A RATIOMETRIC-BASED MEASURE OF GENE CO-EXPRESSION

Thesis by

Anna C.T. Abelin

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2014

(Defended 30th of April 2014)

ACKNOWLEDGEMENTS

First and most I want to thank my family. Being on the other side of the Atlantic Ocean has not been easy but you have always been there for me. Especially my father, **Harald Abelin,** PhD, to whom my entire scientific career owes great gratitude. His support and shared interest in understanding not only 'how' but also 'why', has nourished my curiosity and thirst for knowledge. But above all he has taught me to never doubt my capabilities, only my hypothesis. I would also like to thank my great friend **Agnes Lukaszewicz,** PhD, for her relentless support in all stages of my PhD. The third person that has been very important to me is **Jan-Erik Nyström,** PhD. He has encouraged and mentored me as the times got tough, which I am very thankful for. Finally, I also want to thank my supervisor Barbara Wold and my committee members: Paul Sternberg, Marianne Bronner, and Mitchell Guttman.

ABSTRACT

Current measures of global gene expression analyses, such as correlation and mutual information-based approaches, largely depend on the degree of association between mRNA levels and to a lesser extent on variability. I develop and implement a new approach, called the Ratiometric method, which is based on the coefficient of variation of the expression ratio of two genes, relying more on variation than previous methods. The advantage of such *modus operandi* is the ability to detect possible gene pair interactions regardless of the degree of expression dispersion across the sample group. Gene pairs with low expression dispersion, i.e., their absolute expressions remain constant across the sample group, are systematically missed by correlation and mutual information analyses. The superiority of the Ratiometric method in finding these gene pair interactions is demonstrated in a data set of RNA-seq B-cell samples from the 1000 Genomes Project Consortium (*1*). The Ratiometric method renders a more comprehensive recovery of KEGG pathways and GO-terms.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS AND TABLES

Number

NOMENCLATURE

CV. Coefficient of variation

FPKM. Fragments per kilobase of transcript per million fragments sequenced

MI. Mutual information

PCA. Principal component analysis

PE. Pearson correlation

RA. Ratiometric method

SP. Spearman correlation

Chapter I

GLOBAL GENE EXPRESSION ANALYSIS

Introduction

One of the fundamental and longstanding goals of modern biology is to fully comprehend the role of each gene in the genome. How the genes contribute to the identity and functionality of a cell type. A main focus is to understand the specific mechanisms by which the genome and its genes act like a blueprint of the cell. It was early known that not only the sequence identity of a gene determined its role but also its temporal and spatial expression pattern. Furthermore, genes are influenced by their environment, for example, concurrently expressed genes. Their observed functions are sometimes determined or at least often calibrated by which interaction partners are contemporary.

In early genetic studies, the function of a gene was inferred by its spatio- and temporal pattern of expression. A classical approach still used today, is to perturb the expression of a gene, historically applied one gene at a time. The gene's function is then deduced from the observed effects on the cell state. Such single gene analysis limited the way functional pathways were studied. They were described as isolated and linear courses of events. Each gene's effect was seen as an independent action neutral to the remaining expressed genes. The study of Cyclin A, a gene involved in cell-cycle regulation, tells such a tale. From the early experiments Cyclin A was deemed essential in the cell-cycle machinery as its depletion was observed to be lethal both in Drosophila embryos (*2, 3*) and in knockout mice (*4*). More recent experiments, including multiple genes of interest, have now modified this notion. Its first reported lethality rather reflected its essential function at a specific stage of embryo development than being

ubiquitously essential in the cell-cycle (*5*). The simplified picture from single gene experiments was thus soon not entirely satisfactory. The cell should preferably be seen as an intricate network of cellular functions. As an average gene in higher metazoan organisms is participating in multiple cellular functions (*6-8*), gene-to-gene interactions are rather complex. The complexity arising not only from the sheer number of genes expressed but also from the huge number of gene-to-gene interactions within and between the cellular functions. The ribosome is an excellent example of a group of specific genes coming together in the cell to form a functional complex (*9-13*). Another type of gene-groupings is underlined as pathways. For example, signaling pathways, such as TOR-signaling (*14, 15*), receive and process information, generating a cellular response.

To explore this systematic approach a series of methods were invented that can measure gene expression in parallel, including microarrays (*16-19*), nanoString (*20*), Serial Analysis of Gene Expression (SAGE) (*21*), and RNA-seq (*22*). With these approaches a couple of hundred genes, up to entire transcriptomes, can be measured simultaneously providing a snapshot of global expression profiles. Alongside the experimental breakthroughs, a demand for analytical tools for processing these huge datasets grew. One of the major aims is to be able to precipitate out the subset of genes relevant to each active biological process and/or pathway from the thousands of expressed genes. The standard approach for discovering such gene groupings has been to study differential gene expression profiles. These are derived from collecting data from multiple cellular states extracting out co-fluctuating gene groups. The fruit of these efforts have fostered not only our general knowledge about cellular states (*23, 24*), but also disease and genetic disorders (*25, 26*), alongside resolving the intricate mechanisms conforming cellular processes (*27-34*). Thus global gene expression analyses made it possible

to apply a network approach, acknowledging the collective behavior of individual genes. One common denominator among these studies is the inclusion of multiple cellular conditions, such as tissue type, disease and control samples and/or developmental stages. By framing the questions in relation to the cellular differences across the sample groups, the resulting gene pair relationships will only include those having a differential expression pattern. This is additionally reinforced by the analytical methods applied, which only can detect gene pair associations if there are enough inter-sample fluctuations (*35*), in other words a differential expression pattern. Thus by current standards, gene pair relationships that do not fluctuate across the observed cellular states remains unreported, leaving a potentially important part of the cells' systematic structure undetected and deemed non-present. Despite researches effort in constructing alternative approaches that would capture previously undetected gene pair relationships, few differences in detection yield have been reported so far. Indeed, repeated investigations have concluded that available methods in many biological settings produce extremely similar results (*36, 37*). With this as the starting point, I set out to develop and implement a conceptually different approach where gene pair detection is made in a single cellular state without requirements of inter-sample expression fluctuations. By being able to analysis a singular cellular state alone, the detection of its gene pair relationships is decoupled from whichever other cellular states are included in the study. Analyzing a singular cell state enables gene pair detection purely based on that observed state. The detection is thus not induced, nor deterred, by the relative fluctuations compared to whichever other cellular states included in the study.

The limitations of standard methods

The multitude of methods currently available for gene expression analysis includes a wide variety of approaches and definitions. There are inherent limitations in these methods, decreasing their detection yield. Generalizing, the methods can be divided up into two groups: 1) gene pair association defining and 2) gene-grouping methods. The second group, including methods such as Principal component analysis, factor analysis and clustering, intent to simplify the information produced by the first group. The later containing methods like Pearson/Spearman correlation, Euclidean distance and Mutual information, in turn evaluate each individual gene pair, based on their specific association criteria. In a sense, the gene-grouping methods can therefore not be 'better' than what the association methods provide them with.

The standard methods for gene expression analysis are all mathematically rigorous and have repeatedly been shown to be of great value when interpreting biological data. Pearson correlation measures the similarity between pairwise expressions by evaluating the degree of covariance, Spearman judges according to the conformity to a monotonic function, Euclidean distance assess the 'bird'-distance between the two data points or vectors, while mutual information measures the level of information existing between the two gene expressions. At the end of this chapter, a section entitled *Common analytical methods* will provide a more detailed description of the mathematical foundation and examples of studies in which these methods have been applied.

One of the driving forces for doing this thesis work is the ongoing discussion of which methods are appropriate when and how the results depend on the data and method used (*38*).

This is a discussion worth being elaborated. Conclusions reached by studies involving method comparisons are contradictory. Some conclude that the methods analyzed gave mainly same results while other studies conclude that there are some differences. Steuer et al. compared Pearson correlation with mutual information when analyzing 300 mutations and chemical treatments of *S. cerevisiae*, and found that the two methods are almost completely in agreement regarding gene pair association strength (*36*). In more detail, they confirmed the well-known results of Pearson correlation distinguishing between positive and negative correlations, while mutual information does not. More importantly, the Pearson correlation is bound by the mutual information, in other words: Pearson did not find any gene pair relationships that mutual information did not, disregarding numerical and statistical errors. Furthermore, they did not find any non-linear relationships of higher association strength in the data set, and thus they conclude that the use of Pearson correlation would be justified, as it does not fail to detect a significant portion of the possible interactions. A similar study by Daub et al., compared the different results when applying mutual information and Pearson correlation on two large datasets (*37*), the general agreement between Pearson correlation and mutual information observed in last mentioned study is recovered, but additionally they make further findings that are more informative both on an experiential- and biological basis. In the first data set from 300 experimental conditions in *S. cerevisiae* (*39*), they encounter a smaller subset of gene pairs exhibiting high Pearson correlation and low mutual information index, ascribed to outliers in the data rendering false positive Pearson correlations. This suggests that if one cell type (in this case yeast) is perturbed chemically or by mutations and they are analyzed as part of an amalgamate of cell types, there is a substantial risk in the experimental design that one single perturbation produces an outlier large enough for the Pearson correlation to report a false positive. This erroneous reporting is not a risk when using mutual information, as it is robust

against outliers. From the second data set, including 102 experiments from 20 different human tissues, they report gene pairs having a high mutual information index and a low Pearson correlation, generated by two-regimes in the data set. Thus it became apparent that for certain gene pair combinations the tissue types can be divided into two different expression regimes. That cannot be picked up by Pearson correlation and thus is reported as uncorrelated, but the mutual information recognizes the dependence and accredits it as such. As a concluding note from this study; it is pertinent to observe these types of divergent gene pair recordings would not be made explicit to the investigator if not the expression scatter plot of each gene pair combination is individually surveyed. Thus if Pearson correlation or mutual information index is used straight up without further examination, which is often the case when several thousands of genes are analyzed, false positives could be included by Pearson correlation and two-regime relationships by mutual information under false pretends.

The high similarity between Pearson correlation and mutual information, Pearson correlation's sensitivity to outliers, and the 2-regime expression patterns detected by mutual information, are all observed in this study and their implications are discussed in following chapters.

General agreement between mutual information and Euclidean distance as similarity measure has also been shown, as in the earlier mentioned small 112-genes-study in rat cervical spinal cord (*40*). Even though mutual information captured potential functional relationships not detected by Euclidean distance, the authors concluded the overall results project a high degree of correspondence between the two methods. In 2002 Gibbons et al. explored the impact of choice of dissimilarity calculation, clustering algorithm, and number of pre-set clusters (*41*), including among others the more commonly used Pearson correlation, Euclidean distance, k-

mean clustering, self-organizing maps (SOM), and hierarchical clustering. They used 4 different data sets from studies in yeast and scoring the different choices by how well they clustered genes according to functionality set by *Saccharomyces* Genome Database (developed from the Gene Ontology Consortium). The results show little difference found between Pearson correlation, Euclidean distance and SOM. If anything, Pearson correlation is equally good or better analyzing non-ratio-style data, while Euclidean performed better on ratio-styled data.

The choice of method has been observed as important. In these cases the analysis has aimed towards grouping/clustering the samples according to cellular conditions rather than the genes according to expression profiles. For example, Priness et al. used four gene expression datasets; each one includes two distinct types of samples (for example, tumor tissue versus controls), thus providing a clear 'real' bi-clustering (*42*). In this setting the mutual information approach outperformed both Pearson correlation and Euclidean distance, with the two latter being indistinguishable.

The subgroup of gene expression analyses, which is dependent of previous gene pair interaction knowledge, will not be included in this thesis beyond the general method description given in this chapter. Their exclusion is not due to their lack of success, to the contrary, it has been shown that known gene interactions can be used to construct gene networks with the potential of sorting, for example, cancer types into informative sub groups. One such study showed that human breast and ovarian cancer types can be precipitated into subgroups by using receptor tyrosine kinase-triggered pathway signatures (*43*). The reason for excluding these types of approaches is that a dependence on previous knowledge can be fickle (*44*). Both in the uncertainty of the knowledge being correct, but also, of fully being aware

under which circumstances it is valid. Under certain cellular conditions some signature profiles could be correct while under others they are no longer recognized.

Another point worth discussing is the biological information actually revealed by the current methods. As example of how detailed the findings are. Methods like PCA and K-mean clustering both generate a small number of gene expression profiles, which potentially explain the discrete expression characteristics discerned in the data. But what they don't do is to give pairwise gene expression dependencies. Thus, the results provide indications of how groups of genes behave similarly but they lack how these genes are related to each on a gene-to-gene level. To this end, Pearson correlation, Euclidean distance, and mutual information are proficient, but studies harvesting this type of information are sparse. In general, the latter are used to produce a similarity matrix, which in turn, is further analyzed by generalizing methods, such as PCA or clustering. By following such a procedure, valuable biological information is lost. Firstly, if a few vital gene pair interactions happens to be associated with a large group of weak and semi-random gene interactions, they would fail to be detected in generalizing methods, such as PCA, as such a vector would be weakly loaded. Secondly, the ranking of the detected gene pair interactions is lost, thus a dimension of importance on the gene-to-gene level is absent. A ranking of the gene pair associations would provide further information of which interactions are strongest in the cell state, thus in which biological processes high expression calibration is found and by which genes it is personified.

A gene pair association can be interpreted as the degree of dependence two genes demonstrate, which presently has been translated into the degree two-gene expressions deviate from a mathematical condition of probabilistic independence. This, also called similarity

measurement, is defined slightly different by correlation, Pearson or Spearman, Euclidean distance, and mutual information, but their foundation in mathematical independence is the same. Thus there exists a wide range of approaches for interpreting biological relationships and sifting out the pertinent ones. Their foundation in divergence from mathematical independence is a statistical sound and well-founded philosophical argument. But what if the association defining of a gene pair is approaches differently? What biology would be revealed if you would deviate from the norm of mathematical independence and instead analyze gene expression data based on gene pairwise predictability? Simply put, a strong gene pair association is one where the two genes can be predicted by each other, such that their expression ratios across the samples are stable. Thus, the mathematical independence is of no importance, only the degree of predictability. This distinction might appear as slim but it will turn out to be of major importance shown in later chapter of this thesis.

The approach invented and developed in this thesis, called the Ratiometric method (RA), comprises one of the simplest conceptions of a gene pair association, namely the ratio of the two genes. It evaluates and ranks each gene pair A and B based on the stability of the ratio of their expression values A/B and B/A across samples. The more stable the ratio is, the more highly scored association is reported. The beauty of evaluating ratios is, once it has been calculated the two original gene expression values are irrelevant. This renders RA unaffected by the genes expression ranges, thus no specific degree of expression variation across the samples is required for detection. Thus RA can potentially analyze a single cellular state and extract biologically relevant gene pairs independently of the presence other cell conditions, distinguishing RA from standard analytical methods.

The introduction of RA and its potential is discussed in this thesis both as a straightforward explanation of the RA approach but also, and mainly, as a comparison to commonly used methods. As RA file under gene pair association-defining methods, the comparison will be made with these, more specifically with Pearson correlation and mutual information. These two classical approaches give a wide general inclusion of most current methods. RA will be explored mathematically in chapter 2, followed by its biological implications in chapter 3, and finally the future directions are discussed in chapter 4.

Common analytical methods

There is a multitude of methods for analyzing gene co-expression. The measure of association most often applied is based on, or a derivate of, correlation coefficient, squared correlation coefficient (R^2) or mutual information. They have been the basis for the commonly used clustering techniques (*17*), network motif and inference algorithms ((*38, 42, 45*) and references therein). There is a current effort in constructing new alternative approaches in defining gene expression relatedness and how to measure it. Even if newcomers have been presented, there is often a following debate of their originality relative to standard approaches and the degree of new biological information they actually extract (*45, 46*). The exclusiveness of current approaches in producing unique results, is even more so under debate as a new study indicates that the correlation coefficient and mutual information generally generate the same findings in numerous types of analysis (*47*).

There is a wide range of existing methods for analyzing and interpreting large amounts of gene expression data. The classic aim is to simplify the originally immense quantities of information by extracting the most prominent expression profiles and the cohorts of genes that constitutes

them (*17*). In doing so, biological process and mechanisms currently active in the samples become more apprehensible. These include studies examining time series, for example, cell cycle progression (*31*) and cell differentiation (*48*), between two cellular conditions be it tissue samples (*49*), cell lines (*50, 51*), or disease and control samples (*52*). The gene groupings from the generated expression profiles can further serve as a first step in of identifying the biological role of genes with currently no known function.

Here follows a short description of the most commonly used approaches. Their strengths are discussed through the kind of biological inferences that have been drawn from applying them. The following described methods can be divided into two categories:

1. Gene pair association defining
2. Gene group clustering

The first group defines what a gene pair association is and give the degree each gene pair conforms to that criteria. This group includes Euclidean distance, Pearson and Spearman correlation, regression models, and mutual information. The second type of methods uses those association scores and precipitate out the simplified expression profiles, which are presumed to describe the biological processes currently effective in the data. Examples described here are hierarchical clustering and k-mean clustering and principal component analysis. Before describing these, one commonly applied approach, average fold-change, will be discussed. While not exactly conforming to the above-mentioned categories as no pairwise evaluation of the gene expressions is executed, it is regularly used to precipitate out differentially expressed genes and thus is worth addressing.

Average fold-change analysis

This approach aims to identify individual genes that are differentially expressed in two or more cell conditions. The strategy entails to calculating the fold-change of the expression levels per gene between two cellular states. When more than one biological replicate is available, the expression mean or median is used for each state. Genes are considered differentially expressed if the fold-change exceeds the pre-set cut-off. As a result candidates are identified as possible differentiators between the cell states. Identifiers of a perturbed cell state were, for example, reported in a human melanoma cell line by determining genes differentially expressed in the cell line before and after introduction of a normal human chromosome 6 (*16*). The authors successfully identified several previously reported cancer related genes, among them, the key mediator of tumor suppression by p53, WAF1 (P21), validating the results as relevant to cancer biology. In one of the first cDNA microarray studies, Heller et al. both confirmed the imperative role played by TNF as an early key player in the course of rheumatoid arthritis and identified four new gene candidates, HME, IL-3, ICE, and Groα not previously known to actively participate in the inflammation process (*19*). A more recent study analyzed the differentially expressed genes comparing erythropoietic, granulopoietic, and megakaryopoietic cells further improving our understanding of hematopoietic differentiation (*50*). Differentially expressed genes were observed in cellular processes such as cell motility, immune system development and cell signaling, which are expected and of interest during the maturation of these cell types. The average fold-change approach is powerful in its simplicity, both regarding its application and in understanding what is tested. The simplicity comes though with some costs that under certain circumstances can lead to complications. These are addressed and discussed in the *Discussion* below.

Correlation methods

There are several types of correlation approaches; here Pearson and Spearman will be discussed, as they are preferred in biological studies. In general, correlation can be seen as the divergence from two variables being independent of each other. In the case of Pearson correlation the dependence is required to be of a linear form while as for Spearman correlation even non-linear relationship are accepted. Another difference is the nonparametric nature of the Spearman correlation, which renders it useful when no prior knowledge about the data population is obtainable or preferred.

Pearson correlation $\rho_{X,Y}$ is calculated by

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where *cov(X,Y)* is the covariance of the two variables *X* and *Y* and σ_X σ_Y is the product of the two variables' standard deviations, σ. Covariance is the measure of the degree two random variables fluctuate together. The better the high values of the first variable correspond to the high values of the second variable, and the same with the smaller values; the two variables' behaviors are comparable, giving a positive covariance. A negative covariance comes from the larger values of one of variables being comparable with the smaller values of the other variable and vice versa.

Spearman correlation assesses the dependence between the two variables by using a monotonic function (*53*). A function is termed monotonic if, when the two variables' data

points are ordered and the given order is preserved, so for all x and y that $x \leq y$ ones has $f(x) \leq f(y)$, and f preserves the order (in this case a monotonically increasing, increasing or non-decreasing function). Simply put, if all the values for x are ordered and all the values for y are ordered, the order, in reference to the data points, is the same. The difference between an ascending monotonic function (Spearman) and calculating the covariance (Pearson) is that, in the case of an increasing function, the increase/decrease going along the ordered data points does not need to be proportional. The absence of the requirement of proportionality in the Spearman correlation renders it accepting non-linear relationships.

Comparing the Pearson and Spearman correlation, the latter has a better built-in capacity to handle outliers. Spearman does not report to the same extent, false positives produced by a few large outliers, which can drive Pearson correlation to report a high false dependence. By the same capacity, Spearman detects true dependence even if a few outliers would distort the Pearson correlation reporting a low dependence. One important thing to note for both these correlation methods is that a pair of independent (non-linear for Pearson and non-monotonic for Spearman) variables has correlation 0, but the contrary is not always the case: zero correlation does not automatically entail independence.

