as a sum of products of random variables, $\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} x_{ai} x_{bj} \sigma_{ij}^2$ where σ_{ij}^2 is regarded as a random variable according to assumption 2. As x_{ai}, x_{bj} and σ_{ij}^2 are mutually independent, zero-mean random variables, the expection of each term and of the final sum, is 0.

   The variance calculation is slightly more complicated but nonetheless very managable since in 1962, Goodman derived exact expressions for the variance of a product of random variables [9], which in the case of mutually independent, zero mean random variables is simply the product of variances. Applying this to the summands in the integrative covariance term yields $Var(x_{ai} x_{bj} \sigma_{ij}^2) = sigma^2/(n_a n_b)$. Although the entire set of such products is not mututally independent, they are pairwise independent, so $Var(\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} x_{ai} x_{bj} \rho_{ij}) = n_a n_b \sigma^2/(n_a n_b) = \sigma^2$.

3. Finally, since the null integrative correlation is a sum random variables with finite first-fourth moments, the C.L.T. ensures that the limiting distribution is Gaussian. ∎

We can extend the illustrative example to demonstrate rates of convergence to the asymptotic null distribution. Recollect that we used a publically available data set to simulate two studies, each with 100 samples, and 3000 common genes. The quantile-quantile plots in Figure 4 shows how well the asymptotic null distribution approximates the observed null distribution for 5, 25, or 100 samples per study, and 300, 1000 or 300 genes.

What is perhaps most interesting here that the variance of the null distribution does not shrink as the sample size increases. In most testing situations, it is possible to reach statistical significance by beating a null distribution in a large-sample experiment without achieving the practical signficance of a meaningfully large effect size, but this setting is very different. With appropriately chosen values for the cutoff, the measure of reproducibility suggested here can be used to identify genes that are both statistically signficant and practically significant.

3.3 A Variation on Integrative Correlation

As it is described in Section 2.2, "the cross-study, sample covariance matrix $A^t \mathcal{I} B$ describes the similarity between samples, and pre- and post-multiplication by x_a and x_b^t scores the extent to which similar samples have similar expression for gene x". This points the way to a class of extensions of standard ICC in which a different measure of sample similarity is used in place of the sample covariance matrix. This could be as simple as restricting the set of variables used in calculating the sample covariance matrix or could utilize external variables to establish sample similarity, perhaps non-molecular clinical characteristics of each patient, for example. Depending on how the similarity matrix is calculated, the connections to canonical correlation theory and regression analysis, as well as the asymptotic, null distribution theory may not be valid, though the suggested null distribution can still easily be determined in all cases.