Both Pearson and Spearman correlation are commonly used in biological studies to determine gene pair expression relationship strength, which often is followed by clustering. The biological information extracted by such studies is plenty. For example, in a study from Alizadeh et al. they identify two molecularly distinct forms of *diffuse large B-cell lymphoma* (DLBCL) (*25*). By analyzing global similarities in gene expression patterns among 96 normal and malignant lymphocyte samples, they found expression 'signatures', which conformed to characteristics of

either germinal center B-cells or *in vitro* activation of peripheral blood B cells. While investigating seven alveolar rhabdomyosarcoma (ARMS) all exhibiting the PAX3-FKHR fusion gene, Khan et al. found 37 genes forming a gene expression 'signature' of this type of cancer (*54*). Furthermore, the Pearson correlation managed to cluster (hierarchical clustering) all seven ARMS samples as a single cluster among other non-ARMS tumor samples. In an analysis of bone marrow plasma cells from patients with multiple myeloma and healthy subjects, hierarchical clustering of the Pearson correlation matrix produced a dendrogram with two major branches: healthy subjects and disease-affected, where the latter was divided further into four subgroups (*55*). In a study from Bittner et al. they discern a subgroup (by hierarchical clustering) of malignant neoplasm samples having an aberrant gene expression profile enrichment with genes, differentially expressed in invasive melanomas involved in primitive tubular networks *in vitro* (*44*). Regarding Spearman correlation there are studies showing its capability of capturing similar biological information. For example, a study in rats, where gene expression data from spinal cord development was compared to spinal cord injury, the gene pair association across tissue condition was studied (*56*). Fifteen percent of the genes were found in at least one pair that was correlated in both conditions, which the authors concluded strengthened the hypothesis that those gene pair relationships are of biological interest for tissue repair mechanisms.

Euclidean distance

Besides correlation as a similarity measure, Euclidean distance is commonly applied in biology to assess gene pair's association strength. Euclidean distance is as the name suggests, the distance between two points, x and y, that 'a bird would fly', simply x-y. For higher dimensions, for n subjects, the Euclidean distance $d(x,y)$ is equal to

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + .. + (x_n - y_n)^2}$$

. Euclidean distance is often used in biological studies as the measure for expression similarity between genes. An example is the study from Keuschnigg et al. (*50*). They studied plasticity of blood- and lymphatic endothelial cells to discover that the two commonly used cell lines HMEC-1 and TIME do not represent the phenotype of microvascular blood vascular as previously assumed. Instead, they are hybrid cells with characteristics from both vascular and lymphatic cells. Euclidean distance is often used to determine how similar a diverse group of samples are. For example, Luo et al. could for the first time show a distinct widespread differential gene expression pattern comparing benign and malignant growth of the prostate gland by performing Euclidean distance analysis of both conditions (*57*). One of the first studies of batch gene expression analysis using RT-PCR of 112 genes in rat cervical spinal cord showed that the expression pattern for neurotransmitter receptors depends more on the receptor ligand class than the gene sequence homology (*40*).

Mutual information

Information theory, first described by Shannon in 1948 (*58*), quantifies the amount of information in a random variable by calculating its entropy. In the field of gene expression analysis, the derivate, *mutual information,* is more commonly used, which gives a general measurement of the dependence between two variables. In other words, the mutual information for a gene pair measures how much information one gene's expression level gives about the other gene's expression level. Non-randomly associated gene pairs have a high mutual information score. Butte et al. showed that by calculating the mutual information

between every possible gene pair combination in a data set from *S. cerevisiae*, they extract 22 relevant gene expression clusters (*59*). The largest cluster included genes of the large and small ribosomal subunits and translation initiation factors. In another elegant study, information theory was used to calculate the association strength between all possible gene pairs in a data set consisting of a panel of different B cell phenotypes, ranging from normal, transformed to experimentally perturbed cells related to the germinal center (*60*). After additional analytical steps, gene clusters emerged with interesting hub characteristics. In other words, the results indicated that there are a few genes that interact with most other genes under these cell conditions, with MYC being the largest hub of them all.

Clustering

When clustering gene expression data, pre-calculated gene pair associations (variable distances) are sorted and grouped. Thus the interpretation of the association itself is not defined, calculated or altered, rather, clustering attempts to display the intrinsic relativeness among the variables by sorting the variables according to the set criteria. The two most commonly used clustering approaches in biology is hierarchical and K-mean clustering.

Hierarchical clustering groups the genes by iteratively merging the closest pair of genes, expression wise, until all genes are included (*61*). The choice of gene pair to merge is determined by the gene pairwise distances, which can be reported as complete linkage, average linkage, and single linkage. Hierarchical clustering has proven to successfully extract gene clusters in which the gene members share similar roles in the cell. For example, in a study of *S. cerevisiae* one of the clusters contained 126 genes, which were strongly down-regulated in response to stress and co-varied in the cell cycle (*17*). This cluster included previously known

genes from both the ribosomal complex and genes involved in translation, among others. Core processes vital to a functional cell repeatedly show up as clusters, another example is the clustering observed by Whitfield et al. during cell cycle progression, which included DNA replication, chromosome segregation, and cell adhesion (*31*).

The second frequently applied clustering method is K-mean clustering. This method takes a pre-set number of clusters, given by the user, and tries to assign each data point to one of them. First the assignments are random but for every iteration thereafter, the overall fitness of the previous assignments is calculated and based on that adjustments of the new cluster rosters are generated, until a steady solution has been reached or the maximum number of iterations wanted by the user has been reached. An example of how k-mean clustering can provide insight into previously unknown genes' function is the study of *Plasmodium falciparum* (malaria producing parasite) by Le Roch et al. (*62*). By clustering expression profiles from human and mosquito stages of the malaria parasite's life cycle, they ascribed possible functionality to the parasite's genes.

Principal Component Analysis

Principal component analysis (PCA) strives to simplify the description of data containing large numbers of variables by finding orthogonal vectors, principal components (PCs), extracting the most important information from the observations (*63*). The principal components are calculated by eigenvalue decomposition of the correlation matrix. The first component exhibits the largest possible variance, inertia, and thus describes the strongest observation in the data. The remaining components are calculated such that they are orthogonal to the preceding vectors and explain the maximum portion of the remaining inertia. The proportion of variance each component shows dictates its importance. The PCA results can be summarized in two

parts: the *factor scores* and the *loading scores*. The factor scores denote how strongly each PC describes an observation, gene expression. The loading score gives the influence each data point, gene, had on the PC. The characteristics of each component are determined by examining which genes load on it. For example, if the first component is heavily enriched in genes regulating cell cycle processes, the likely conclusion is that the strongest gene expression pattern in the data stems from cell cycle activities. Once the PCs are described, their factor scores for each gene describe how much each one of them influences the gene's observed expression pattern. Furthermore, unknown genes' role in the cell can be studied by observing which PCs load strongly on them.

The approach is appealing as it compresses large data sets into a smaller new set of variables that captures the strongest gene expression trends in the sample group. In a study re-analyzing gene expression data from wild type mice and heat-shock transcription factor 1 (HSF1) mutant mice, Jonnalagadda et al. reports that the first PC, capturing 42.12% of the inertia, responds to up-regulated genes during the stress response (*64*). The second PC (24.75% of the inertia) described the down-regulated genes, giving two broad general expression profiles occurring during the first 8 hours after the initial heat shock. In another study, Strakova et al. show that there is a general agreement of the expression profiles between the larger PCs of mRNA and protein levels during germination in *S. coelicolor* (*65*). Principal components of gene expression data can also be used as input to a second analysis, where large number of unique genes would simple not be manageable. Khan et al. calculate PCs describing the largest expression patterns seen in cancer samples compared to controls. These PCs are then input to an artificial neural network that extracts, in this case, 96 genes, which can successfully diagnose the existence of cancer in test samples (*66*).

It should be added that PCA is closely related to *factor analysis* (FA) with the difference of FA being based on regression modeling of the components, in this case called factors, which produce an additional error term per factor.

Conclusion

Classical gene co-expression analyses have been shown to generate highly similar results in biological settings. Furthermore, they require a certain degree of expression variation across samples, traditionally meet by including multiple cellular conditions in the analysis, for successful detecting of gene pair associations. This leaves gene pairs exhibiting little or no expression variation undetected and potentially part of the cellular processes underrepresented. The RA approach developed in this thesis aims to capture this group of gene pairs additionally to the gene pairs already being detected. RA measures the gene pair association based on the stability of the ratio between the two genes' expressions. Thus RA is conceptually different from standard methods both in its definition of a gene pair association but also in that it can be applied to a single cellular state.

References

1. C. Genomes Project et al., An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (Nov 1, 2012).
2. C. F. Lehner, P. H. O'Farrell, Expression and function of Drosophila cyclin A during embryonic cell cycle progression. *Cell* **56**, 957 (Mar 24, 1989).
3. C. F. Lehner, P. H. O'Farrell, The roles of Drosophila cyclins A and B in mitotic control. *Cell* **61**, 535 (May 4, 1990).
4. M. Murphy *et al.*, Delayed early embryonic lethality following disruption of the murine cyclin A2 gene. *Nature Genetics* **15**, 83 (Jan, 1997).
5. I. Kalaszczynska *et al.*, Cyclin A is redundant in fibroblasts but essential in hematopoietic and embryonic stem cells. *Cell* **138**, 352 (Jul 23, 2009).
6. M. I. Arnone, E. H. Davidson, The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851 (May, 1997).
7. G. L. Miklos, G. M. Rubin, The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**, 521 (Aug 23, 1996).
8. A. C. Gavin *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141 (Jan 10, 2002).
9. R. P. Perry, Balanced production of ribosomal proteins. *Gene* **401**, 1 (Oct 15, 2007).
10. C. Gorenstein, J. R. Warner, Coordinate regulation of the synthesis of eukaryotic ribosomal proteins. *Proceedings of the National Academy of Sciences of the United States of America* **73**, 1547 (May, 1976).
11. A. Rosenberg, L. Sinai, Y. Smith, S. Ben-Yehuda, Dynamic expression of the translational machinery during Bacillus subtilis life cycle at a single cell level. *PloS one* **7**, e41921 (2012).
12. Z. Shajani, M. T. Sykes, J. R. Williamson, Assembly of bacterial ribosomes. *Annual review of biochemistry* **80**, 501 (2011).
13. M. Nomura, R. Gourse, G. Baughman, Regulation of the synthesis of ribosomes and ribosomal components. *Annual review of biochemistry* **53**, 75 (1984).
14. L. Xiao, A. Grove, Coordination of Ribosomal Protein and Ribosomal RNA Gene Expression in Response to TOR Signaling. *Current genomics* **10**, 198 (May, 2009).
15. M. Laplante, D. M. Sabatini, mTOR signaling at a glance. *Journal of cell science* **122**, 3589 (Oct 15, 2009).
16. J. DeRisi *et al.*, Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature genetics* **14**, 457 (Dec, 1996).
17. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863 (Dec 8, 1998).
18. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467 (Oct 20, 1995).
19. R. A. Heller *et al.*, Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2150 (Mar 18, 1997).
20. G. K. Geiss *et al.*, Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology* **26**, 317 (Mar, 2008).
21. J. Marti, D. Piquemal, L. Manchon, T. Commes, [Transcriptomes for serial analysis of gene expression]. *Journal de la Societe de biologie* **196**, 303 (2002).

22. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621 (Jul, 2008).
23. T. Kayo, D. B. Allison, R. Weindruch, T. A. Prolla, Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5093 (Apr 24, 2001).
24. C. K. Lee, D. B. Allison, J. Brand, R. Weindruch, T. A. Prolla, Transcriptional profiles associated with aging and middle age-onset caloric restriction in mouse hearts. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14988 (Nov 12, 2002).
25. A. A. Alizadeh *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503 (Feb 3, 2000).
26. A. Thakur, A. Bollig, J. Wu, D. J. Liao, Gene expression profiles in primary pancreatic tumors and metastatic lesions of Ela-c-myc transgenic mice. *Molecular cancer* **7**, 11 (2008).
27. F. Luo, J. Liu, J. Li, Discovering conditional co-regulated protein complexes by integrating diverse data sources. *BMC systems biology* **4 Suppl 2**, S4 (2010).
28. A. Rawat, G. J. Seifert, Y. Deng, Novel implementation of conditional co-regulation by graph theory to derive co-expressed genes from microarray data. *BMC bioinformatics* **9 Suppl 9**, S7 (2008).
29. B. Andreopoulos, C. Winter, D. Labudde, M. Schroeder, Triangle network motifs predict complexes by complementing high-error interactomes with structural information. *BMC bioinformatics* **10**, 196 (2009).
30. P. T. Spellman *et al.*, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* **9**, 3273 (Dec, 1998).
31. M. L. Whitfield *et al.*, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977 (Jun, 2002).
32. W. C. Reinhold *et al.*, Identification of a predominant co-regulation among kinetochore genes, prospective regulatory elements, and association with genomic instability. *PloS one* **6**, e25991 (2011).
33. C. van Waveren, C. T. Moraes, Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. *BMC genomics* **9**, 18 (2008).
34. W. H. Mager, R. J. Planta, Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate. *Molecular and cellular biochemistry* **104**, 181 (May 29-Jun 12, 1991).
35. J. M. Bland, D. G. Altman, Correlation in restricted ranges of data. *Bmj* **342**, d556 (2011).
36. R. Steuer, J. Kurths, C. O. Daub, J. Weise, J. Selbig, The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18 Suppl 2**, S231 (2002).
37. C. O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC bioinformatics* **5**, 118 (Aug 31, 2004).
38. D. K. Slonim, From patterns to pathways: gene expression data analysis comes of age. *Nature genetics* **32 Suppl**, 502 (Dec, 2002).
39. T. R. Hughes *et al.*, Functional discovery via a compendium of expression profiles. *Cell* **102**, 109 (Jul 7, 2000).

40. G. S. Michaels *et al.*, Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 42 (1998).
41. F. D. Gibbons, F. P. Roth, Judging the quality of gene expression-based clustering methods using gene annotation. *Genome research* **12**, 1574 (Oct, 2002).
42. I. Priness, O. Maimon, I. Ben-Gal, Evaluation of gene-expression clustering via mutual information distance measure. *BMC bioinformatics* **8**, 111 (2007).
43. T. Kessler, H. Hache, C. Wierling, Integrative analysis of cancer-related signaling pathways. *Frontiers in physiology* **4**, 124 (2013).
44. M. Bittner *et al.*, Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536 (Aug 3, 2000).
45. D. N. Reshef *et al.*, Detecting novel associations in large data sets. *Science* **334**, 1518 (Dec 16, 2011).
46. J. B. Kinney, G. S. Atwal, Equitability, mutual information, and the maximal information coefficient. *arXiv.org*, (2013).
47. L. Song, P. Langfelder, S. Horvath, Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* **13**, 328 (2012).
48. S. Yung *et al.*, Large-scale transcriptional profiling and functional assays reveal important roles for Rho-GTPase signalling and SCL during haematopoietic differentiation of human embryonic stem cells. *Human molecular genetics* **20**, 4932 (Dec 15, 2011).
49. W. Xie *et al.*, Tissue-specific transcriptome profiling of Plutella xylostella third instar larval midgut. *International journal of biological sciences* **8**, 1142 (2012).
50. J. Keuschnigg *et al.*, Plasticity of blood- and lymphatic endothelial cells and marker identification. *PloS one* **8**, e74293 (2013).
51. P. Liu *et al.*, Transcriptome profiling and sequencing of differentiated human hematopoietic stem cells reveal lineage-specific expression and alternative splicing of genes. *Physiological genomics* **43**, 1117 (Oct 20, 2011).
52. O. M. Sessions *et al.*, Host cell transcriptome profile during wild-type and attenuated dengue virus infection. *PLoS neglected tropical diseases* **7**, e2107 (2013).
53. M. Hazewinkel, 'Monotone function', Encyclopedia of Mathematics. *Springer*, (2001).
54. J. Khan *et al.*, Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer research* **58**, 5009 (Nov 15, 1998).
55. F. Zhan *et al.*, Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood* **99**, 1745 (Mar 1, 2002).
56. M. Kotlyar, S. Fuhrman, A. Ableson, R. Somogyi, Spearman correlation identifies statistically significant gene expression clusters in spinal cord development and injury. *Neurochemical research* **27**, 1133 (Oct, 2002).
57. J. Luo *et al.*, Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer research* **61**, 4683 (Jun 15, 2001).
58. C. Shannon, A mathematical theory of communication. *Bell System Technical Journal* **27**, 623 (1948).
59. A. J. Butte, I. S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418 (2000).
60. K. Basso *et al.*, Reverse engineering of regulatory networks in human B cells. *Nature genetics* **37**, 382 (Apr, 2005).

61. T. T. Hastie, R., 14.3.12 Hierarchical clustering: The Elements of Statistical Learning (2nd ed.). *New York: Springer*, 520 (2009).
62. K. G. Le Roch *et al.*, Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503 (Sep 12, 2003).
63. H. a. W. Abdi, L.J, Principal component analysis. *WIREs Comp Stat* **2**, 433 (2010).
64. S. Jonnalagadda, R. Srinivasan, Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. *BMC bioinformatics* **9**, 267 (2008).
65. E. Strakova, J. Bobek, A. Zikova, J. Vohradsky, Global Features of Gene Expression on the Proteome and Transcriptome Levels in S. coelicolor during Germination. *PloS one* **8**, e72842 (2013).
66. J. Khan *et al.*, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine* **7**, 673 (Jun, 2001).

Chapter II

MATHEMATICAL DISCUSSION OF THE RATIOMETRIC METHOD

Abstract

The Ratiometric method (RA) is based on measuring the variation of the ratio between two genes' expressions. When compared to current methods such as Pearson correlation (PE), Spearman correlation (SP) and Mutual information (MI), it is shown that narrowing expression ranges does not impair RA. This particular nature of ratiometricity is first demonstrated by a residual analysis of ratiometrically generated data and data produced by a linear relationship. Furthermore, by measuring the ratio of two expressions, the expression values themselves are removed from the RA assessment with following beneficial consequences. Outliers are treated after their degree of deviation from the average ratio and their influence on the score does not depend on the size of their expression value. In conclusion, RA has built-in characteristics of treating all biological samples as equally important when evaluating a gene pair's association strength to a higher degree than PE and MI.

Introduction

A common basis for global gene expression analyses today is measure of association, including the most prominent methods such as the squared Pearson correlation coefficient (R^2), often used in hierarchical clustering (*1*), and Mutual Information (with mutual information statistic I). As the effort of designing new approaches to discover gene pair expression relatedness continues, the propositions made are many times questioned as of

how distinctive they are from present methods (*2, 3*). At the same time the uniqueness, between already established methods, especially the correlation coefficient and mutual information, is debated regarding their comparable results in biological studies (*4*). The sensation that the main body of analytical tools available gives, in essence the same results, promotes the search for alternative methods. For even if the current strategies have shown to be efficient in identifying substantial numbers of important classes of gene relationships (*5-15*), there is a possibility that they fail to identify all types of relationships present in biological samples. One such type of undetected gene pair relationships relates to the often unappreciated effect variability has on the measure of association (*16*), rendering gene pairs reported as not co-expressed when they are.

This advocates for a search for alternative methods of gene expression analysis that is to a smaller extent focused on measure of association, and rather centers on less explored statistical approach in the biological field, such as variability. Here, the new approach RA ranks each gene pair (genes *A* and *B*) based on the stability of the ratios of their expression values (A/B and B/A) across the sample data. From a biology viewpoint, the objective is to appreciate each gene pair relationship based on their two genes' relative stability, regardless of their absolute expression levels' dispersion rates across the samples.

The Ratiometric method

RA evaluates the dependency between two gene expressions, gene *A* and *B*, based on how stable the expression ratios, $\frac{A}{B}$ and $\frac{B}{A}$, are across the samples, such that a stable ratio, of $\frac{A}{B}$, for 4 hypothetical samples is

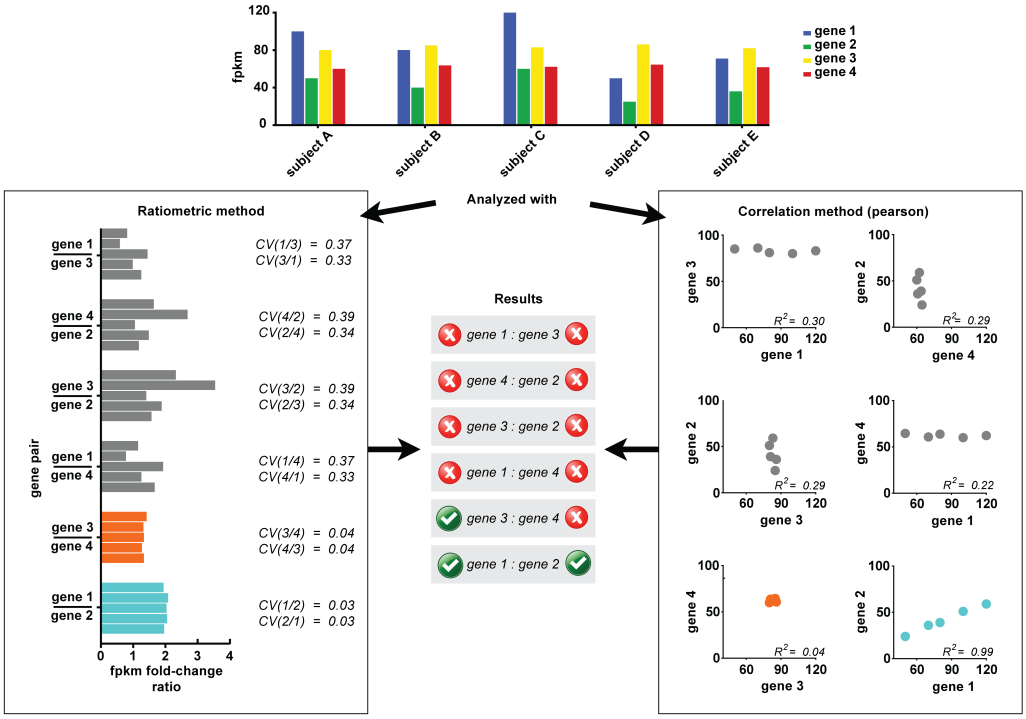
$$\frac{a_1}{b_1} \approx \frac{a_2}{b_2} \approx \frac{a_3}{b_3} \approx \frac{a_4}{b_4}$$

where a_i is the expression value for gene *A* and b_i is the expression value for gene *B* in sample *i*. The stability of ratio, *r*, is measured by calculating the coefficient of variation (CV), the standard deviation divided by the mean, of the gene pair's ratios

$$CV(r) = \frac{stdev(r_1,\ r_2,\ r_3, r_4)}{mean(r_1,\ r_2,\ r_3, r_4)}$$

where r_i is the ratio $\frac{a_i}{b_i}$ in sample *i*. The lower the CV(*r*) the less the ratio fluctuates among the samples and thus the higher prediction power does that particular gene pair combination has for predicting one of the genes in the pair using the other gene. One important point to be made here is the relationship between a ratiometric regime and a linear regime (high squared Pearson correlation), PE. At a first glance the proposed approach could appear as an analysis of a linear regime. A gene pair displaying a linear regime can exhibit a low CV(*r*) and thus be of high prediction power. But the opposite is not by default true: a gene pair having a low CV(*r*) does not automatically have to a high Pearson correlation, in other words follow a linear regime. To schematically demonstrate how RA differs from PE, a hypothetical case with 5 subjects (*A-E*) and 4 genes (*1-4*) is drawn out (Figure 1A).

A


fpkm
120
80
40
0
subject A
subject B
subject C
subject D
subject E
gene 1
gene 2
gene 3
gene 4
Analyzed with
Ratiometric method
gene pair
gene 1
gene 3
CV(1/3) = 0.37
CV(3/1) = 0.33
gene 4
gene 2
CV(4/2) = 0.39
CV(2/4) = 0.34
gene 3
gene 2
CV(3/2) = 0.39
CV(2/3) = 0.34
gene 1
gene 4
CV(1/4) = 0.37
CV(4/1) = 0.33
gene 3
gene 4
CV(3/4) = 0.04
CV(4/3) = 0.04
gene 1
gene 2
CV(1/2) = 0.03
CV(2/1) = 0.03
0 1 2 3 4
fpkm fold-change ratio
Results
gene 1 : gene 3
gene 4 : gene 2
gene 3 : gene 2
gene 1 : gene 4
gene 3 : gene 4
gene 1 : gene 2
Correlation method (pearson)
gene 3
gene 1
R2 = 0.30
gene 2
gene 4
R2 = 0.29
gene 2
gene 3
R2 = 0.29
gene 4
gene 1
R2 = 0.22
gene 4
gene 3
R2 = 0.04
gene 2
gene 1
R2 = 0.99
100
50
0
60 90 120

B

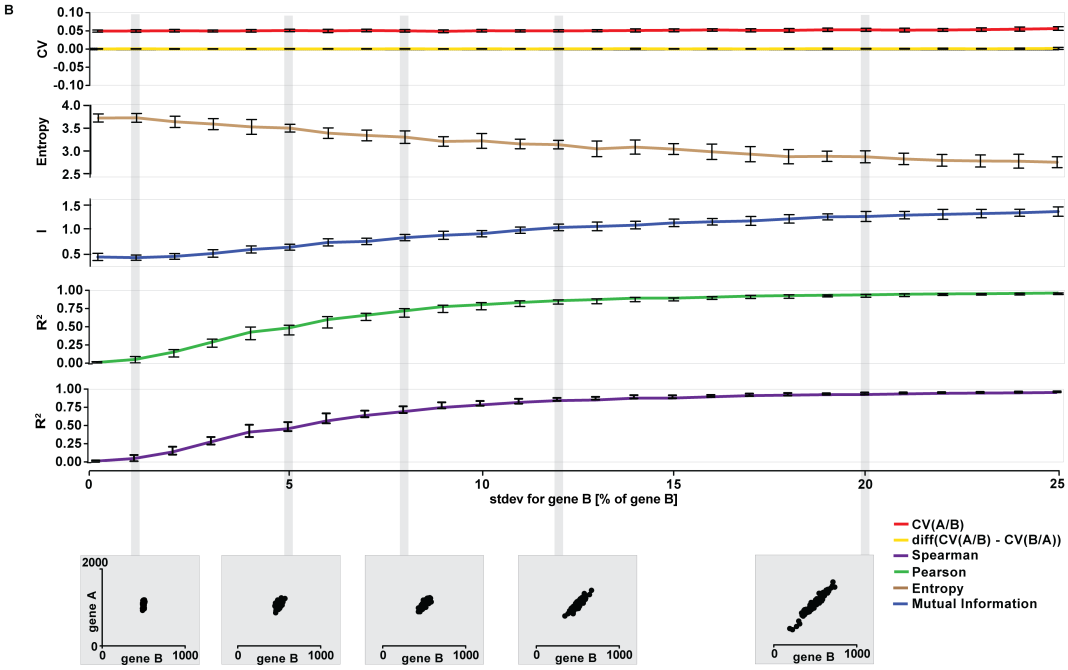

CV
0.10
0.05
0.00
-0.05
-0.10
Entropy
4.0
3.5
3.0
2.5
I
1.5
1.0
0.5
R²
1.00
0.75
0.50
0.25
0.00
0 5 10 15 20 25
stdev for gene B [% of gene B]
CV(A/B)
diff(CV(A/B) - CV(B/A))
Spearman
Pearson
Entropy
Mutual Information
gene A
gene B
2000
0
1000

Figure 1: (Previous page) **Schematic of the differences between RA and correlation metrics.** A) For five samples (A-E), the expression levels of genes 1-4 are measured (top graph). The box on the right shows the results of an analysis of expression relationships using a Pearson correlation. Only gene pair 1:2 is identified as a significant interaction, with $r^2 = 0.99$. In contrast, the ratiometric method (box on the left) identifies both gene pairs 1:2 and 3:4 as significant. The PE method does not capture the second relationship (gene pair 3:4) as the FPKM ranges of the two genes are too narrow for a regression line to be stable. On the other hand, the RA model assesses only the FPKM fold-change across samples, is thus much less sensitive to narrow FPKM ranges, and identifies both gene pairs. B) Shown is simulated expression data for two hypothetical genes, A and B. The expression levels of *A* were generated from those of *B* using the following equation: $a_i = 2b_i + u_i$. For each dataset, the expression range of B was varied by increasing *CV(B)* from 0 to 25% of the mean expression level of $B(\mu_B, = 500)$. The expression level of *B* is thus normally distributed following $B \sim N(500, \%B)$. For each value of *CV(B)*, 10 datasets with 100 samples each were generated. The Pearson and Spearman r^2, the entropy, mutual information as well as the *CV(A/B)*, *CV(B/A)* and $\Delta_{CV} = |CV(A/B) - CV(B/A)| = 0.02$ values were calculated for each dataset and the mean and standard error are shown. Note that the gene pair association does not change along the *x* axis and the expression level of gene B can be used to predict the expression level of gene A equally well in all runs. As the expression range for B narrows, the Pearson and Spearman r^2-values decrease, along with the mutual information index. In contrast, the ratiometric CV is constant and the relationship between the expression levels of the two genes is always recovered.

The gene pair combinations, plotted as the expression levels of gene A versus gene B, are analyzed with PE (right box) and the coefficient of determination, R^2, reveals only one gene pair, *1:2*, is highly correlated. On the other hand, when the pairwise combinations are evaluated with RA, both gene pairs *1:2* and 3:4 pass as observing a RA expression pattern and having stable ratios. Thus the second gene pair, *3:4*, is not detected as mathematically dependent according to PE but yet it has a high prediction power, very similar to the *1:2* pair, which is also picked up and reported by RA.

The Ratiometric association is defined as a function of the relative dispersion of the two ratios $\frac{A}{B}$ and $\frac{B}{A}$, calculated by their *CV:s*. For further explanation and motivation of this choice see appendix 1 generated by Dr. Kenneth McCue. For evaluating a gene pair's

expression pattern, two criteria are used. The first is the stringency of fit to a ratiometric relationship given by Δ_{CV} as

$$\Delta_{CV} = |CV(r) - CV(r^{-1})|$$

, where r is $\frac{A}{B}$ and r^{-1} is $\frac{B}{A}$. The better the fit the closer to 0 the Δ_{CV} is. As Δ_{CV} is explicitly modeling the variability in the expression values, the possible convoluting affects caused by variability seen in traditional methods, is averted. This distinction is most prominent when variability is lowered caused by the range of gene expression being restricted. This is further investigated and analyzed in *Expression range effect*. Once the selected gene pairs passing the stringency of fit, Δ_{CV}, are selected, the second criterion is applied: the strength of the association. This is based on the *CV(A/B)* and *CV(B/A)*, where the lower value the more stable ratio. The two CV:s are given as CV, as they are approximately the same when the Δ_{C} is set close to 0. For the remaining of this thesis, the application of RA corresponds to the two-step combination of the Δ_{CV} and CV.

Results

Expression range effect

To demonstrate the sensitivity to the expression range exhibited by RA compared to PE, MI, and SP, a simulation was performed where the expression range of one hypothetical

gene was varied and each method's reported score was determined (Figure 1B). For two hypothetical genes, *A* and *B*, their expression levels were generated as followed:

$$a_i = 2b_i + u_i$$

where $u_i \sim N(0,50)$. The expression range of *B* was stepwise altered by increasing the standard deviation of *B*, from 0 to 25 percent of the mean expression level of *B*. This gives, when $E[B] = 500$, the expression level of *B* normally distributed following $B \sim N(500, s^2)$, with *s* being the standard deviation. Ten iterations per run with $n = 100$. It is important to notice that throughout the different expression ranges of gene *B*, the prediction power was maintained. Thus for any of the runs either of the two genes can be used to predict the other one with the same precision. This was captured by RA as its scoring is unchanged across all runs, and thus the underlying relationship is always recovered. The same is not true for PE, MI or SP, where their scoring is changing as the expression range of gene *B* is decreasing. All reports the gene pair association as weakening the narrower the expression range becomes.

In summary, the expression level range has a pivotal effect on the reported association strength by PE, MI and SP but not RA. This is due to the fact that as the expression range narrows the two expression profiles deviate less and less from the statistical definition of mathematical independence. As it is the deviation from the latter that PE, MI, and SP report they by default score the gene pair relationships worse and worse. On the other hand, RA is not derived from mathematical independence but is based on prediction strength, which is not altered by the narrowing of the expression range in this case, and thus RA scores all of the gene pair relationships as equally strong.

Average fold-change study

Analyzing gene expression profiles by calculating average fold-change between two conditions, e.g., controls and disease, is one of the most simple and straightforward approaches for discovering differential expression patterns between two cellular states. As it analyzes each gene as a separate entity, it cannot extract information about gene pair interaction and thus neither generate general co-expression profiles characteristic of the samples. Therefore, it is not easily comparable to more elaborate methods aimed towards co-expression analysis such as PE, MI, and SP. But as it is still commonly practiced, a brief demonstration of its limitations, here compared to RA but which conclusion can be extrapolated to co-expression methods in general, is given.

The schematic case, figure 2A, depicts expression levels for one gene in 4 control subjects and 4 altered subjects (this example, disease affected). The average and standard variation for the two conditions are the same, Figure 2B, which would indicates this is a gene that has an unaltered expression between the two conditions and in accordance with the average fold-change analysis would be of little interest.

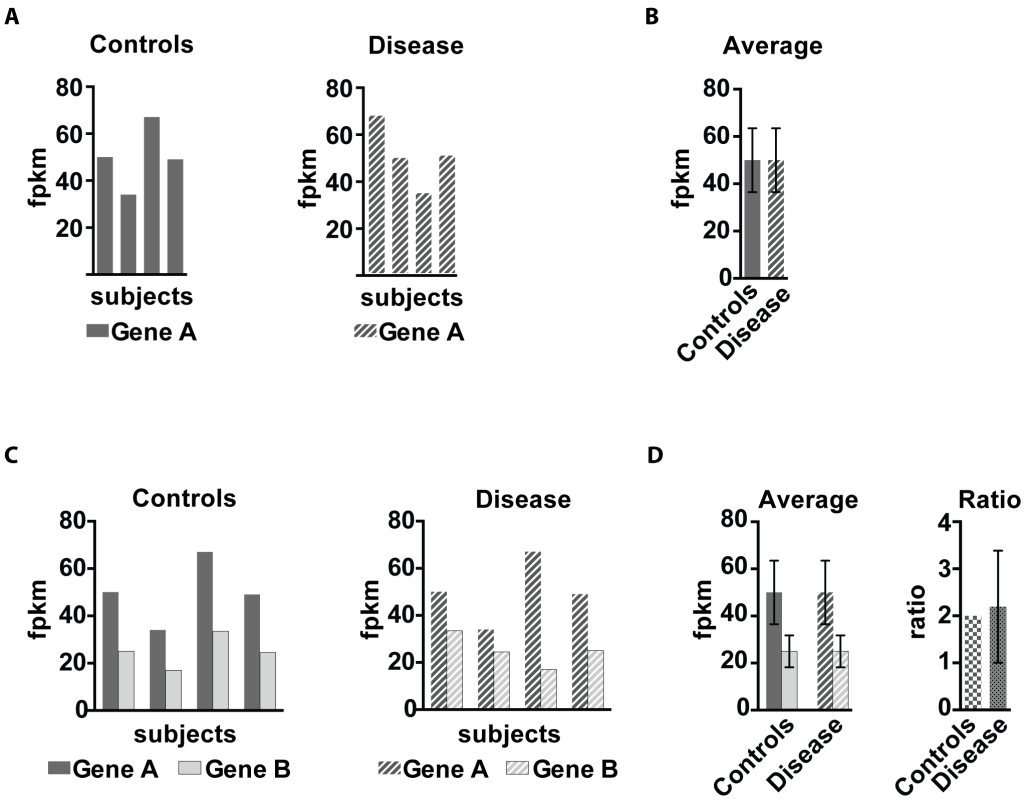


Figure 2: **Average fold-change versus the ratiometric approach. A**. The expression levels [fpkm] for 4 control and disease affected subjects for gene A. Their average and variation of expression is the same, **B**. In C, the same gene A expression values as in **A**, with a second gene B added. **D,** the average and variation show no difference between controls and disease-affected subjects. But the expression ratio between gene A and B is stable in the control group but not the disease-affected group. Thus the RA method detects the difference between control and disease state concerning the two genes while the average fold-change method does not.

But if the first gene is observed in reference to a second gene, Figure 2C, it becomes clear that in the control subjects the two genes are maintained in a perfect ratio of 2:1, Figure 2D. While in the disease group the ratio is no longer stable and fluctuates between the subjects. The loss of the stable ratio could either be the cause, or the side effect of the disease state, thus being of interest for further investigation. In conclusion, when average fold-change is used these two genes would not be reported as altered between the two conditions. While a

RA analysis would detect the loss of the stable ratio and report the gene pair as possible candidate of interest.

As average fold-change is such a crude method, not emphasizing co-expression patterns, it is of no interest for the following method comparisons in this thesis.

A comparison of the ratiometric equation and a linear relationship

As a first step in elaborating on how the mathematical and statistical foundation of RA is distinct from current methods, its definition of a pairwise expression relationship is compared to what could be seen as its 'closest' relative: a linear relationship. Thus how the general ratiometric equation

$$\frac{A}{B} = r$$

where A and B are expression values of two genes (A and B), with ratio r, differs from a linear equation

$$A = B * c$$

where A and B are again expression values of the two genes (A and B), with the linear constant c. For this discussion the intercept is set to 0. It is not immediately obvious why RA would produce different results compared to methods based on linear equations as the ratiometric equation can be given as the linear equation when $r = c$ by

$$\frac{A}{B} = r$$

$$A = r * B = c * B$$

. But when the relationship is measured in more than one replicate, the fit of the applied equation can be observed. To this end the error term is introduced, defining the individual

fluctuation in each sample from the norm of the entire sample group. Such that the ratiometric equation becomes

$$\frac{a_i}{b_i} = \bar{r} + u_i$$

for sample *i*, where a_i *is* gene A's expression, b_i *is* gene B's expression, $\bar{r}$ is the average ratio in the sample group, and u_i is the error term or the individual deviation from the norm. In the same fashion the linear equation becomes

$$a_i = b_i * c + u_i$$

for sample *i*, where a_i *is* gene A's expression, b_i *is* gene B's expression, c is the constant (slope), and u_i is the error term. Now if the transformation of the ratiometric equation into the linear is intended, the discrepancy is revealed:

$$\frac{a_i}{b_i} = \bar{r} + u_i$$

$$a_i = (\bar{r} + u_i) * b_i$$

$$\neq$$

$$a_i = (b_i * c) + u_i$$

if $c = r$. Therefore, there should be a difference between data sets generated by a ratiometric equation compared to a linear equation. To determine this, a simulation was performed where two data sets were generated and analyzed using a ratiometric and a linear approach separately, figure 3. The ratiometrically generated data set, (*RATIO*, left column) was produced with the ratiometric equation $A = (r + u_r) * B$, where $r = 2$, $B \sim N(50, 10)$, and $u_r \sim N(0, 0.2)$. The data set generated from a linear relationship (*LINEAR*, right column) followed $A = c * B + u_l$, where $c = 2$, $B \sim N(50, 10)$, and $u_l \sim N(0, 10)$. Both data sets consist of 10.000 data points. First, the fit to a linear relationship was tested by a residual analysis applying a linear regression on both data sets, figure 3A.

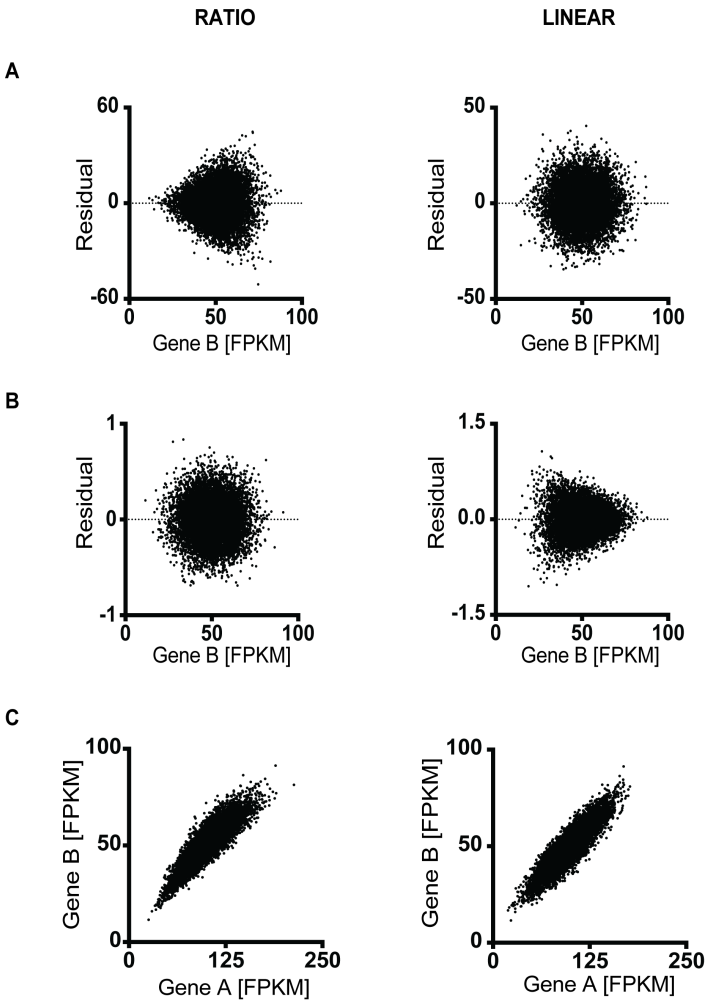


Figure 3: **Data simulated with a linear equation versus a ratiometric equation.** The data in the left column was generated with the ratiometric equation $A = (r + u_r) * B$, where $r = 2$, $B \sim N(50, 10)$, and error $u_r \sim N(0, 0.2)$.The data in the right was generated with the linear equation $A = c * B + u_l$, where $c = 2$, $B \sim N(50, 10)$, and error $u_l \sim N(0, 10)$. Both data sets have an N=10.000. **A**. The residuals plotted against B, calculated as the orthogonal residual using a linear regression. **B**. The residual ratio plotted against B, calculated as $residual_i = r_i - \bar{r}$. **C**. Gene A plotted against gene B, illustrating the narrowing of the data points at the lower end of the expression ranges in the RA data set.