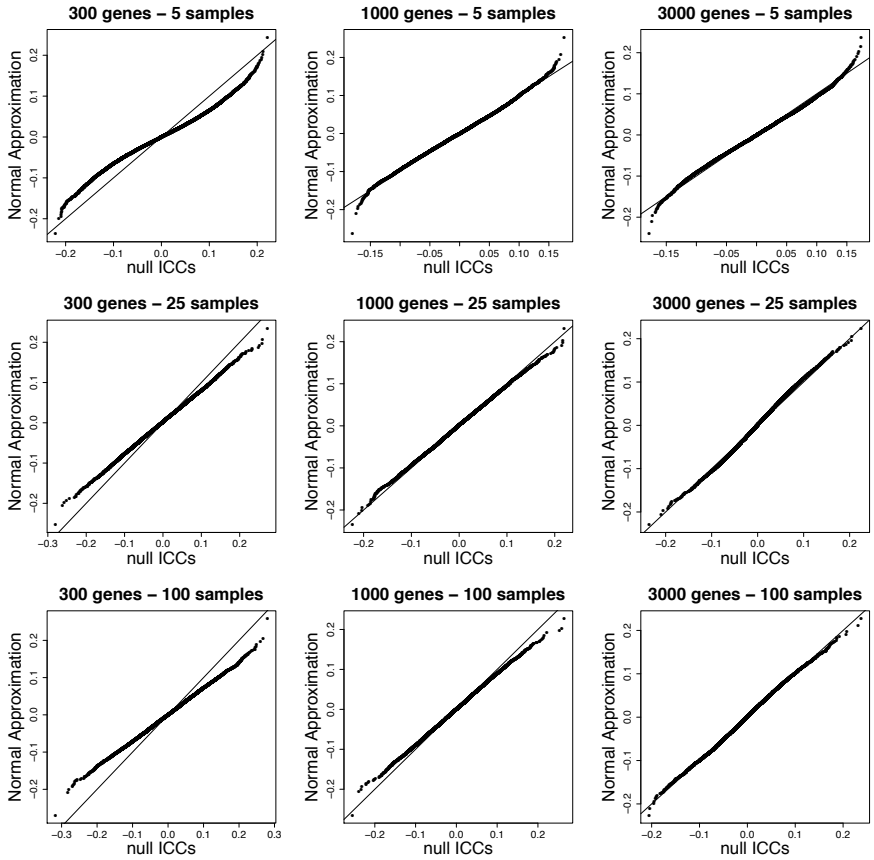

Figure 4: Quantile-quantile plots compare the observed null distribution, on the x-axis, to the asymptotic Gaussian null, on the y-axis of each plot in the figure. The number of genes common to the two studies increases by row, from 300 to 3000, while the number of samples in each study increases by column, from 5 to 100. The variance for the asymptotic null distribution was determined by using all 3000 common genes, and the total sample sizes of 100 samples per study. The distribution converges quickly to normality as the sample size increases, regardless of the number of common genes. However, when there are few common genes, the null variance is increased.

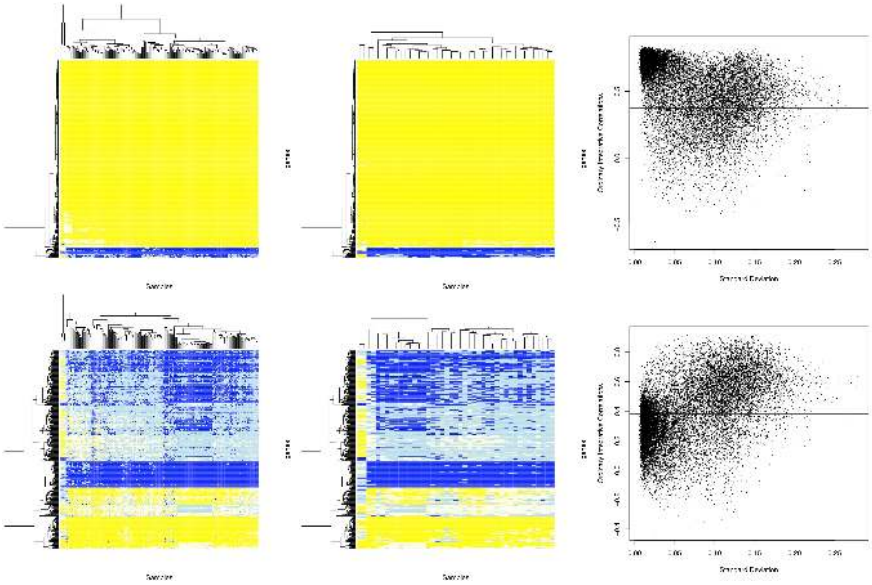

Figure 5: Each heatmap shows TCGA lung cancer methylation data. Squamous cell carcinomas are on the left, adenocarcinomas on the right. In each, genes are represented as rows while samples are represented as columns. Methylation level is represented by color intensity on a yellow-blue scale, where bright yellow spots have very low methylation levels ($\leq 20\%$), while bright blue spots have very high levels of methylation ($\geq 70\%$). The top row of figures illustrates the difficulty that can arise when standard integrative correlations are used in this data. The genes with the highest integrative correlations, shown here, include many that are completely unmethylated in all samples, in which batch effects too small to be seen on this 4 color scale dominate the correlation structure. The bottom row includes a similar number of genes, again those having the highest integtative correlations, where this time the sample similarity matrix is calculated without scaling genes so that gene contributions are proportional to their variance. The scatterplots in the rightmost column show the relationship between the s.d. of each probe, shown on the horizontal axis, and the integrative correlations on the vertical axis.

We see this as a viable solution to some of the limitations described in Section 2.2. The case of a single gene determining a phenotype on its own may be a lost cause, but in a less extreme case, one might base the sample covariance matrix on genes in a particular pathway, to focus on the co-expression patterns most relevant to pathway function, and elevate related genes, whose expression patterns may be very rare within the whole genome, to more prominent positions in the integrative correlation analysis.