When plotting the residuals against the expression value of gene B, the behavior of the error term can be observed, in other words, the noise in the data remaining once a linear approach

has been applied. If the applied equation completely describes the expression pattern, the noise should exhibit a normal random distribution, which is, not surprisingly, the case for the *LINEAR* data set. In the case of the *RATIO* data set, the noise takes the form of an increasing cone, thus displaying heteroskedasticity, implying that the linear equation does not capture all the information present in the data. Similar but reverse behavior was seen when the residuals of the individual ratios where plotted against gene *B*'s expression value, figure 3B. Here the noise from the *RATIO* data set, non-surprisingly, displays a normal random distribution, while the *LINEAR* data set exhibits a 'flipped' heteroskedasticity. For lower expression values of gene *B*, the ratio-residuals are greater. Thus the ratiometric approach cannot fully explain all the information in the *LINEAR* data set. These results suggest that there would be a distinction between the *RATIO* and *LINEAR* data sets when the two genes are plotted against each other for each set, figure 3C. This is confirmed as the *RATIO* graph displays a narrowing at the lower end of the expression ranges, forming a slight cone-shaped expression pattern. This is intuitive if the data points' positions are interpreted in the light of how their ratio is affected by their absolute values. If the ratio, $A/B=2$, is varying by 50%, its measured value could range from 1 to 3, thus if gene *B* has an expression of 100, the expression of gene *A* could be 100 or 300, and if gene *B* has an expression of 1, the expression of gene *A* can be 1 or 3. Therefore gene *A* can only vary in expression between 1-3 if gene *B* is 1 (lower end of the expression range), but if gene *B* has a higher expression value, in this example 100, gene *A*'s expression can vary between 100-300 and still be within the given ratio fluctuations. The consequence is a cone-shaped expression profile, not observed in the *LINEAR* data set.

To continue exploring how the magnitude of an expression value effects gene pair relationships strengths, following example was constructed using outliers, figure 4.

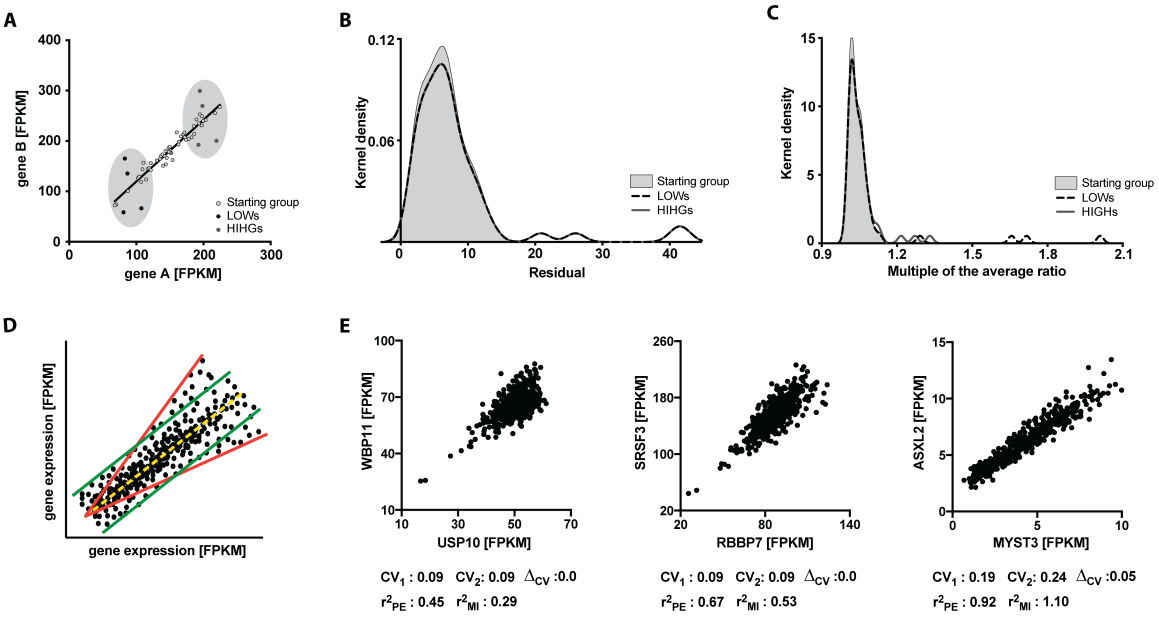


Figure 4: **The effect of outliers on reported relationship**. **A**. Starting with two genes, *A* and *B*, with the expression depicted by group 1 (light gray), with the linear relationship shown by the black line. The two additional 4 data points: group 2 (black) is in the lower end of range of the expression range for both genes and group 3 (dark gray) is in the higher end of the range. Note that the two additional groups exhibit the exact same internal position pattern of the 4 data points. The kernel densities of the orthogonal residuals (**B**) and the multiple of the average ratio (**C**) for the three groups: group 1, group 1 + 2, and group 1 + 3. **D**. Illustration of how the RA method contrasts in evaluating data points based on the position of the latter. **E**. Three real examples of gene pairs, where WBP11:USP10 exhibits a strong cone shape, ASXL2:MYST3 displays an even band across the entire expression range, and SRSF3:RBBP7 is in between the two types.

Starting with a larger data set that exhibits a linear behavior, two sets of outliers were applied: 4 highly expressed data points (HIGHs) and 4 lowly expressed data points (LOWs). The two outliers groups were constructed such that they have identical internal positions and distances among themselves and the regression line. They only differences was that the

LOWs are at the low end of the expression range of the starting data set and the HIGHs are at the top end of the expression range, figure 4A. From these, three data sets could be composed: 1) starting group, 2) starting group + HIGHs, and 3) starting group + LOWs. To begin with, the distributions of the residuals for the three data sets were examined, figure 4B. Markedly, the HIGHs and LOWs are completely overlapping, which is to be expected as they have the same orthogonal distance to the regression line. Then the outliers' ratios were determined and the distributions of the ratios for the three data sets were compared, figure 4C. There is a clear distinction between the ratios belonging to the HIGHs and the LOWs. The HIGHs are closer to the starting group's distribution curve then the LOWs, which are much further away. Thus more lowly expressed data points are seen as major outliers when they are positioned off the average ratio line, as their actual ratiometric value then is very far from the average ratio. Highly expressed data points can appear to the naked eye to be very much outside the 'mainstream data group' but their ratios are not altered to any larger extent and thus RA does not see them as major outliers. This is to say that in a regression the data points are equally treated along parallel lines (green lines) of the regression line (yellow dotted line), figure 4D. While in a RA approach, the data points along the sides of a cone (red lines) would be equally treated. Finally, just to demonstrate that such behaviors are observed in real gene expression data 3 gene pairs are shown, figure 4E. The cone-shaped expression pattern can be seen with WBP11:USP10, meanwhile ASXL2:MYST3 displays an even band across the entire expression range, and SRSF3:RBBP7 is in between the two types. Noteworthy is also the evaluations each of RA, PE, and MI reports for the 3 gene pairs. A strong cone shape has a strong RA score, while both PE and MI are giving such a gene pair a very low score. The 'even band'-expression shape produces high PE and MI scores while RA even rejects it completely as not conforming to the ratiometric definition.

Thus RA evaluates gene pair relationships differently than PE and MI, which can be observed in the shape of and distribution within the expression profiles.

Gene expression stability generating false positives

There is a general concern regarding the use of ratio stability for evaluating gene pair relationships, which is that two biologically completely unrelated genes invariantly expressed would generate a high RA-scoring per automatic. To test for such false positive reporting by RA, a simulation was performed where two genes, A and B, expressions were generated independently of each other according to a normal distribution $A \sim N(\mu,\sigma)$ and $B \sim N(\mu,\sigma)$, μ = [1, 2, 5, 10, 20, 50, 100, 500, 1000] and $\sigma = \mu\%$ where % = [0, 5, 10, 15, 20, 25], see Figure 5.

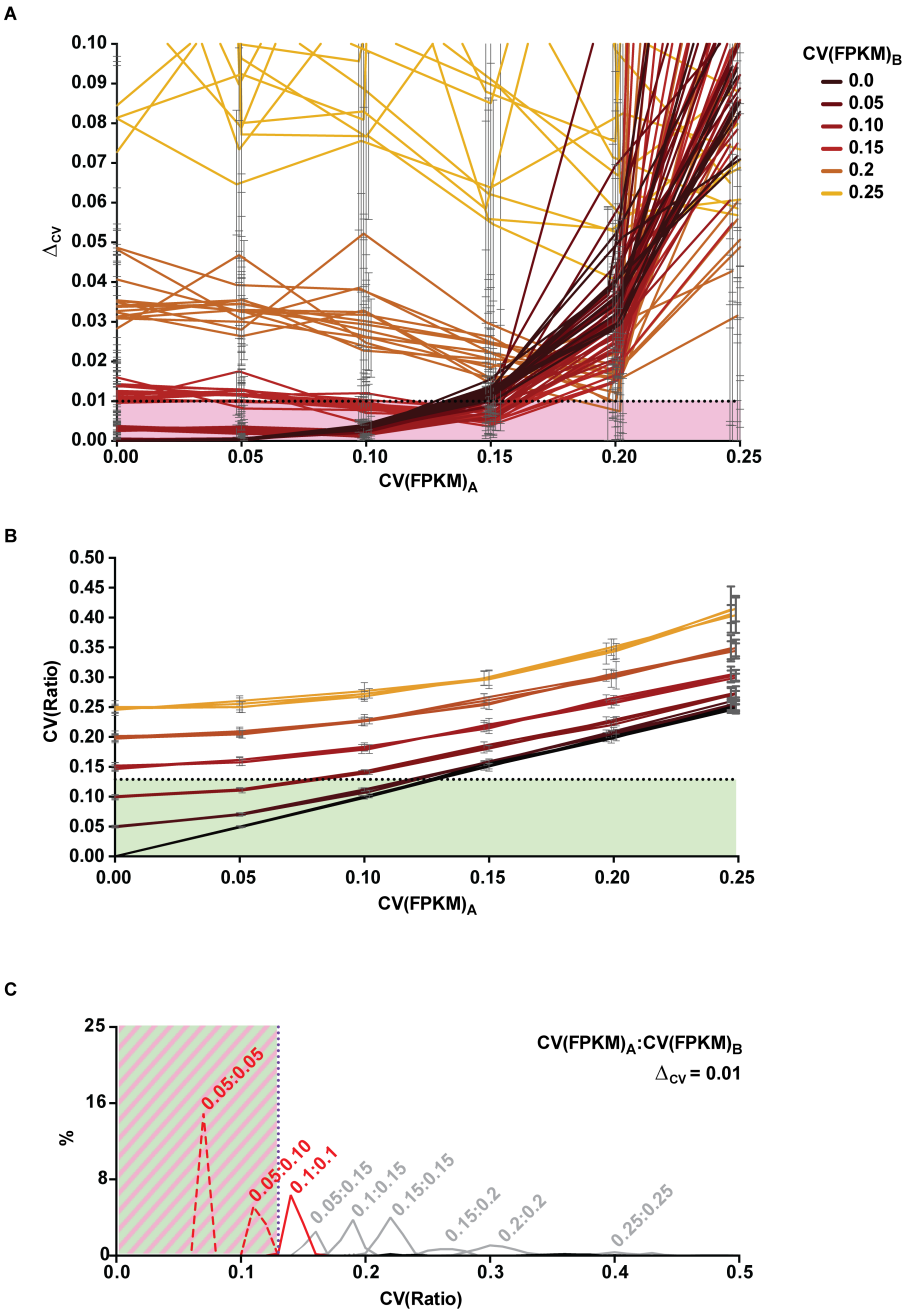


Figure 5: (Previous page) **How gene's expression stability influences scoring of ratio stability**. Two genes, *A* and *B*, expressions were generated independently of each other according to $A \sim N(\mu,\sigma)$ and $B \sim N(\mu,\sigma)$, μ = [1, 2, 5, 10, 20, 50, 100, 500, 1000] and $\sigma = \%\mu$ where % = [0, 5, 10, 15, 20, 25]. Each run included 10 iterations and each iteration had an n = 462 (the number was chosen to mirror the dataset size used in chapter 3). The runs are colored according to the $CV(FPKM)_B$. **A**. Δ_{CV} versus $CV(FPKM)_A$ when varying $CV(FPKM)_B$, demonstrating that Δ_{CV} both increases in value and fluctuation as the CV(FPKM) increases. Pink area indicates the Δ_{CV}-cut-off used in chapter 3. **B**. CV(Ratio) versus $CV(FPKM)_A$ when varying $CV(FPKM)_B$, demonstrating that CV(Ratio) increases as the CV(FPKM) increases. Green area indicates the CV(Ratio)-range plotted in the KEGG-

analysis in chapter 3. **C**. The percent of accepted gene pairs for the different CV(FPKM)-values tested (given over each peak as the $CV(FPKM)_A$:$CV(FPKM)_B$) at Δ_{CV}= 0.01. Plotting only curves demonstrating at least one peak ≥1%. The striped area indicates the chosen cut-offs used for the B-cell dataset in chapter 3. Red color indicate CV(FPKM) potentially in the risk zone of producing false positives. The dashed red lines are the subgroup of the latter exhibiting a CV(FPKM) not present in the B-cell dataset. The solid red line is the only scenario of CV(FPKM)s presenting a possible risk of introducing false positives in the B-cell dataset. Out of 2430 possible gene pairs with 0.1:0.1, only five were reported by RA, thus an error rate of 0.2%. The gray curves are outside the cut-off levels.

Each run included 10 iterations and each iteration had an n = 462 (the number was chosen so to mirror the dataset size used in chapter 3). As the results indicate the determining factor for reporting a false positive gene pair is the size of the expression variation of its two genes, and not the expression levels themselves, see Figure 6.

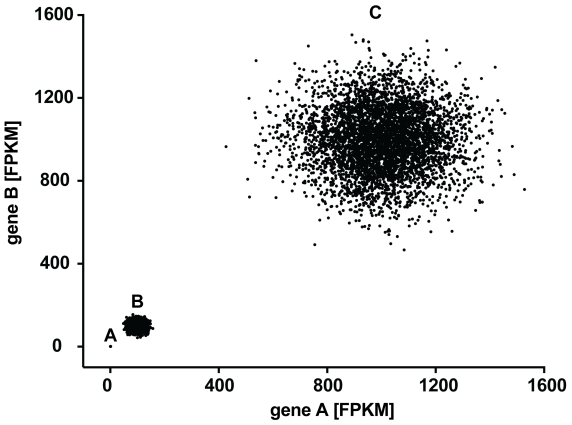


Figure 6: **Δ_{CV} relates to CV(FPKM)**. A, B, and C, have a mean expression level of 1, 100, and 1000, respectively. They all have CV(FPKM) = 0.15 for both their genes thus demonstrating the expression spread at different expression levels generating a $\Delta_{CV} < 0.01$.

As the variation increases the both the reported Δ_{CV}-value increases but also the range of Δ_{CV} reported. Thus by setting the Δ_{CV} to close to zero the false positive rate is lowered. Furthermore, the CV(Ratio) is similarly affected by the CV(FPKM) of the two genes, Figure 6B. The higher the CV(FPKM) the higher the CV(Ratio). Thus depending on the chosen Δ_{CV} and the stability of the genes' expression in a dataset, the false positive yield might vary.

For example, a scenario such as the B-cell dataset in chapter 3, the ~ 3000 top-ranked genes (cut-off at CV(FPKM) = 0.13) would possibly include false positives generated from gene pairs where both genes have CV(FPKM) $\lesssim$ 0.75 (green area). The average of false positives reported in this simulation, Figure 6C, display an extremely small false positive rate when the Δ_{CV} is set stringent and the CV(Ratio)-range includes the top-ranked genes. For example, for the B-cell data set in chapter 3, there is an estimated 0.2% false positive rate among the ~ 3000 top-ranked genes (again with a cut-off at CV(FPKM) = 0.13).

Discussion

The proposed approach, RA, to gene expression analysis is based on a simple mathematical expression of $A/B = ratio$ for the two genes A and B. It can at first appear as just a simple rewrite of the linear relationship of $A = B * ratio$ where *ratio* can also be termed c (the slope of a linear equation with the intercept of 0). If only a single value for gene A and gene B, respectively, are used, the ratiometric equation is ergo equivalent to the linear equation. But once there are multiple values for genes A and B, an error term is introduced depicting how well each pair of a_i and b_i fit to the relationship. Thus when applying the equations to empirical data it becomes clear that they are distinct and as a result conceivably could generate divergent results when evaluating gene pair relationships. By plotting the error terms, the noise, generated from implementing one of the equations, the extent to which the applied equation describes the data can be observed. An equation that fully describes the expression relationship in the data would produce noise, which exhibits a normal random distribution for both variables. This would be exhibited by a uniform spherical pattern, see figure 3A right graph, for example, called random noise. As the results depict a data set

generated by a linear equation will, when analyzed using a linear relationship, produce random noise. While when the data is created by a ratiometric equation, the noise generated from a linear analysis is no longer random but feature heteroskedasticity. The same reasoning holds true when the residuals of the individual ratios are plotted. The ratiometrically generated data set display a random noise pattern, while the noise from the linearly produced data set still exhibit some degree of dependence on gene *B*. In conclusion, a ratiometric equation is not fully interchangeable with a linear equation, and vice versa, when it comes to evaluating a gene pair relationship.

There is a multitude of approaches currently available for gene expression analysis, each one implementing their definition of how a gene pair relationship should be evaluated. There are two venues for discussing how RA differs from the currently available methods, mainly correlation and mutual information-based analyses.

The first venue focuses on the interpretation of dependence between two gene expressions. In contemporary methods, such as correlation and mutual information-based approaches, dependence has been translated as the degree two gene expressions deviate from a mathematical condition of probabilistic independence. The implication of such translation is worth contemplating, as the reverse is not automatically true: mathematical independence does not necessarily have to imply biological independence. Two genes that are both constantly expressed, are said to be mathematical independent, but their jointly calibrated expression could be imperative for the cell; they just do not fluctuate under the observed condition or the tested sample group.

The second venue involves measure of association versus variability, which here will be a summary of Dr. Kenneth McCue's work presented in the supplementary text and methods found in the accompanying article to this thesis. Association is the definition used by a method to define the pairwise expression relationship, while variability is the dispersion of expression values within the sample group. Methods, such as PE, include both components of association and variability, and therefor there is a risk of convoluting effect of the latter altering the degree of association reported. This has been discussed in the statistical community by Bland and Altman (*16*), by addressing how range restrictions in the expressed genes may lead to reduced correlation coefficient;

> *Correlation coefficients are a property of the variables and also the population in which they are measured. If we look at a restricted population, we should not conclude that there is little or no relation between the variables because the correlation coefficient is small.*

This is known as truncation and results from the statistical literature demonstrate how such alteration of the data can alter the correlation coefficient of the bivariate normal (*17*). While truncation insinuates an alteration of the data, causing the measured data points to not completely represent the actual population, and consequently the report correlation coefficient is 'modified', narrow expression ranges can arise naturally. Maybe the most straightforward condition is when the data consists of a rather homogeneous sample group. Such convoluting effects are not seen in RA, as it is not a measure of association. It analyzes relatedness by measuring directly and solely the variation in ratios rendering it insensitive to expression dispersion.

Every new method presents a challenge in understanding how it relates and complements established methods. RA is not a new approach to, per se, increase the detection success rate under the regime of using a correlation coefficient or mutual information. Even if there is a high likelihood that RA will overlap to some extent in its detection of gene pairs with PE and MI, it should not be perceived that RA functions under the same premises as PE and MI. For a proper application and appreciation of RA in future studies it is imperative to have an understanding of how the alternative statistical approach presented in RA affects the type of findings it generates from biological data. Thus it is important to have an awareness of how the mathematical difference of RA, compared to PE/MI, affects which biological questions can be answered and how they should be phrased to not misinterpret results or impede the full potential of RA. By analyzing the measured stability of a ratio, by calculating *CV(ratio)*, the absolute expression values of such a ratio have no longer any influence on the scoring of the gene pair relationship in question. Thus RA is not under same constraints as PE/MI, where the latter require a certain degree of variation of the absolute expression values across the samples to be able to detect the relationship. Philosophically this is an interesting aspect of RA, as there is nothing in biology that would indicate that all gene pair relationships of interest have to vary across samples. Consequently, if you then apply a method that has variation as a requirement for successful detection, you impose such an assumption. This is not always incorrect, as the investigation might concern gene pairs that remain in relationship despite expression fluctuations across samples. Then methods such as PE and MI are more than sufficient for such detection. But if the question of interest is to obtain an, as comprehensive as possible, roster over gene pair relationships present in a certain sample group, then it is of value to use a method that is not limited to detect only gene pairs with variation present in them.

Another interesting aspect of RA is the equal weight it gives each sample in the analysis regardless of their absolute expression values. When applying a linear relationship on expression data, it can be seen as testing for a constant fold-change (if the intercept is 0); a linear expression with a slope of 2 can also be perceived as there is a 2-fold-change between the two variables. Assuming that all the samples have expression values over experimental noise levels, each one of their measured fold-change should thus be equally important for determining if a fold-change is present in the sample group or not. This follows from the fact there are no indications that samples with higher expression values, per se, depict a truer form of the observed biological relationship than samples with very low expression values (keep in mind that all expression values are above noise-level). When applying methods such as PE to expression data, the samples will be given different weight to the final scores depending on their absolute expression values. Highly expressed samples are influencing the results to a larger extent than the lowest expressed samples. Thus, the use of PE produces gene pair relationships, which are by the scientist perceived, as universal in the sample group, but in reality might be solely present, to reported precision, in high-expressed samples. The lowest expressed samples could exhibit a much higher spread of fold-change-values among themselves than displayed by the highest expressed samples. But as PE gives less weight to the samples in the lower end of the expression range, this incoherency is not detected. As RA only analyzes the ratios (fold-changes) among the samples there is no bias towards neither highly nor lowly expressed samples, thus not giving either of them a disproportionally large influence on the resulting relationship score. In other words, those samples with lower expression values also need to have the ratio for the relationship to be reported by RA. In the end, the importance of how each method treats samples at varying

points in the expression range, goes back to what question the researcher is interested in. If there is a risk that the expression data originate from experimental set-ups where a lot of inherent noise and uncertainty were inevitable, then caution should be taken as RA might reject gene pair relationships that are biologically interesting but due to data quality are deemed not expression-related. On the other hand, if it is more important to be able to say that the gene pair relationships reported are present in all samples, regardless of absolute expression values, and with the same ratio/fold-change variation in the lower end of the expression range as in the higher end, then RA is suitable. The view of a gene pair relationship being present as a fold-change/ratio might not be very common but it has been studied before. Recent findings showed that there are biological decisions made based on fold-change detection of gene product concentrations and not their absolute concentrations (*18*).

Another aspect of RA worth reflecting upon is how a ratio or fold-change restricts the value of the intercept. This discussion revolves around the subgroup of accepted RA gene pair relationships where the expression pattern exhibits larger expression dispersions such that the data points are stretched out in a cone-shape manner, for example, see figure 4E left graph. In these cases an imaginary line can be drawn through the data points giving an intercept of 0. Note that by its nature RA only accepts expression patterns where the 'observed' intercept is 0. This is only observed when the expression pattern has a stretched out feature as expression patterns with a spherical shape has no 'observed' single unique intercept. There are implications of an intercept of 0 in a cellular setting pertaining to what type of expression relationships are captured. Theoretically, an intercept of 0 could be seen as if there is an amount of gene *A* there is always an amount of gene *B* present in the cell.

There are no cells, in the studied cellular condition, where only one of the two genes is observed. In comparison, an expression relationship where the intercept is, for example, 4 ($a = 4 + b * slope$), would theoretically imply there could hypothetically be cells where there is no RNA from gene *B* present but from gene *A*. The conclusion from this is, as before more in the aspect of how the different methods answer the questions asked by the researcher. RA will not detect gene pair expression patterns where the intercept is not 0, while PE would. In contrast, RA will only report gene pairs where both genes are required to be represented in RNA in the cells possibly indicating a more mutual co-expression.

Finally, the number of false positive reported by RA was estimated using a simulated model where expression data for two independent genes were generated according to normal distributions. From the simulation three findings could be seen:

1) The degree of expression variation, measured in CV(FPKM), determines the likelihood of a false positive reporting.

2) The more invariantly expressed genes are, the lower Δ_{CV} and CV(Ratio) they exhibit in the relationships.

3) Depending on the expression variation distribution of a dataset, it is possible to select a Δ_{CV}-level and a range of CV(Ratio) such that there is close to zero false positive reported.

The first finding dictates that the Δ_{CV} and CV(Ratio) are independent of the genes' individual expression-levels and are only correlated with the genes' CV(FPKM). Thus a highly expressed gene is equally likely to produce a false positive with another highly expressed gene if both of them exhibit extremely small fluctuations in their expressions across samples.

The second finding is that only highly invariantly expressed genes generate $\Delta_{CV} \approx 0$ and very low CV(Ratio). This implies that when RA is applied to a dataset the number of false positives reported can be estimated by determining the portion of the analyzed gene population exhibiting highly invariant expression, thus very low CV(FPKM)s. The larger the portion the more caution should be observed when drawing conclusions from the data analysis. The third and possibly the most interesting finding is that as the number of reported false positives depends on how stringently the Δ_{CV}-cut-off is set and what range of the CV(Ratio) is analyzed, these two parameters can be calibrated to lower the number of false positives reported to the minimum level possible for each dataset. Depending on the expression variation distribution in a dataset the cut-offs can be set such that the number of false positives reported is close zero. For example, from the simulation it is estimated that, for a dataset where the CV(FPKM)s are >0.10, setting the $\Delta_{CV} = 0.01$ and studying the gene and gene pairs reported within the CV(Ratio)-range of 0 to 0.13, would put the estimated portion of false positives being detected at ~0.015% (for the gene pairs). For any study using RA, it is recommended that the distribution of the expression variations is determined. Then by simulating a series of individually expressed genes with varying expression variation (lowest set to the lowest found in the dataset), the false positives can be estimated for varying Δ_{CV} and CV(Ratio)-ranges, which in turn can guide the researchers in the choice of cut-off-levels.

Conclusion

The RA method is not impaired by the effects of narrowing expression ranges that affect common methods, such as PE, SP, and MI, which for the latter leave gene pair relationships undetected. This is achieved by using the variability of the expression ratio between two

genes for evaluating gene expression pair relationships, conceptually different from traditional methods. The comprehension of RA compared to current methods, can both be discussed in the form of predictability versus that of probabilistic independence, and the degree of association versus variability. Regardless of which one is a more 'correct' venue in pursue of the exact uniqueness of RA, the concluding point is that RA is an analytical approach that statistically differentiates itself from extant methods which, in turn, calls for a continued study of how such an approach carry itself when analyzing biological data.

Methods

Expression range effect

The simulation includes expression data for two hypothetical genes, *A* and *B*. For each run the expression level of *B* is normally distributed following $B \sim N(500, \%B)$, with an $n = 100$, and there is 26 runs for which the %B is ranging from 0% to 25%. The expression levels of *A* are generated from those of *B* exercising the following equation: $a_i = 2b_i + u_i$. Each run is reported as the mean score with standard errors showing the spread from 100 iterations for the Pearson and Spearman r^2, the entropy, mutual information as well as the *CV(A/B)*, *CV(B/A)* and Δ_{CV}. The simulation was design such that the gene pair association, prediction strength, does not change along the *x* axis and thus the expression level of gene B can be used to predict the expression level of gene A equally well in all runs.

A comparison study between a ratiometric equation and a linear equation

The simulation was performed with two generated data sets. The first data set was generated using the ratiometric equation $A = (r + u_r) * B$, where $r = 2$, $B \sim N(50, 10)$, and error $u_r \sim N(0,$

0.2). The second data set was generated using the linear equation $A = c * B + u_i$, where $c = 2$, $B \sim N(50, 10)$, and error $u_i \sim N(0, 10)$. Both data sets have an N=10.000. The residuals were calculated using a linear regression from which the slope and intercept was used to determine the residual as:

$$residual_i = a_{predicted} - a_i$$

where $a_{predicted}$ is

$$a_{predicted} = intercept + slope * b_i$$

for $residual_i$ in sample i. The $ratio\ residual_i$ in sample i was calculated as residual

$$ratio\ residual_i = r_i - \tilde{r}$$

.

References

1. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863 (Dec 8, 1998).
2. J. B. Kinney, G. S. Atwal, Equitability, mutual information, and the maximal information coefficient. *arXiv.org*, (2013).
3. D. N. Reshef *et al.*, Detecting novel associations in large data sets. *Science* **334**, 1518 (Dec 16, 2011).
4. L. Song, P. Langfelder, S. Horvath, Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* **13**, 328 (2012).
5. T. Kayo, D. B. Allison, R. Weindruch, T. A. Prolla, Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5093 (Apr 24, 2001).
6. N. J. Severs, The cardiac muscle cell. *BioEssays : news and reviews in molecular, cellular and developmental biology* **22**, 188 (Feb, 2000).
7. A. A. Alizadeh *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503 (Feb 3, 2000).
8. F. Luo, J. Liu, J. Li, Discovering conditional co-regulated protein complexes by integrating diverse data sources. *BMC systems biology* **4 Suppl 2**, S4 (2010).
9. A. Rawat, G. J. Seifert, Y. Deng, Novel implementation of conditional co-regulation by graph theory to derive co-expressed genes from microarray data. *BMC bioinformatics* **9 Suppl 9**, S7 (2008).
10. B. Andreopoulos, C. Winter, D. Labudde, M. Schroeder, Triangle network motifs predict complexes by complementing high-error interactomes with structural information. *BMC bioinformatics* **10**, 196 (2009).
11. P. T. Spellman *et al.*, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* **9**, 3273 (Dec, 1998).
12. M. L. Whitfield *et al.*, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977 (Jun, 2002).
13. W. C. Reinhold *et al.*, Identification of a predominant co-regulation among kinetochore genes, prospective regulatory elements, and association with genomic instability. *PloS one* **6**, e25991 (2011).
14. C. van Waveren, C. T. Moraes, Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. *BMC genomics* **9**, 18 (2008).
15. W. H. Mager, R. J. Planta, Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate. *Molecular and cellular biochemistry* **104**, 181 (May 29-Jun 12, 1991).

16. J. M. Bland, D. G. Altman, Correlation in restricted ranges of data. *Bmj* **342**, d556 (2011).
17. M. Regier, M. A. Hamdan, Correlation in a Bivariate Normal Distribution with Truncation in Both Variables. *Australian Journal of Statistics* **13**, 77 (1971).
18. L. Goentoro, M. W. Kirschner, Evidence that fold-change, and not absolute level, of beta-catenin dictates Wnt signaling. *Molecular cell* **36**, 872 (Dec 11, 2009).

Chapter III

BIOLOGICAL APPLICATION

Abstract

The ratiometric approach, RA, was tested and compared to Pearson correlation, PE, and mutual information, MI, using a large homogenous sample group consisting of B-cell lymphoblastoid lines from 462 human individuals (*1*). The RA gene and gene pair rankings were substantially different from both PE and MI, which in turn were largely the same. The RA top-ranked genes were enriched in core biological processes, such as the ribosome, spliceosome, and mRNA transport. PE and MI exhibited a degree of enrichment but it both involved fewer pathways and yielded often a weaker enrichment. A correlation was found between pathways being strongly detected by RA and weakly detected by PE and MI, and the gene expression dispersion range among the pathways' gene populations. The narrower the expression dispersion was, the weaker the detection of PE and MI. This finding is in agreement with the simulation results (Chapter 2), where PE and MI scoring were shown to depend on the expression dispersion range of the simulated gene in such a manner that the more invariant expression the lower rating was reported.

Introduction

The underlying concept of the RA approach to transcriptome analysis is rather simple from a biological perspective. To evaluate the strength of a gene pair's relationship based on its ratiometric qualities is similar to studying fold-change in its most simple form. The

fluctuations of a ratio between two genes' expressions from one sample to the next, is the same as the fold-change of those expression values from one sample to the next. Giving that, the smaller the difference in fold-change is between samples, the more stable that ratio is. The additional criteria that the RA method applies, limits the gene pairs of interest to only those that exhibit a ratiometric expression profile (see chapter 2 for details and justification). Thus, RA focuses on the gene pair relationships that have ratiometric characteristics instead of studying all possible gene pairs, as in the more classical form of analyzing fold-changes in expression data (*2-4*).

The notion of the cellular mechanisms utilizing fold-changes, or ratios, as their *modus operandi* for regulating appropriate gene expression levels has been studied recently, for example, in the Wnt-signaling pathway (*5*). It was shown the expression levels of the target genes, thus the output of the Wnt-signaling pathway, can be robust to expression level fluctuations. Specifically, they studied the expression levels of the pathway regulator, β-catenin. As long as the fold-change of the β-catenin expression levels remained unchanged before and after Wnt stimulation, the output also remained the same. Thus the fold-change, ratio, of β-catenin is pivotal and not its absolute expression levels. The possible benefits of operating under a fold-change/ratio regime are, both the resilience against cellular fluctuations in gene expression and quickly varying noise in their activity levels (*6*). It could be seen as a buffering mechanism against fluctuations arising from stochasticity, genetics, and/or environmental variation.

RA is well suited for detecting the occurrences of ratiometrically conformed gene pair relationships as it solely evaluates the ratios between two genes and thus without taking into account their absolute expression levels and the expression ranges. This enables a fair comparison of ratio stability between gene pairs, not biased by one pair's expression ranges compared to the others (see chapter 2 for further details). Thus, RA complements current methods since it identifies gene correlations independently of variability of a gene's expression across a sample cohort. A gene exhibiting the same absolute expression value in all samples can still be involved in interactions with other genes that require its expression level to be jointly calibrated with the interaction partners. Thus it would be preferable if these genes would not be automatically excluded merely on the based that they do not fluctuate enough between samples.

To evaluate how RA performs, it was compared to the two most common general approaches for global gene expression analysis: Pearson correlation, PE, and mutual information, MI. By comparing with these two methods, the vast majority of analytical approaches currently available are considered, as most of them are derivatives of PE or MI.

RA was applied to a publically available dataset containing RNA-seq data from 462 individual human lymphoblastoid cell lines (*1*). This particular dataset was chosen as it has a large sample number, *N*, thus decreasing the risk of results becoming skewed unintentionally by non-random sampling of the biological population. Additionally, it is a rather homogenous sample group compared to, for example, fresh tissue samples. The

fact that these samples are all from the same type of cells and have gone through immortalization process to become a cell line and then are grown in as identical conditions as possible, further removes confounding variation commonly present in biological data.

Results

Data and gene selection

The downloaded raw sequencing reads were processed by uniform expression qualification using eXpress (*7*) (See Methods for further details). A gene selection was perform such that only genes with FPKM values ≥ 1 in at least 95% of samples were included (Figure 1), in total 9752 genes.

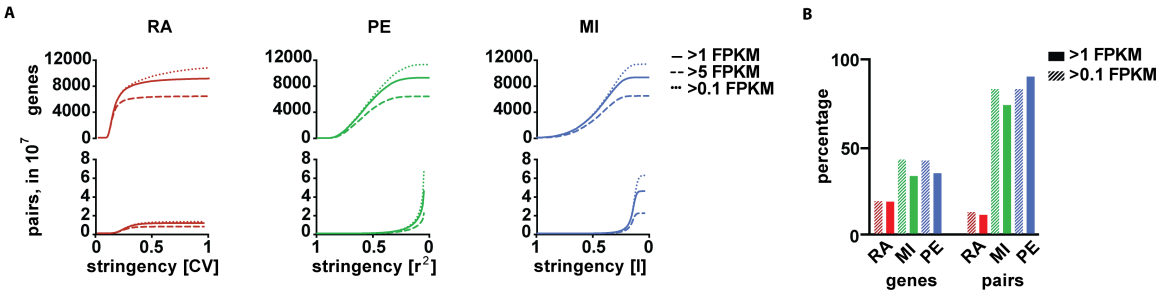


Figure 1: **Gene set selection**. The gene set analyzed includes all genes with a FPKM ≥ 1 for at least 95% of the samples. To explore the effect of FPKM cut-off used, we analyzed a cut off at 0.1 and 5 FPKM. A) Each set was analyzed according to the RA, PE and MI reporting how many gene pairs and genes found at each stringency level, RA [0-1], PE [1-0] and MI [1-0]. B) An example of the difference between the gene sets for the 3 methods. Here is shown, the difference (in percentage) of gene pairs and genes included for gene set ≥ 1 FPKM and ≥ 0.1 FPKM, at the same stringency level that gives ~5000 incorporated genes for gene set ≥ 5 FPKM. The RA method shows a lower degree of variation across the three gene sets than PE and MI.

To determine the effect the FPKM cut-off has on the analysis, the data selection is performed with two additional cut-offs: one stricter (FPKM ≥ 5) and one more lenient (FPKM ≥ 0.1) cut-off. The distributions of scored gene pairs across the coefficient ranges, RA: CV[0-1], PE: r^2[1-0], and MI: I[2-0], for each method are reported to determine to what degree they vary for the three FPKM cut-offs. What is of highest interest is how much the pair distribution changes at the most stringent end of the coefficient range as it is generally those gene pairs that are chosen for further empirical testing in studies. As indicated in Figure 1A, the FPKM cut-off has little effect on the gene pair distribution along the RA coefficient range (CV [0-0.5]) while the distributions for PE and MI display a higher fluctuation when the FPKM cut-off is changed. This is further demonstrated in Figure 1B, where the differences of gene pairs and genes for the gene sets created by ≥ 1 FPKM and ≥ 0.1 FPKM are shown, at the same stringency cut-off for which ~5000 genes are detected when the FPKM cut-off ≥ 5 FPKM. In detail, at the given stringency cut-off PE includes 225,093 gene pairs as being correlated from the most stringent FPKM cut-off (≥ 5 FPKM), PE reports 410,168 gene pairs for the most lenient FPKM cut-off (≥ 0.1 FPKM), an increase of 82%. The same 82% increase is seen for MI with the same parameter settings. On the other hand, RA only reports an increase of 12%, same parameter settings. Thus RA is more robust against inclusion and/or exclusion of genes with low expression.

Summarizing, RA shows a lower degree of variation across the three gene sets than PE and MI. Thus the conclusions drawn in this report will not heavily change if a different

FPKM cut-off is chosen and the small changes observed will be less pronounced than the changes occurring in the results from PE and MI analyses.

Δ_{CV} *cut-off analysis*

In theory a gene pair conforming to a ratiometric expression profile has a $\Delta_{CV} = 0$ (see chapter 2 for further details). To allow some wiggle room for empirical noise, the Δ_{CV} in this biological study is set to 0.01. Thus any gene pair having a Δ_{CV} lower or equal to 0.01 is included. To determine the effect the chosen Δ_{CV} cut-off has on the reported gene pair distribution, the Δ_{CV} cut-off is varied, 0.005, 0.01, 0.02, 0.03, and the results compared, Figure 2.

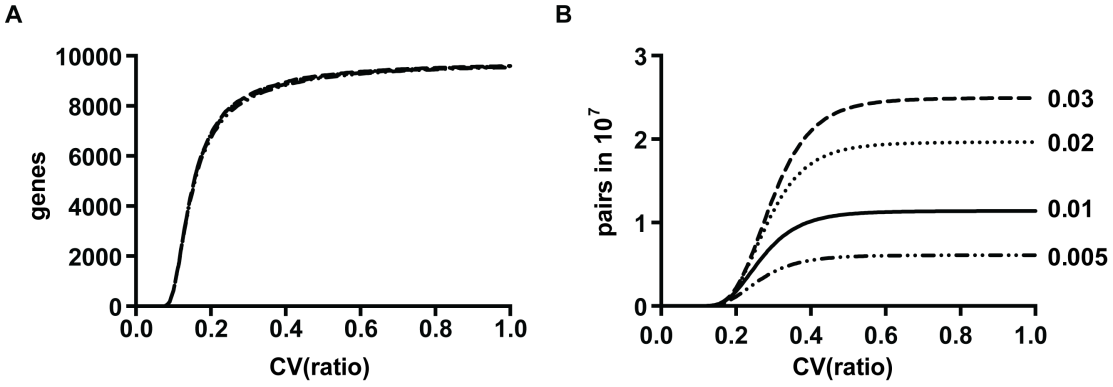


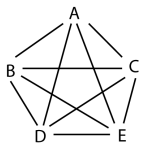
Figure 2: **Connectivity trends in RA graphs at different Δ_{CV} cut-offs.** Analyzing the effect of Δ_{CV} cut-off has on the inclusion rate per stringency level by plotting A) the number of unique genes and B) the number of gene pairs found at every CV level ranging from 0 to 1. The Δ_{CV} has no noticeable effect on the rate of gene inclusion to the cluster. The effect can be seen in the number of gene pairs found, though first after ~50% of the genes are already included. As the interest of most gene expression analysis focus on the top ranked interactions the effect of Δ_{CV} is minimal.

The differential effect between the different Δ_{CV} cut-offs is close to none regarding the number of unique genes included along the stringency range CV [0-1], Figure 2A. There

is a noticeable higher number of gene pairs reported along the CV-range, Figure 2B. The difference in inclusion rate for the gene pairs is though first observed when ~50% of the genes are already included. Thus the Δ_{CV} cut-off has little effect on the top-ranked genes and as it is those that most often are of interest for further investigation, the used Δ_{CV} cut-off is not pivotal within the tested range of Δ_{CV} [0.005-0.03] for this data set.

Gene pair relationship landscape across the stringency ranges

To determine the extent to which RA detects a different subgroup of gene pair relationships compared to PE and MI, the number and identity of the pairwise relationships detected at several stringency levels were established for each method. Each ranking list was divided into a series of 100 steps in accordance to the method's stringency range, thus PE r^2[1→0], MI I[2→0], and RA CV[0→1]. For each stringency step or level, i, the interactions can be drawn out as a graph, G_i, with the genes as nodes and the gene pair relationships as undirected edges. The set of edges is thus $E(G_i)$ and the number of edges $|E(G_i)|$. All genes with no interaction at a given stringency level, vertex degree $d_g(v) = 0$, are excluded and the remaining number of nodes are denoted as $|V(G_i)|$. See figure 3 for schematics.



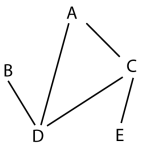
Five hypothetical genes (A-E) drawn out as nodes in a graph, G, with every possible gene pair relationship drawn out as edges between the nodes (genes).