This approach is demonstrably effective when the integrative correlation coefficient is applied in the presence of batch effects, which, though they may be of small absolute magnitude, can create the illusion of widespread coordination among large numbers of otherwise unrelated, even unexpressed, genes. This seems to be a particular problem in copy number or promoter methylation array data, where the vast majority of genes can be expected to have 2 copies in every sample (or to be unmethylated in every sample). A simple but crude fix would be to base the sample covariance matrix on the subset of the genes with the highest sample to sample variation, excluding those in which small batch effects represent the only variation from the sample similarity calculation, although integrative correlations are still calculated for all genes. Slightly more elegant is to reduce the influence of these genes by centering, but not scaling each gene before calculating the sample covariance matrices, so each gene makes a contribution proportional to its variance. This is what we do below.

We will illustrate the method using data from the Cancer Genome Atlas (TCGA) data, where samples are processed in several batches, with well documented effects[20]. Figure 5 illustrates the problem, and

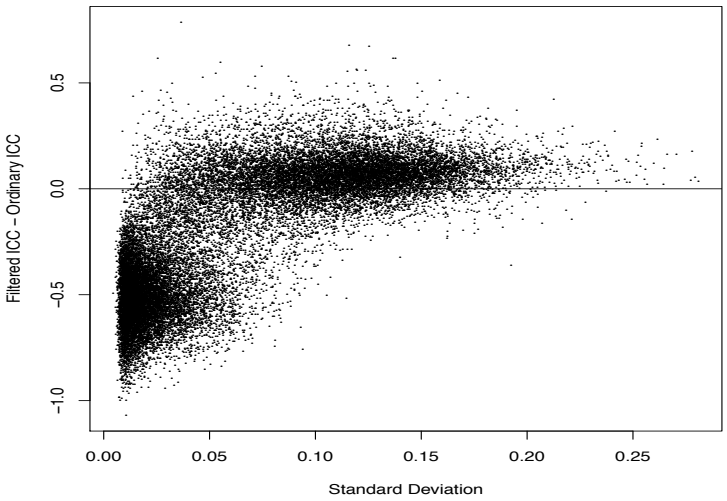

Figure 6: The difference between filtered ICC and ordinary ICC for each gene is plotted on the vertical axis, the standard deviation on the horizontal axis.

our solution, in TCGA lung cancer promoter methylation data. In this data, methylation takes values in the interval [0,1], interpreted as the proprotion of tumor cells in which the gene is methylated, and ICC is calculated between two tumor types, adenocarcinomas and squamous cell carcinomas. Fewer than 25% of the genes show any appreciable variation in this index and the vast majority are in fact completely unmethylated in all normal and tumor samples in either of the tumor types. The batch effects show up very clearly as patterns common to all of these genes, completely determing the sample covariance matrix. The standard integrative correlation exhibits a truly embarassing preference for low variance genes, assigning high coefficients to nearly all genes with standard deviation below 0.03, a value that can be achieved if 1 sample in 100 is 30% methylated, the rest completely unmethylated.

By weighting genes in proportion to their variance, however, we are able to redefine sample similarity completely, so that genes with substantial sample to sample variation, and potentially interesting methylation patterns common to both studies achieve the highest integrative correlations. Figure 6 shows the differences more clearly. ICC drops dramatically for nearly all genes with standard deviations below 0.03 and increases slightly for the vast majority of genes with standard deviations above 0.10. This elegant solution establishes a clear proof of principle for a valuable class of extenstions to the standard integrative correlation coefficient.

4 Discussion

The meta-analysis of genetic data, and more generally, the integration of data from multiple sources, is a rapidly developing research problem in bioinformatics, with a number of recent publications discussing principles, risks and benefits, or proposing or evaluating methods. [13, 21, 3, 10, 26, 32, 14, 15]. These publications reflect a variety of goals: not only do we wish to increase sample sizes by combining similar studies, but also want to translate findings from cell lines to primary tumors, integrate expression data with copy number, protein or methylation and even from animal models to humans. At the same time, the very concept of the gene itself is becoming more complex[2, 25, 22] with recent studies suggesting that genes can be spread of very large genomic regions, with substantial overlap of other genes and regulation by transcription factors binding to sometimes quite distant sites[8]. Other studies report that great deal

of non-coding RNA is expressed and apparently plays regulatory roles [6], and identifying possible roles for pseudogenes [1]. It is more difficult and more important than we ever understood before, to be sure that the 'gene' we identifed in one study is the same gene we are now looking at in another.