The strength of association for each edge (gene pair) is evaluated by each method

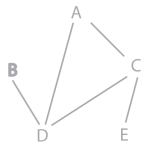
Ratiometric approach (RA)
CV [0 -> 1], 100 steps

Pearson correlation (PE)
r^2 [1 -> 0], 100 steps

Mutual information (MI)
I [2 -> 0], 100 steps

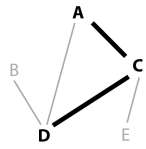


Pre-processing
The edges (gene pair combinations) that has ΔCV > 0.01 is discarded as they don't conform to the ratiometric definition

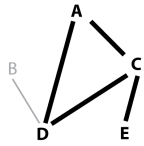


For each CV-level

nodes (genes) and edges (gene pairs) are counted. Starting with the most stringent CV = 0.

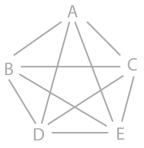


nodes (genes) and edges (gene pairs) are counted. Next step, stringency CV = 0.5.



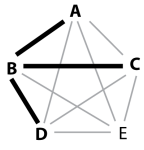
nodes (genes) and edges (gene pairs) are counted. Last step is least stringent CV = 1.

stringnecy [CV]	# nodes (genes)	# edges (gene pairs)
0	0	0
.	.	.
.	.	.
.	.	.
0.5	3	2
.	.	.
.	.	.
.	.	.
1.0	4	4



For each r^2-level

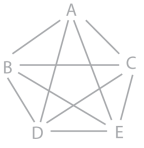
nodes (genes) and edges (gene pairs) are counted. Starting with the most stringent r^2 = 1.



nodes (genes) and edges (gene pairs) are counted. Next step, stringency r^2= 0.5.

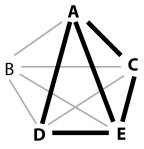
nodes (genes) and edges (gene pairs) are counted. Last step is least stringent r^2 = 0.

stringnecy [r^2]	# nodes (genes)	# edges (gene pairs)
1	0	0
.	.	.
.	.	.
.	.	.
0.5	4	3
.	.	.
.	.	.
.	.	.
0	5	10

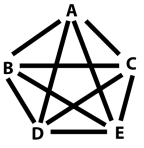


For each I-level

nodes (genes) and edges (gene pairs) are counted. Starting with the most stringent I = 2.



nodes (genes) and edges (gene pairs) are counted. Next step, stringency I = 1.



nodes (genes) and edges (gene pairs) are counted. Last step is least stringent I = 0.

stringnecy [I]	# nodes (genes)	# edges (gene pairs)
2	0	0
.	.	.
.	.	.
.	.	.
1	4	5
.	.	.
.	.	.
.	.	.
0	5	10

Figure 3: (Previous page) **Schematics of how the methods process a hypothetical dataset**. Using five fictive genes, which create 10 possible gene pair combinations, we can construct a graph where the genes are nodes and the pairwise gene interactions are edges. The processing of the expression data by Pearson correlation, PE, and mutual information, MI, is relatively straightforward. By stepwise decreasing the stringency level, for PE r^2 and MI I, from most stringent, $r^2 = 0$ and I=2, to least stringent, $r^2 = 1$ and I=0, and at each step count the number of nodes and edges (thicker lines), a continuous reporting of how the graph increases is produced. Note that for both PE and MI all genes and gene pair combinations will be counted at the lowest stringency level. For the ratiometric method, RA, there is a pre-processing step to select for gene pairs that exhibit a ratiometric expression pattern. Only these relationships are included in the analysis. Then, the same approach is applied, where the stringency is measured in CV. The most stringent level is CV = 0 and the least stringent level used is CV = 1. Notice that due to the pre-processing step and the fact that CV goes to infinite, there is a possibility that not all genes and all gene pairs are included in the analysis.

First, the rates of increase of $|V(G_i)|$, number of unique genes involved in a relationship, as a function of CV (RA), R^2 (PE), and I (MI) were calculated, Figure 4A solid lines.

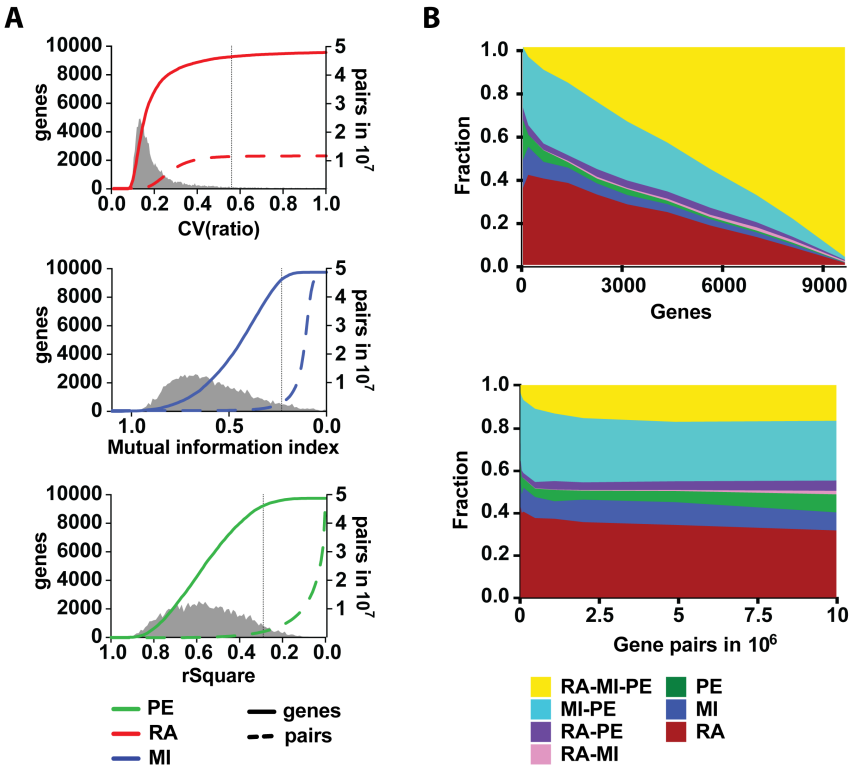


Figure 4: (Previous page) **Connectivity trends in graphs built by each model**. Analyzing the increase of connectivity coverage as the stringency level (coefficient cut-off) decreases. A and B, for every cut-off, the accepted portion of the graph is plotted as the percentage of all genes (solid line, left y-axis) and the number of gene pair connections (dashed line, right y-axis). All three graphs include >98% of the genes at the lowest plotted stringency level, though the dynamics of how the graphs grow are different, separating the RA method from PE and MI. Likewise in gene pair relationship inclusion, with the additional distinction that as the ratiometric model does only consider connections that exhibit a ratiometric profile it will have a smaller total number of possible connections (24% of total gene pairs possible). The dashed black vertical line marks the stringency level at which 95% of the genes are included, $|V(G)| = 9244$. B) The overlap of genes and gene pairs between the three methods at different vertex sizes, plotted against number of genes accepted (upper graph) or number of gene pairs accepted (lower graph). For each size of the graph, the following fractions of the total number of discoveries (genes or gene pairs) are given: RA, MI, PE, RA-MI, RA-PE, RA-MI-PE, MI-PE, where RA = ratiometric method, PE = Pearson correlation and MI = mutual information. PE and MI have a larger overlap compared to the RA method that differs compared to both former ones.

As expected the number of genes increase as the stringency levels are relaxed for all three methods. What differ between the methods are the characteristics of the increase. For RA, there is a 2-step increase profile, where most of the genes are included during the first short interval of high stringency. The inclusion rate then quickly levels off and the remaining few genes are included over a much larger stringency range, demonstrated by the gray filled curve in Figure 4A. On the contrary, the inclusion profile for PE is very close to a bell-shaped curve, spread out across almost the entire stringency range. MI has an inclusion profile in between the RA and PE, it is more bell-shaped than the RA inclusion curve but still has a slight tail towards the lower stringencies as RA has.

Second, the inclusion rates of the pairwise interactions, $|E(G_i)|$, across the stringency range were plotted, Figure 4A dotted lines. Here the difference between RA and, PE and MI on the other hand, becomes even more distinctive. Similar as with the gene inclusion

rate, RA has a steep inclusion followed by a leveling off. The observation made is that when genes are incorporated into the RA graph, they do so by interacting with multiple genes already included. In comparison, the PE and MI graphs demonstrate an exponential increase of the inclusion rate, which starts at a rather low stringency level. The exponential increase happens after the majority of the genes are already included, thus describing a different scenario for PE and MI; as new genes are incorporated they do so predominately by creating fewer interactions at first and often with other newly added genes. This can be seen by the black vertical line in the graphs in Figure 4A, which denotes when 95% of the genes ($|V(G)| = 9244$), are included in the graphs. At this stringency level, RA reports ~5-fold, ~4-fold more gene interactions compared to the PE and MI, respectively. Thus, the mass of the gene pair relationships are incorporated by PE and MI after the majority of the genes are already detected, while for RA the bulk of the relationships are discovered simultaneously as the most of the genes are included.

These findings are corroborated by an analysis conducted by Dr. Kenneth McCue, where the distributions of the Pearson correlation coefficient r, R^2, Δ_{CV}, and I, were found to have generally the same shape, appendix 2, with the exception of the correlation coefficient. Despite the similar shapes, the co-variation between these measures displayed another story, Table 3. I and R^2 were close to the same with a correlation coefficient of > 0.9. In turn, r was moderately correlated with I and R^2. While, Δ_{CV} was highly correlated with $CV(A/B)$ and $CV(B/A)$, none of them was strongly correlated with r, R^2, or I.

Even if the inclusion rates have been shown to be different between RA and, PE and MI, the question remains if the ranking order of the genes is the same for the three approaches. The different inclusion rates are of little interest if the ranking of the genes remains the same. To determine this, Venn diagrams for same-sized graphs from RA, PE and MI, genes ($|V(G_i)_{RA}| \approx |V(G_i)_{MI}| \approx |V(G_i)_{PE}|$) and gene pairs ($|E(G_i)_{RA}| \approx |E(G_i)_{MI}| \approx |E(G_i)_{PE}|$) separately, were plotted at different graph sizes, Figure 4B. The graph sizes used can be found in Table 1 for genes and Table 2 for gene pairs. As the results demonstrate the RA order, both for gene and for gene pairs, is to a much higher extent unique, while PE and MI are strongly overlapping.

It can be concluded from these analyses of the ranking orders generated by RA, PE, and MI, that indeed RA detects a distinct set of gene pair interactions, unique both in nature of gene identities and pairwise interactions, especially at the more higher part of the stringency range. These findings are encouraging and implore further investigation into what kind of relationships are uniquely discovered by RA and if they can reveal biological information previously undetected.

GO category differential enrichment among top ranked genes

As a first step in shedding light on the biological relevance of the top-ranked genes by RA, a GO category analysis was performed on the first 4 sets of same-sized graphs, G_i for $i = 1,..,4$ (see Table 1 for graph sizes). Using DAVID, http://david.abcc.ncifcrf.gov (*8, 9*), the GO category enrichment for each interaction graph was determined, with following modifications. There are a number of GO categories with very similar

functionalities; therefore a reduction of plotted GO categories was performed maintaining the resolution of the biological information as much as possible. The condensation of the GO categories involved smaller GO categories being absorbed into the next GO category a level up until the resulting aggregated GO category contained no more than 200 genes out of the 9752 genes included in this analysis. GO categories with a Bonferroni-corrected p-value $\leq 10^{-4}$ were considered enriched and plotted, Figure 5.

Figure 5: (Previous page) **GO category enrichment for the top vertex sets**. The V(G) sizes plotted for which, the aggregated GO categories (biological processes) have a GO enrichment with a Bonferroni-corrected $p \leq 10^{-4}$. The aggregated GO categories are created by using the complete graph (|V(G)|=9752) and grouping all GO categories with less than 200 detected genes together with its closest GO category upwards in the GO tree hierarchy having a detected gene set of ≥ 200.

A general grouping of the given GO categories is further given to elucidating the general trends. The results indicate that genes involved in interactions detected by RA display a higher number of GO category enrichments than PE and MI. The GO categories for which all three methods show enrichment for is often detected earlier in the stringency cut-offs by RA (i.e., smaller $|V(G_i)|$ thus lower i), than PE and MI. Stronger enrichments for RA are seen among genes involved in transcription and RNA processing, translation, and ubiquitination. Only the general grouping of 'cell-cycle' has a slightly stronger enrichment for PE and MI. This observation is further discussed in *Discover subgroupings among sample population*. The enrichments among the 'ribosome' grouping is confirmed by previously published findings, where a mutual information analysis of *S. cerevisiae* expression data discovered several gene networks of which the largest contained mainly ribosomal genes and transcription initiation factors (*10*). The RA findings reported here becomes thus further interesting, as they do not only agree with, here reported PE and MI and the above mentioned study, but they demonstrate a stronger enrichment of ribosome related genes among RA top gene rankings.

KEGG pathways analysis

To corroborate the findings from the GO category analysis, where the top-ranked RA genes exhibited a strong enrichment in biologically validated information, based on the

GO database; a second enrichment study was performed using the Kyoto Encyclopedia of Genes and Genomes database (KEGG, www.kegg.jp, (*11*)). First a general analysis was done determining the degree each method detects relationships, in which both genes are annotated in KEGG, i.e., $|E(G) \cap E(KEGG)|/|E(G)|$, using same graph vertex set sizes $|V(G)|$ for RA, PE and MI, Figure 6.

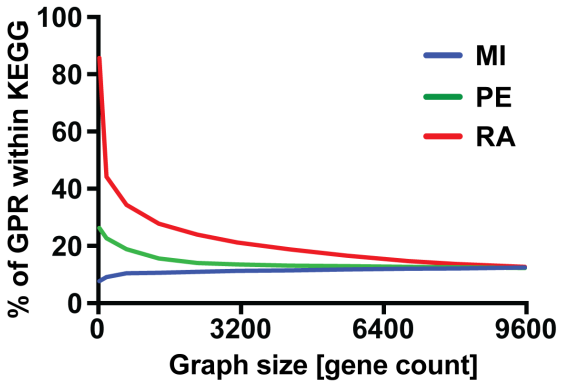


Figure 6: **Percent of KEGG-annotated gene pair detected by each method.** The percent of gene pair relationships (GPR) between two annotated genes in KEGG of total number of relationships reported for each graph size (given in number of genes included). The RA method detects to a higher degree than PE and MI gene pair relationships that are within the KEGG annotated database.

The results confirm the GO category conclusions as the top-ranked gene pairs from RA is strongly enriched in KEGG annotated genes, while the strongest gene pair associations from PE and MI are not involving KEGG annotated genes to the same extent. When ~1000 genes are incorporated into the graphs, 40% of the relationships are between KEGG annotated genes, while for PE it is 20% and MI 10%. Thus the top-ranked RA gene pairs captured, to a higher extent, known pathways and functions in the cell,

indicating that the RA approach possibly is a valuable tool for capturing biologically pertinent relationships.

The general enrichment of KEGG annotated genes leads to the next question of how these gene pair relationships are distributed across and within the individual KEGG pathways. In other words, which ones and how well, does each method recover individual pathways described in KEGG? To investigate this, the number of gene pairs, Figure 7, and the percentage of genes, Figure 8, in which both genes (in the relationship) are annotated in the KEGG pathway, is plotted against the graph size for the three methods.

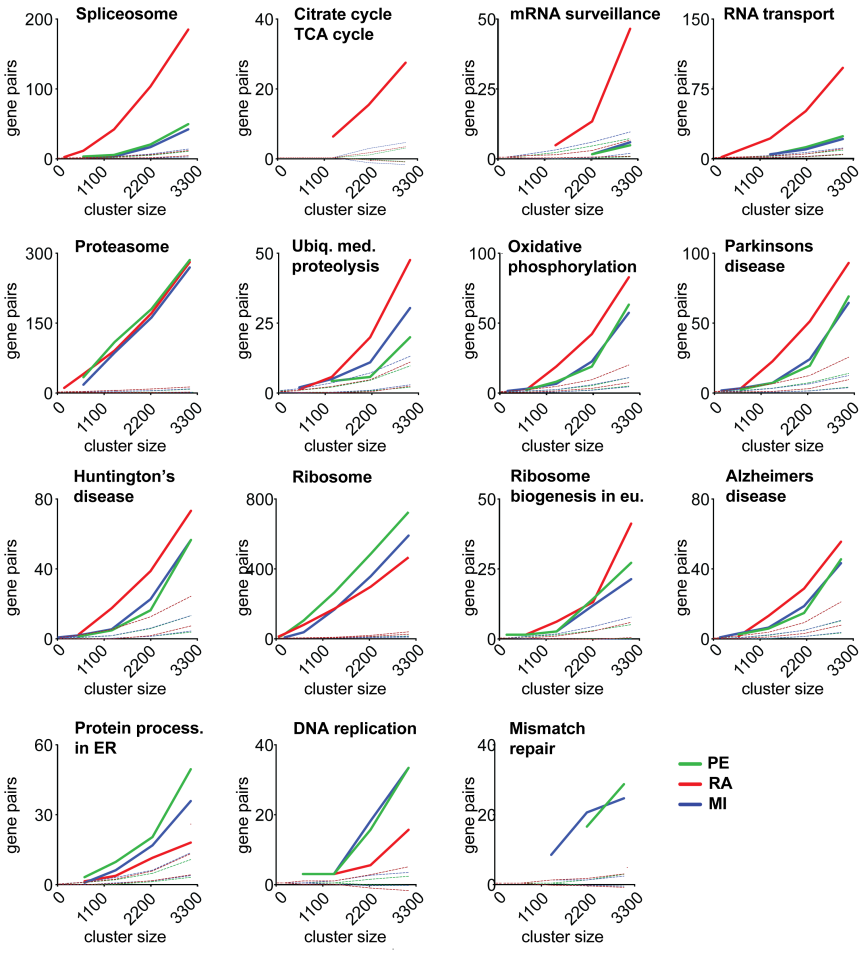


Figure 7: (Previous page) **KEGG pathway enrichment analysis**. The number of gene pair relationships within the same pathway plotted against an increasing graph size. The RA method, red, MI, blue, PE, green. Only pathways with gene % ≥ 25 for at least one method at graph size ~3000 genes are included in the figure. Pathways detected early in the graphs are: ribosome, proteasome, spliceosome, RNA transport. Pathways picked up by the RA method more strongly than PE and MI are both basic cell function pathways as well as diseases, such as: spliceosome, RNA transport, mRNA surveillance, and citrate cycle TCA cycle. Slightly stronger: proteasome, ubiquitin mediated proteolysis, Parkinson's disease, oxidative phosphorylation, and Huntington's disease. Pathways picked up equally well by all three methods are ribosome, ribosome biogenesis in eukaryotes, and Alzheimer's disease. Where PE and MI methods do better than the RA model is in pathways protein processing in ER, mismatch repair, and DNA replication.

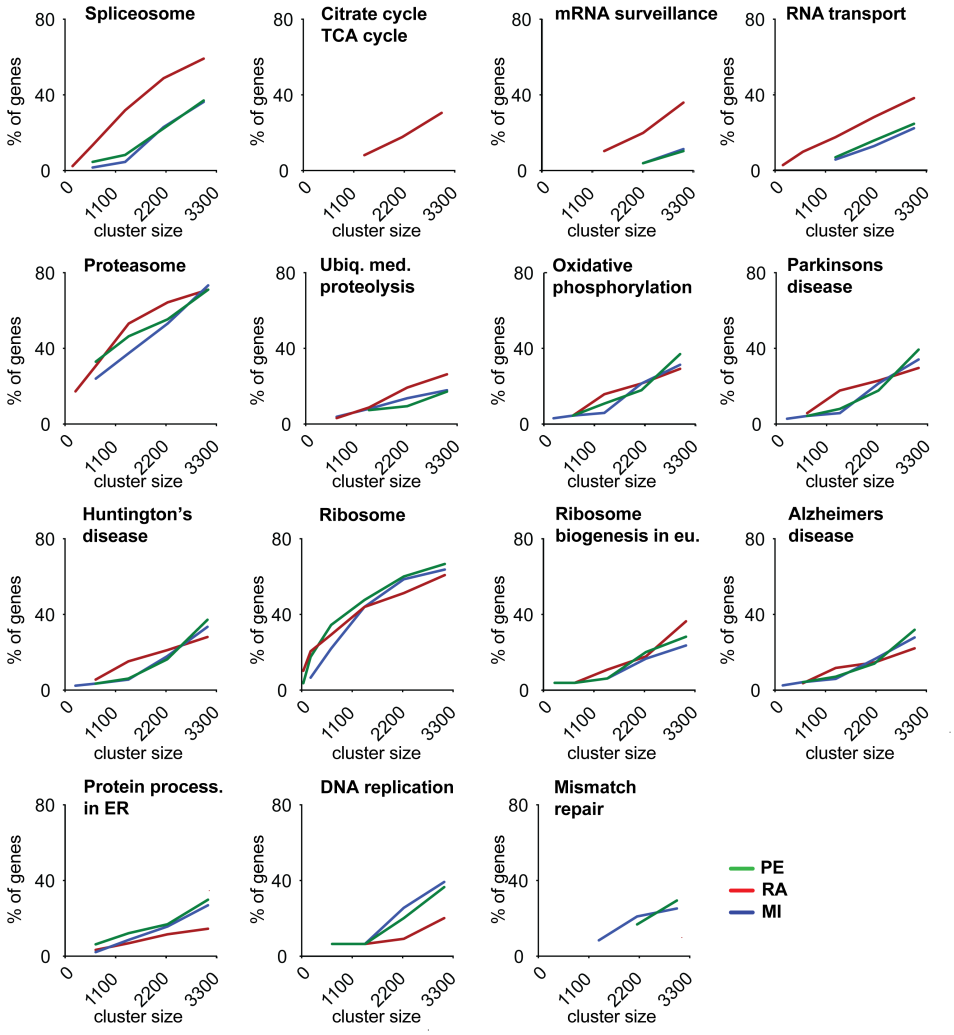


Figure 8: **KEGG pathway enrichment analysis, gene coverage**. The percentage of genes in a pathway discovered plotted against an increasing graph size. The RA method, red, MI, blue, PE, green. A gene is included if it is in at least one relationship with another gene from the same pathway. The same pathways as in figure 5 are plotted.

Only pathways, where at least 25% of the genes were part of at least one of the three graphs at $|V(G)| = 3100$ were considered. Fifteen pathways in total were recovered by at least one of the methods. Overall, RA discovers more pathways and generally to equal or a higher degree than PE and MI. With the largest differences seen in the spliceosome, mRNA surveillance, and RNA transport pathways, where RA outperformed PE and MI. The few pathways where RA was worse at detecting the pathways are protein processing in ER, DNA replication, and mismatch repair. These outcomes are in accordance with the GO category analysis where RA had the strongest presence in core biological processes, while switched roles were seen in cell cycle related categories (see section *Discovering subgroupings among sample populations* for plausible explanation). The question is if the differentiation of pathway detection has its roots in the biological characteristics of the pathways or if there is another possible explanation to why RA recovers certain pathways undetected by PE and MI?

Expression characteristics among detected KEGG pathways

The difference between RA and, PE and MI, in being able to detect gene pair relationships despite narrow expression ranges, makes expression dispersion a topic of interest when investigating the different pathway-discovery rates. In an effort to investigate the extent gene expression range sizes are correlated to pathway discovery, the detected pathways in Figure 8, were grouped into 4 groups according to how well the methods detected gen pair relationships. Groupings were made as followed: 1) pathways identified much better by RA compared to PE or MI; 2) pathways identified comparably by all three methods, but slightly better by RA; 3) pathways in which an approximately

equal number of genes was detected by all three methods; and 4) pathways better identified by PE and MI. For each group separately the *CV(FPKM)* distributions of the pathways were plotted as kernel densities, Figure 9A, group 1 at the top and group 4 at the bottom.

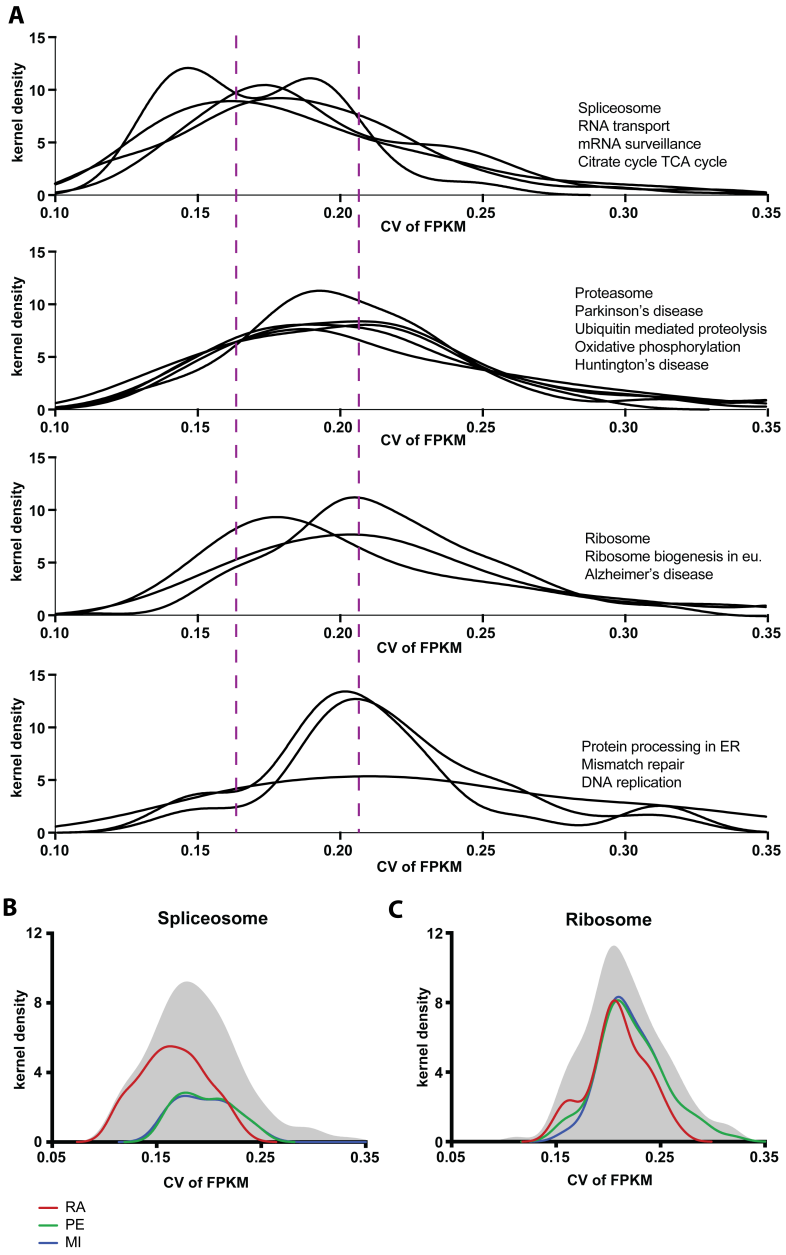


Figure 9: (Previous page) **Distribution of CV(FPKM) per KEGG pathway.** A) The kernel density, per pathway, of CV(FPKM) for the pathway's genes found in our data set. Top graph are pathways detected much more strongly by the RA method than by the PE and MI. Second to top graph are pathways picked up slightly stronger. Second to bottom graph are pathways detected equally by all three methods. Bottom graph contains curves better picked up PE and MI compared to the RA method. The better the RA method detects a pathway compared to the other two, the further left its curve sits. The dotted purple lines indicate how the distributions shift towards the left as you go up the graph series. B and C, two pathways, one from the top graph and one from third down graph, showing the densities for the subgroup of genes each method detects at $V(G)_5 = 2220$. The pathway CV(FPKM) density (gray), RA (red), MI (blue), PE (green).

There is a broad trend across the plots where a diminishing discovery rate by PE and MI is accompanied with pathways exhibiting an expression range distribution curve centered further towards smaller *CV(FPKM)*. In appendix 1 section *Statistical test of CV(FPKM) groupings* Dr. Kenneth McCue shows that group 1 has statistically significant smaller (at less than the .001 level) average CV(FPKM) compared to groups 2, 3, and 4. In addition, group 4 is significantly larger (at less than the .005 level) then group 1, 2, and 3. Group 2 and 3 are deemed indistinguishable from each other. Two arguments can be made here; they might have little biological difference but are worth discussing as they highlight the distinction between RA and PE and MI in slightly alternative ways, which in turn can facilitate the appreciation of the rudimentary difference in the approach of RA. One argument would be that PE and MI finds the basic set of pathways present and for unknown reasons RA discoveries an additional set of pathways, assuming PE and MI give the ‘standard result’. This would then indicate that it is something unique with the pathways that RA uniquely recovered. But no such evidence could be found. The other argument is that the extended set of pathways RA discoveries is all equally ‘standard’ in terms of biological prudence. The fact that PE and MI do not report certain pathways is a

failure in their capability to detect gene groups heavily populated by genes with highly invariant expression. As the results indicate, PE/MI performances are increased as the expression ranges increase. In contrast, the expression range sizes are not imperative for the success for RA, embodied by the comprehensive identification of the pathways. This is further shown by the *CV(FPKM)* distributions for the detected genes per method (at $V(G)_5 = 2200$) for two pathways: ribosome and spliceosome, Figure 9B. The ribosome is an example of a pathway detected equally well by all three methods and the gene expression ranges for the genes detected have the same size range as the pathway's genes in general (gray curve). While for spliceosome, a pathway that exhibit much more narrow expression ranges overall (gray curve), display a higher degree of detection of RA, detecting the genes with the lower *CV(FPKM)*, while PE and MI detect fewer genes and those detected have larger *CV(FPKM)*.

The correlation between RA unique success in detecting a pathway and the pathway's expression range sizes could have a simple explanation of little interest: the reported gene pairs are detected because the genes have a constant expression across samples which produce a false strong association using RA. Two non-fluctuating genes can exhibit a stable ratio without having biological significance, they just appear to be jointly calibrated. Two analyses were executed to address this concern. The first estimates the percentage of gene pairs, Figure 7, that would be detected by each method in a randomly assembled pathway (dotted lines in each sub graph). The scrambled pathways are assembled by randomly choosing, from the complete dataset, the same number of genes as the original pathway and such that the FPKM distribution of its original genes is

maintained. The scores demonstrate a clear separation between real RA detection and the scrambled data. Thus the gene pair relationships RA discovers within KEGG pathways are not reported as an effect of the genes having low expression dispersion across samples. Furthermore, it is noteworthy to comment on how PE and MI perform on the scrambled data as there are a couple of pathways where the real results are much closer or overlap with the perturbed pathways, the clearest example is 'mRNA surveillance'. Thus there is a smaller margin, when PE or MI is applied, between the measured pathway yield and what is the baseline detection occurring inadvertently. A second analysis was implemented to further eliminate this possible predicament of the stability of the gene expressions boost the ratiometric association strengths they participate in. For the complete data set of 9752 genes, the *CV(FPKM)* was plotted against the strongest relationship's *CV(ratio)* for each gene, Figure 10.

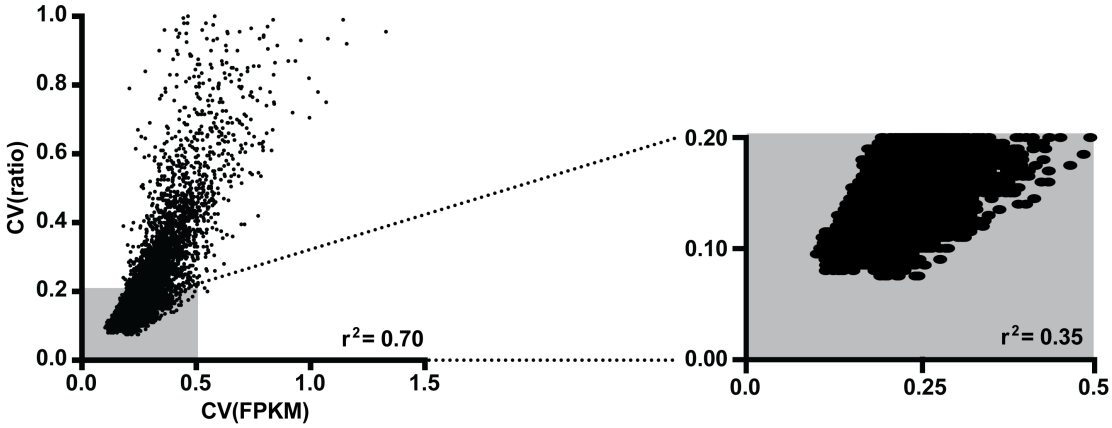


Figure 10: **CV(FPKM) versus CV(ratio)**. Plotting the CV(FPKM) against the CV(ratio) from the most stable ratio per gene, including the 9563 genes detected by RA. For the entire gene set CV(FPKM) correlates with CV(ratio), but not among the top 7000 RA-ranked genes, seen zoomed graph (CV(FPKM) < 0.5, and CV(ratio) < 0.2). Thus for the majority of the genes and especially the top RA-ranked genes, the degree of gene expression dispersion does not predict the RA-ranking order.

For the entire gene set the correlation has an r^2 of 0.7 indicating that there is a general dependency between more stable gene expressions and how stable ratios they exhibit. Though, this pattern vanishes quickly and it suffices to remove less than 30% of the genes, looking at the top RA-ranked 7000 genes, for the correlation to decrease to an r^2 of 0.35, zoomed in graph in Figure 10. In conclusion, a gene's expression range does not anticipate its rank in a RA analysis. The RA method ranks gene pair relationships in a manner distinct from the meek order of the genes expression fluctuations within the sample group. Furthermore, the ranking produced by RA present a biological account, partially missed by PE and MI, centered on core pathways and processes described both in KEGG and GO databases.

Estimating the number of false positives reported

To estimate the number of false positives reported in the bulk B-cell dataset, the distribution of the expression variation in the dataset was determined, Figure 11A, and a simulation was performed as outlined in chapter 2. The expression pattern for two independent genes was generated with a normal distribution $N\sim(\mu, \sigma)$ with μ = [1, 2, 5, 10, 20, 50, 100, 500, 1000] and σ = [0.1, 0.15, 0.2, 0.25]. Each run had an n = 462 and each μ:σ combination was run with 10 iterations. According to the simulation a Δ_{CV} = 0.01 and studying a CV(Ratio)-range of 0 $\rightarrow$ 0.13, limits the number of false positives reported to close to zero (0.015%).

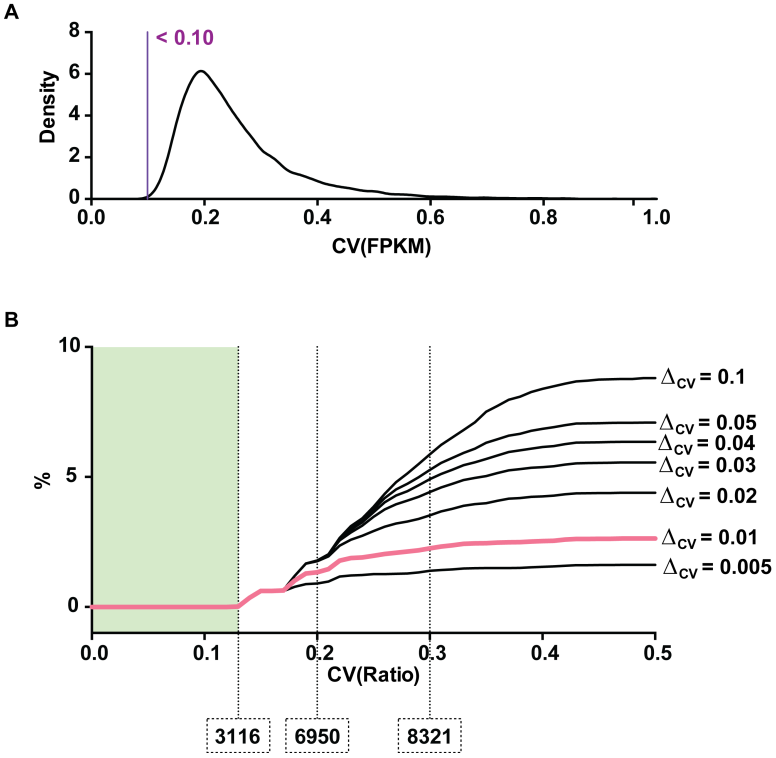


Figure 11: **Estimating false positive in the B-cell bulk dataset**. **A**. The distribution of CV(FPKM) for all genes in the dataset. A CV(FPKM) ≤ 0.10 is indicated (purple line) which in the B-cell dataset includes no genes. **B**. The percent of accepted gene pairs when simulating two individually generated genes' expressions (see chapter 2 for details), for varying Δ_{CV}-levels. The gene counts at CV(Ratio)-levels 0.13, 0.2 and 0.3 are given in dotted boxes. To mimic the dataset the simulation results was adjusted to only include generated genes with CV(FPKM) ≥ 0.1, thus [0.1, 0.15, 0.20, 0.25]. The chosen Δ_{CV} of 0.01 is colored in pink and the CV(Ratio)-range, 0 to 0.13, analyzed in the KEGG-analysis is marked green. As the results indicate such settings minimizes the number of false positives reported to ~0.015%.

Connectivity trends versus FPKM variation

The above KEGG-findings imply that there is a group of KEGG pathways, which have gene populations of overall more narrow relative expression ranges than the rest of the pathways. One reason why certain pathways have genes with less inter-sample expression fluctuations than others could be due to how many other pathway those genes are involved in. The hypothesis would be as followed: a gene that is involved in many

pathways has more interaction partners to consider when its expression level is calibrated. The more partners the less flexibility is possible (see figure 12 for schematics).

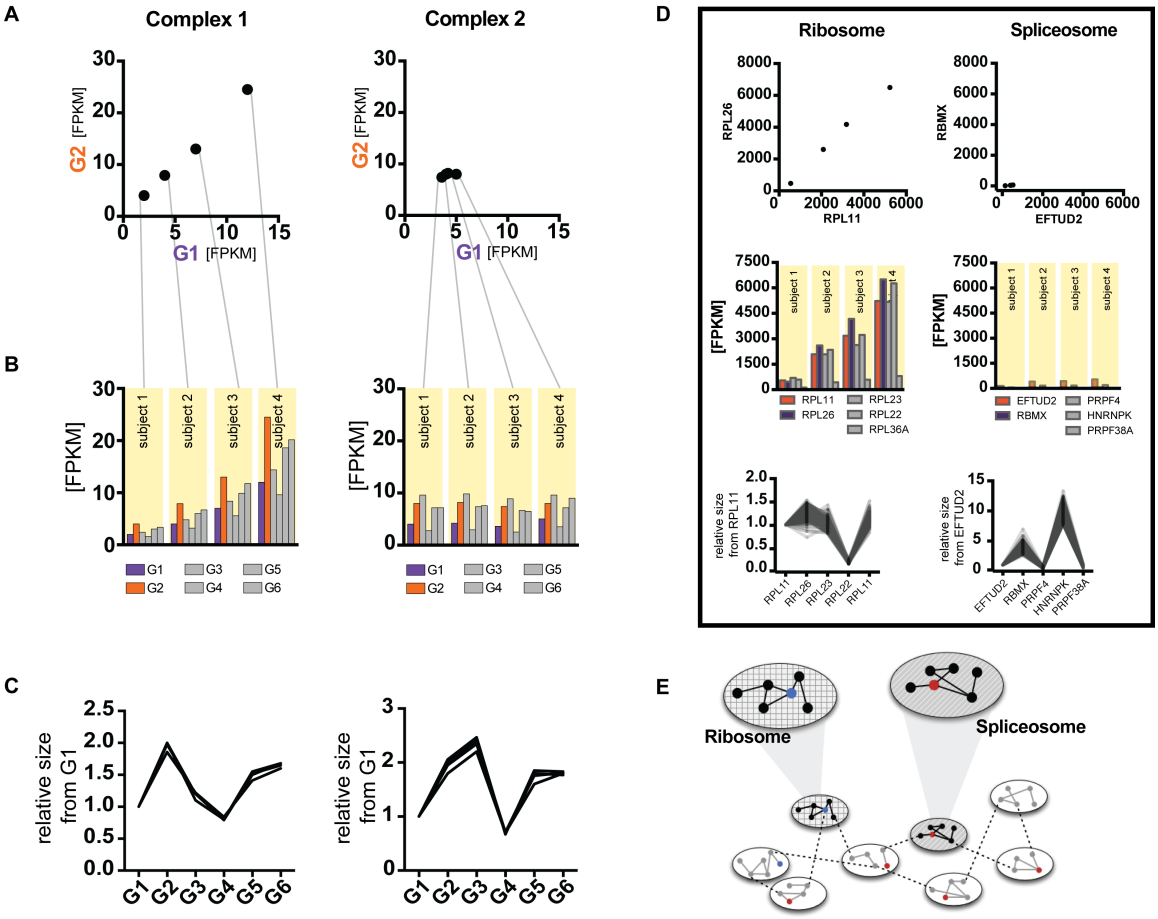


Figure 12: **Hypothesis of multi-pathway occupancy decreases gene expression dispersion.** Two hypothetical complexes, 1 and 2, containing 6 imaginary genes each, exhibiting the contrasting features of high and low, respectively, expression dispersion across 4 subjects. **A**) The expression levels of the two first genes G1-G2 and G1-G2 for complex 1 and complex 2, respectively, plotted against each other. **B**) The expression levels for all 6 genes per pathway in each of the 4 subjects. **C**) The relative expressions, of all 6 genes, compared to the first gene. **D**) The same schematics but with two real examples: the ribosome and the spliceosome. The top and middle graphs displays the overall expression spread using typical expression values across expression range in the sample group. The bottom graphs displays data from all 462 subjects. **E**) Ovals represent pathways and the lines indicate gene pair relationships, solid within pathway and dotted between pathways. There are equal number of intra-pathway gene pair relationships

within the ribosome and the spliceosome complexes. But the spliceosome complex has a much higher number of genes occurring in multiple pathways (blue dot) than the ribosome (red dot) and thus its expression variation as a 'complex', between subjects, is more limited as a higher number of other pathways depend on the same gene to be correctly calibrated. On the other hand the ribosome complex has fewer of this kind of multi-pathway genes and thus is more independent and thus its 'complex' concentrations can be allowed to fluctuate to a higher extent between subjects. As such, both complexes have relative expression levels that are important for the complex to function. The difference is the degree the 'complex' as an entity is varying across subjects. RA, PE and MI will pick up a complex that is fluctuating, but a complex that has smaller fluctuations between subjects can only be picked up by RA.