In this paper we offer an important reformulation of the ICC that brings the concept of reproducibility that is operative here into sharp focus, and makes clear the circumstances under which such genes can be accurately identified. Ultimately, the idea is a very general one, a gene is reproducible if otherwise similar samples have similar gene expression values. The same fundamental concept is operative if one calculates a within-study t-statistics with respect to some binary phenotype, and call a gene reproducible if it has similar t-statistics in the two studies; though the definition of sample similarity, and precise method of scoring reproducibility are different. As we demonstrate, by varying the sample similarity matrix, the ICC can be tuned to specific phenotypes or pathways, or to minimize the influence of batch effects.

We present a unique permutation null distribution that precisely captures the key source of variation, and prove a very interesting asymptotic result for it, showing that the null distribution is asymptotically Normally distributed, but that the variance does not necessarily shrink as the sample size increases. In consequence, the classical disconnect between statistical and practical signficance, whereby vanishingly small effects become statistically significant when the sample size is large enough, is substantially reduced here.

The integrative correlation coefficient reduces the problem of identifying reproducible genes to bare essentials. The concept of reproducibility operative here is very simply that a gene is reproducibly expressed across studies if similar samples in each study have similar gene expression. The method is non-parametric, and easily adapted to use any sample similarity matrix. The method does depend on correlation, or more broadly on inner products, to determine when similar samples have similar gene expression, but as this addresses a practical need to minimize cross-study differences in the location and scale of gene expression measures, we do not think it very restrictive. The default version is based on Pearson correlation, but by calculating rank-based sample correlations, and ranking expression within gene before standarizing, a robust Spearman version is easily implemented.

Although developed for microarray analysis, the method might be applied to any high dimensional data in which the variables can be expected to have complex dependence structure. Examples would include extensive nutrition, health or behavorial surveys, where it might be expected that cultural differences, for example, might make it difficult to compare results for some questions across populations. Nor is it necessary that the variables be continuous as long as their levels are ordered; the interpretation of the correlation coefficient as a measure of similarity is maintained in this situation.

References

[1] Evgeniy S Balakirev and Francisco J Ayala. Pseudogenes: are they "junk" or functional dna? *Annu Rev Genet*, 37:123–151, 2003.

[2] M S Boguski and G D Schuler. Establishing a human transcript map. *Nat Genet*, 10:369–371, 1995.

[3] Anna Campain and Yee Hwa Yang. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, 11:408, 2010.

[4] Gregory C. Chow. A theorem on least squares and vector correlation in multivariate linear regressions. *JASA*, 61(314):413–414, 1966.

[5] Leslie Cope, Xiaogang Zhong, Elizabeth Garrett-Mayer, Edward Gabrielson, and Giovanni Parmigiani. Cross-study validation of the molecular profile of brca1-linked breast cancers. *Unpublished Manuscript*, 2004.

[6] Fabricio F Costa. Non-coding rnas: could they be the answer? *Brief Funct Genomics*, Dec 2010.

[7] Elizabeth Garrett-Mayer, Giovanni Parmigiani, Xiaogang Zhong, Leslie Cope, and Edward Gabrielson. Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, 9(2):333–354, Apr 2008.

[8] Mark B Gerstein, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode? history and updated definition. *Genome Res*, 17(6):669–681, Jun 2007.

[9] Leo A. Goodman. The variance of the product of k random variables. *JASA*, 57(297):54–60, 1962.

[10] Jemila S Hamid, Pingzhao Hu, Nicole M Roslin, Vicki Ling, Celia M T Greenwood, and Joseph Beyene. Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics*, 2009, 2009.

[11] D Neil Hayes, Stefano Monti, Giovanni Parmigiani, C Blake Gilks, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A Socinski, Charles Perou, and Matthew Meyerson. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol*, 24(31):5079–5090, Nov 2006.

[12] Ronald R Hocking. *Methods and Application of Linear Models: Regression and the Analysis of Variance*. Wiley Series in Probability and Statistics, 1996.

[13] Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, Feb 2008.

[14] Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, Nov 2006.

[15] Pingzhao Hu, Celia M T Greenwood, and Joseph Beyene. Using the ratio of means as the effect size measure in combining results of microarray experiments. *BMC Syst Biol*, 3:106, 2009.

[16] Homin K Lee, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094, Jun 2004.

[17] Jae K Lee, Kimberly J Bussey, Fuad G Gwadry, William Reinhold, Gregory Riddick, Sandra L Pelletier, Satoshi Nishizuka, Gergely Szakacs, Jean-Phillipe Annereau, Uma Shankavaram, Samir Lababidi, Lawrence H Smith, Michael M Gottesman, and John N Weinstein. Comparing cdna and oligonucleotide array data: concordance of gene expression across platforms for the nci-60 cancer cells. *Genome Biology*, 4:doi:10.1186/gb–2003–4–12–r82, 2003.

[18] Jae K Lee, Charles Coutant, Young-Chul Kim, Yuan Qi, Dan Theodorescu, W Fraser Symmans, Keith Baggerly, Roman Rouzier, and Lajos Pusztai. Prospective comparison of clinical and genomic multivariate predictors of response to neoadjuvant chemotherapy in breast cancer. *Clin Cancer Res*, 16(2):711–718, Jan 2010.

[19] Jae K Lee, Dmytro M Havaleshko, Hyungjun Cho, John N Weinstein, Eric P Kaldjian, John Karpovich, Andrew Grimshaw, and Dan Theodorescu. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc Natl Acad Sci U S A*, 104(32):13086–13091, Aug 2007.

[20] Jeffrey T Leek, Robert B Scharpf, Hctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–739, Oct 2010.

[21] Shuya Lu, Jia Li, Chi Song, Kui Shen, and George C Tseng. Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340, Feb 2010.

[22] Tim R Mercer, Dagmar Wilhelm, Marcel E Dinger, Giulia Sold, Darren J Korbie, Evgeny A Glazov, Vy Truong, Maren Schwenke, Cas Simons, Klaus I Matthaei, Robert Saint, Peter Koopman, and John S Mattick. Expression of distinct rnas from 3' untranslated regions. *Nucleic Acids Res*, Nov 2010.

[23] Alykhan S Nagji, Sang-Hoon Cho, Yuan Liu, Jae K Lee, and David R Jones. Multigene expression-based predictors for sensitivity to vorinostat and velcade in non-small cell lung cancer. *Mol Cancer Ther*, 9(10):2834–2843, Oct 2010.

[24] Giovanni Parmigiani, Elizabeth S. Garrett-Mayer, Ramaswami Anbazhagan, and Edward Gabrielson. Cross-study comparison of gene expression data sets for the molecular classification of lung cancer. *Clinical Cancer Research*, 10(9):in press, 2004.

[25] Chris P Ponting and T. Grant Belgard. Transcribed dark matter: meaning or myth? *Hum Mol Genet*, 19(R2):R162–R168, Oct 2010.

[26] Andrey A Shabalin, Hakon Tjelmeland, Cheng Fan, Charles M Perou, and Andrew B Nobel. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160, May 2008.

[27] Steven C Smith, Alexander S Baras, Jae K Lee, and Dan Theodorescu. The coxen principle: translating signatures of in vitro chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer. *Cancer Res*, 70(5):1753–1758, Mar 2010.

[28] Steven C Smith, Dmytro M Havaleshko, Kihyuck Moon, Alexander S Baras, Jae Lee, Stefan Bekiranov, Daniel J Burke, and Dan Theodorescu. Use of yeast chemigenomics and coxen informatics in preclinical evaluation of anticancer agents. *Neoplasia*, 13(1):72–80, Jan 2011.

[29] Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, and John A Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, 2005.

[30] Matthew D Wilkerson, Xiaoying Yin, Katherine A Hoadley, Yufeng Liu, Michele C Hayward, Christopher R Cabanski, Kenneth Muldrew, C Ryan Miller, Scott H Randell, Mark A Socinski, Alden M Parsons, William K Funkhouser, Carrie B Lee, Patrick J Roberts, Leigh Thorne, Philip S Bernard, Charles M Perou, and D Neil Hayes. Lung squamous cell carcinoma mrna expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res*, 16(19):4864–4875, Oct 2010.

[31] X Zheng-Bradley, J Rung, H Parkinson, and A Brazma. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*, 11(12):R124, Dec 2010.

[32] Xiaogang Zhong, Luigi Marchionni, Leslie Cope, Edwin S. Iversen, Elizabeth S. Garrett-Mayer, Edward Gabrielson, and Giovanni Parmigiani. Optimized cross-study analysis of microarray-based predictors. Technical Report Working Paper 129, The Johns Hopkins University, Baltimore, Maryland, 2007.