This proposed correlation was tested by plotting all KEGG-annotated genes' *CV(FPKM)* as a function of the number of pathways they are identified in, Figure 13.

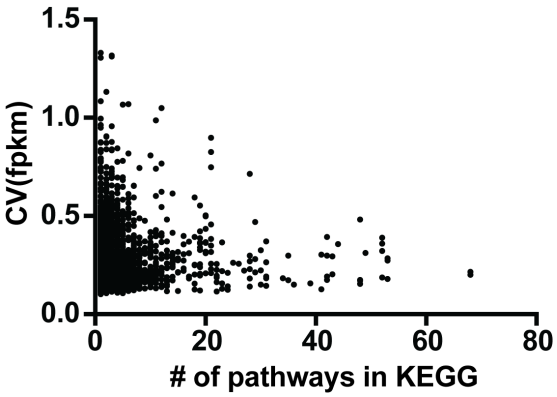


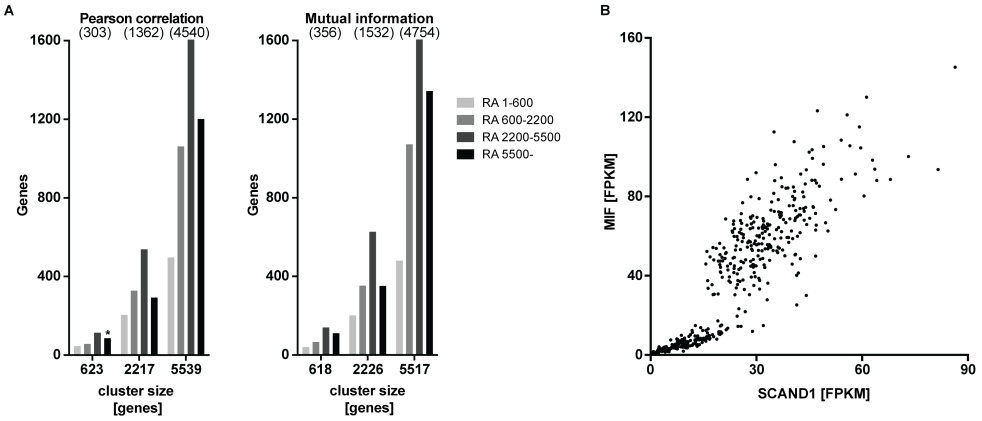
Figure 13: **Number of KEGG annotations correlating with expression dispersion.** Plotting the number of KEGG pathways versus the CV(fpkm) for all genes annotated in KEGG. A negative dependence can be observed where the higher number of KEGG pathways a gene is annotated in the narrower expression range can be anticipated. The trend observed by the eye is not statistically significant due to too few data points at the higher end of the x-axes.

The observed trend supports the hypothesis but it was not statistically significant due to too few data points at the higher end of the x-axes (# of pathways in KEGG). The 'trend' reveals a presence of large expression dispersions only among genes participating in a low number of pathways. As the number of pathways increase, the maximum expression

range decrease. It is important to note here that the dispersion discussed is between samples, i.e., human subjects, in the sense that a highly multi-pathway gene has little room for expression fluctuations between system replicates, i.e., human samples. This does not imply that such a gene cannot have vastly difference expression levels under a variety of biological conditions.

Discover subgroupings among sample population

An analysis of the gene pair relationships scored high by PE and MI and not accepted as conforming to the ratiometric definition (thus discarded by RA), reveals a large portion of the PE and MI reported gene pairs as false positives, Figure 14A.



C

GO term	p-value	Bonferroni corrected
Regulation of apoptosis	$1.3 * 10^{-3}$	0.604
Reg. of prog. cell death	$1.4 * 10^{-3}$	0.632
Reg. of cell death	$1.5 * 10^{-3}$	0.642
Positive reg. of apoptosis	$7.7 * 10^{-3}$	0.996
Positive reg. of prog. cell death	$8.0 * 10^{-3}$	0.996
Positive reg. of cell death	$8.2 * 10^{-3}$	0.997
Reg. of transcription	0.012	1.0
Phosphorylation	0.015	1
Reg. of phosphorylation	0.04	1
Reg. of caspase activity	0.04	1

Figure 14: (Previous page) **Analysis of gene pairs highly ranked by PE and MI but rejected by RA.** A) For 3 cluster sizes (~600, ~2200, ~5500), the number of genes involved in a gene pair relationship not accepted by the Δ_{CV} cut off (total number given in parenthesis above bars) is binned into one of four groups (1-600, 600-2200, 2200-5500, 5500-) according to which RA cluster size it makes its first appearance. B) An example of a gene pair (SCAND1 and MIF) which exhibit a two-subgroup pattern. Roughly 40% of the gene pairs in the top PE and MI clusters found in RA5500- have such behavior. C) For PE, the first 10 GO terms (biological processes) for the gene group in the top cluster size (623), found in the largest cluster size (5500-) in the RA cluster, column marked with * in A. Gene pairs not accepted by the RA method and found in the top cluster in both the PE and the MI, contains genes that to the most part, are low ranked in the RA method. This gene group is enriched in GO terms involved in cell death and its regulation.

Among these gene pairs ~40% have a 2-regime expression pattern, in Figure 14B an example is shown by plotting SCAND1 against MIF. As can be seen (also holding true for most of the 2-regime cases), the smaller subgroup is highly expressed and thus drives the correlation up. Among the genes detected in this analysis, there was an enrichment of genes involved in apoptosis, regulation of cell death and caspase activity, Figure 14C. It could be speculated that if a smaller number of the cell samples was grown under, intentionally or unintentionally, more stressful conditions, that would produce this type of findings. Noteworthy, is that no 2-regime gene pair relationships were found among the top-rankings by RA.

These findings indicate that a cross-examination of the results of either PE or MI with RA, would be a possible approach to discerning subgroupings among a large dataset. By selecting the gene pair relationships, which are highly ranked by PE or MI and not accepted by RA, and from them identify the samples deviating from the main expression pattern, sample subdivision can be found.

Single cell data

Recent developments in RNA-Seq methodology has made it possible to obtain global gene expression data from single cells. This new avenue of transcriptome analysis opens up great possibilities of determining the individual cells gene expression and how the stochastic expression profile for a population of cells is featured. As RA is suited for homogenous sample groups, a pilot experiment was implemented to explore how it would perform in such a setting. The single cell data used was two sets of 10 single cell RNA-Seq libraries of B-cell lymphoblastoid cell lines (*12*). By analyzing the two sets separately, a consistency measure could be obtained for the three methods tested. In the analysis 782 genes were used, which reflected the number of genes expressed in both data sets. First, I determined the ranking consistency between the two data sets for each method separately, Figure 15A.

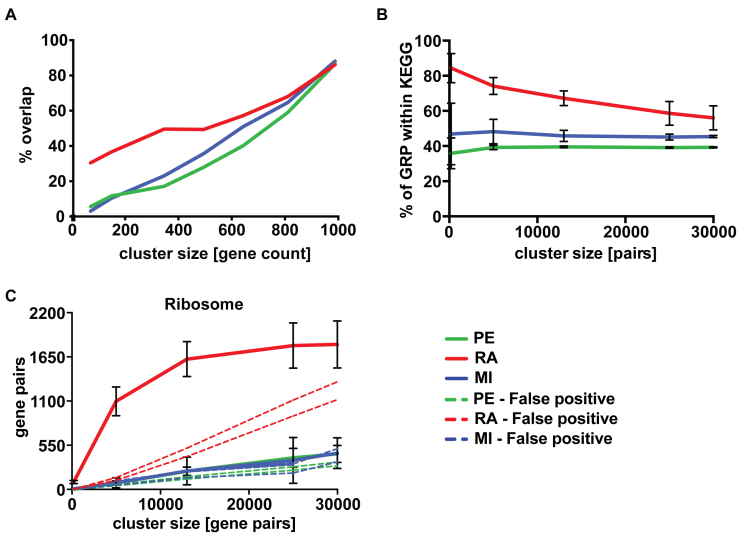


Figure 15: (Previous page) **Single cell analysis.** Comparing the results from two separate batches of 10 single cell samples, including 782 genes expressed in both batches. A) The ranking stability between the two batches for RA, MI and PE. B) The KEGG enrichment per graph size. Error bars describe the variability between batches. The only KEGG pathway sufficiently represented in this data set is the Ribosome. The number of gene pair relationships reported within the ribosome pathway per graph sizes (solid lines) and the false discovery interval (dashed lines).

The ranking consistency for RA was higher, >40%, between data sets compared to both PE and MI. Furthermore, the RA results had a higher enrichment in KEGG annotated genes among the top-ranked genes compared to PE and MI, Figure 15B. This is consistent with the previous results where RA demonstrated a stronger enrichment than PE and MI, in the bulk-RNA. The analysis of the individual KEGG pathways showed only a detection of the ribosome pathway. This might be expected, as there are so few genes included in this analysis, thus rendering most of the pathways missing large portions of their gene populations. The ribosome is still detected as it is populated by very robustly and highly expressed genes and thus will obtain measured expression values even when the measurements are less optimized. The interesting observation from this analysis, Figure 14C, is that compared to the previous bulk-RNA data set, where all three methods detected the ribosome equally well, now in these single cell data sets, only RA picks it up as a pathway with high intra-pathway connectivity. Keeping in mind that single-cell RNA-seq is still in its cradle and a lot of optimization is still to desire, these results indicate that RA is well-suited for such analysis, possibly even to a higher degree than PE and MI.

Discussion

The general approach when designing and executing global gene expression analysis has been to extract functionally interacting gene groups or pairs by identifying which genes co-fluctuate across cellular conditions. A plethora of such multi-state analyses have been published, involving different tissues, different species, time courses but also developmental stages (see chapter 1 for references). As the strategy is, and requires, the gene expressions to fluctuate across the cellular conditions, the findings are limited to discovering only gene cohorts with varying expression profiles. The question is then: are there biological relevant gene pair relationships, which are invariantly expressed, and thus hitherto been missed?

What if there is another type of approach that would render it possible to capture such gene pairs? Let us hypothesis that you could detect all present gene pairs in one cellular state independently of including additional states in the detection process. Then the picture of each cell state can be constructed individually, and then be compared to each other. The roster of each cell state would be solely based on information extracted from its own kind and not be inflated or deflated by influences from the other states. Such a reasoning would require each sample group representing the different states, to contain as 'pure' as possible biological replicates of the state to minimize convoluting input. In other words, a homogenous sample group would be ideal. It could then be theorized that the discoveries could be made by comparing two cellular states, here given as an example of healthy (CN) and disease-affected (DS) samples groups. When describing the *CN* sample group there is a gene pair scoring as a strong relationship, though with very low

expression dispersion rates. But their expressions are completely decoupled from each other in the *DS* state either, causing the condition or as a consequence of it. In this scenario, conventional analytical methods, such as PE and MI, would not have picked up this gene pair relationship as the expression does not fluctuate across the two conditions: in *CN* the expression is invariant and in *DS* the two genes are dis-regulated. On the other hand, if each cellular state could be analyzed separately to determine the gene pair relationships present in each one of them, the disappearance of the relationship would be noticed. RA would render such analysis possible.

This type of new orthogonal approach to gene expression analysis renders it well to the use of the introduced method RA, which is capable of discerning gene pair relationships of both high and low dispersion rate in a single homogenous sample population. To show that such is the case, I applied RA on 462 B-cell lymphoblastoid cell lines. This data was chosen both on the basis of the large sample number and on the characteristics of the cellular state. The desire was to have cells being close to or in a homeostatic state in a controlled environment. Cell cultures are preferred as the conditions can be supervised and the cells can be harvested under similar conditions compared to, for example, fresh tissue collection where the subject's general state could affect the cells' gene expression profile to some extent.

The first step in analyzing the B-cell data set using RA was to determine the effect of the 2 pre-set parameters: the expression [FPKM] cut-off for a gene to be included in the analysis and the Δ_{CV} set to determine if a gene pair relationship is in agreement with the

ratiometric definition and thus should be included in the analysis. Studying their effect on the inclusion rate of the genes and gene pairs tested the influence of these two parameters on the reported rankings. The results indicate that the top-ranked gene and gene pairs in the RA analysis is little effected by the FPKM cut-off used. When more lowly expressed genes are included in the analysis, they tend to get reported first when the stringency cut-off is more relaxed. Thus the findings made by RA are not artifacts produced by inclusion of very lowly expressed genes, which expression profiles might be distorted by experimental noise. Interesting enough, by lowering the FPKM cut-off more genes are included but there is a smaller increase of gene pair relationships than expected, thus indicating that there are fewer RA relationships detected involving very lowly expressed genes than observed among higher expressed genes. This is in contrast with PE and MI where the FPKM cut-off has a stronger impact on the inclusion rates, with almost 50% more genes included at most and roughly 85% gene pairs included, when comparing the most and least stringent FPKM cut-off (comparing inclusion rates at the stringency level where ~5000 genes are included at $\geq$ 5 FPKM). As such PE and MI are more vulnerable to the effect of possible experimental artifacts, and especially if the study calls for inclusion of very lowly expressed genes, compared to RA.

The second parameter, Δ_{CV}, was tested in a similar matter. Both more stringent and relaxed cut-offs were tested by analyzing their effect on the inclusion rate. The results showed the findings made by RA are robust when it comes to gene inclusion rate and gene pair inclusion rate at the more stringent levels. In other words, concerning the top-ranked genes and gene pairs, the Δ_{CV} cut-off level has little to no effect and thus the

results are not dependent on smaller parameter adjustments for this data set. Thus the findings reported by RA are robust and stable, across the whole parameter range tested, in such a way that the top findings are the same and the differences in results are first seen at the lower end of the stringency range. For the remaining analysis the FPKM cut-off was set to ≥ 1 FPKM and the Δ_{CV} to 0.01.

The effect the Δ_{CV}-criteria has on the ranking population of gene pairs and genes is worth contemplating, together with the limitations it brings. In the RA analysis, every possible gene pair combination is first evaluated as being able to be described by a ratiometric definition or not, by using Δ_{CV}. This removes all gene pairs from the succeeding analysis that have an expression profile that is not ratiometric. Note here that it does not imply that the underlying equation is truly ratiometric for the accepted gene pairs, only that the observed expression pattern can be described as complying with the ratiometric definition. In other words, by applying RA we are only interested in what biological information can be extracted by looking at the subgroup of gene pairs that exhibit a ratiometric behavior, out of all gene pairs. Once that subgroup is selected, the ranking occurs within that group based on each gene pair's CV score. This produces a smaller resulting gene pair population than when PE or MI is applied, as the latter is completely inclusive, meaning every gene pair combination is given a score (see figure 3 for schematics). Thus RA captures the subgroup of gene pairs that have a ratiometric expression profile and then produces the internal order of these based upon how stable the observed ratio is for each one of the pairs. Thus, when in proceeding analyses the stringency levels are said to be varied, it refers to the stability of the ratios, meaning the

CV is varied, and not the Δ_{CV}, which determines the compliance to the ratiometric definition.

As a first examination of how RA behaves compared to PE and MI, the inclusion rates, counting genes and gene pairs separately, were studied as a function of the stringency cut-off, using CV for RA, R^2 for PE and I for MI. As the results demonstrate RA differs both in gene and gene pair inclusion profile from PE and MI, which in turn are very similar. In RA the gene inclusion profile has a two state feature, where most genes are rapidly included at high stringency levels and then abruptly the remaining genes are included in a much slower rate. PE has a profile much more similar to a normal distribution shape, while MI is somewhere in between. Furthermore, the gene pair inclusion rates are very different, where RA has a rapid inclusion rate that overlaps to a higher degree with the gene inclusion rate than PE and MI do. The consequence of this is that the majority of the gene pairs accepted by RA are included while genes are still being added to the cluster. While for PE and MI, most of the gene pairs are included once all the genes are already accounted for.

Next question was to determine if the above differential findings only concerned the inclusion rates or if the order of the genes and gene pairs also were different between the three methods. By comparing same-sized portions of the ranking lists generated by the three methods, it became clear that the PE and MI ranking lists both for genes and gene pairs are largely similar, meaning that roughly the same order of genes and gene pairs is observed regardless if the data is analyzed by PE or MI. This is in line with previous

reports, where high similarity between the two has been observed (*13-15*). On the other hand, RA produces a distinctly different ranking order both when looking at genes and gene pairs. In other words, RA draws up a different picture of which genes and gene pair interactions that are potentially of interest in the B-cell lymphoblastoid expression data explored in this study. When the expression data is evaluated based on the individual gene pairs' ratiometricity, there is a re-arrangement of which genes and gene pairs are in strong relationships and which are not, compared to analyses based on common methods such as PE and MI. This manner of comparing the three methods by using same-sized portions of the ranking lists is reoccurring throughout this study and the portion sizes are referred to as cluster sizes.

Subsequently, the next question was to discern if the alternative ranking order of the genes produced by RA had any biological merit to it or not. It is not difficult to generate an analytical method, which produces a shuffled, and thus unique, ranking order but as long as the proposed new order is not biologically meaningful, the method is useless. To test if the RA results capture known, in other words previously published, gene interactions and gene groupings, both GO term and KEGG pathway enrichment analyses were performed. This was to validate the methods' capability of processing biological data as to what degree the detected gene pair relationships reflect the biological processes expected to be in play in the samples. Assuming that the harvested cells were at or close to homeostasis, the majority of their active processes would be core processes and pathways essential for the cells to maintain viability. It is thus desirable that the method ranks highly genes enriching such cellular functions. The higher the enrichment of core

biological processes is among the smaller group of top ranked genes, the better the method depicts the active biological mechanisms present in the cells. Put differently, if a method reports a lot of noise, in other words false positives, the weaker the enrichment will be. Here a dilemma arises as to which gene pair relationships are correct and which ones are false positives. As a previously, in the literature, unreported gene pair relationship cannot be validated as a truly functional one by the proposed method alone, it cannot be claimed either to be nor not to be noise. Thus a method cannot be said to generate false positive relationships as long as the relationships are untested. But what can be argued is that if a method picks up verified gene pair relationships reported previously in the literature, it speaks in favor of the method. Following this reasoning, the results from the GO term and KEGG pathway analyses demonstrate a stronger enrichment in core biological processes among top-ranked genes by RA compared to PE and MI. More of the top-ranked gene pair relationships from RA, involve two genes, which to date, have well-characterized biological functions. The increased enrichment is both present when looking at all KEGG-annotated genes at generally and within many of the individual KEGG pathways. The observation that RA appears to detect more pathways than PE and MI do, give rise to the next question of why. Is there something biologically different between pathways equally well detected by all three methods and pathways only picked up by RA? Or, is the differentiation of a more technical nature, where the limitations of some of the methods are in effect?

Going back to the fundamental differences between RA and, PE and MI on the other side, it could be pertinent to consider their different sensitivities to expression dispersion rates,

see chapter 2 for further details. A possible hypothesis is that RA detects all pathways with a high degree of intra-pathway gene pair relationships regardless of the expression dispersion rates present, as RA is not affected by expression variations. In contrast, PE and MI cannot detect gene pair relationships with too narrow expression ranges and thus if a pathway is population by genes with low expression dispersion rates they will fail to detect it. To assess this hypothesis, the distribution of expression dispersion rates, or relative expression ranges, for each detected pathway was determined. The distribution curves indicated a trend where the curves of the pathways, which all three methods picked up equally well, were shifted towards larger relative expression ranges. The smaller portion of a pathway PE and MI detected, the more that pathway's distribution curve was shifted towards more narrow expression dispersion rates. Thus it appears that the difference in pathway detection is of a technical matter, caused by the limitations of PE and MI, which requires a certain degree of fluctuations across samples to be able to detect relationships of interest. As such, these results should not be seen as RA detecting gene pair relationships with a unique biological characteristic that has failed detection until now. Rather, out of all biologically active gene pair relationships in the cell, there are some that are more difficult or even impossible to detect with commonly used methods, such as PE and MI, due to their high degree of invariant expression across samples. These can be detected be RA, which in turn enables RA to depict a more complete picture of the biological processes currently active.

In a biology point of view the above findings are intriguing, as it implies that there is a group of KEGG pathways, which have a gene population of overall narrow expression

ranges. These pathways are at one end of the spectrum with a continuous positioning of pathways until the other end where pathways are mostly populated by genes with highly fluctuating expression across samples. Drawing from system logistics and intricate mechanistic modeling, it is reasonable to hypothesis that a pathway's expression range distribution is linked to the number of pathways its genes are involved in. A gene that is required to calibrate its expression level against genes in many other pathways and processes has a decreased degree of freedom to fluctuate. While, the cell has potentially less to loose when it comes to interrupted biological processes, if a gene involved in few pathways exhibits a looser regulation as it has fewer gene partners to consider. Thus it is possible for the gene to have a broader expression range among the samples. It could be imagined that pathways or complexes that are required to fluctuate heavily across different biological conditions, are under a less strict expression constraint in homeostasis as that could render it more difficult to exhibit the fluctuation capacity when needed, for example, the ribosome. The results indicate a trend between number of KEGG pathways a gene is involved in and its expression dispersion rate. There are no KEGG-annotated genes that are highly multi-pathway members that have large expression ranges. As was mentioned in the result section but it is important to emphasize it again: the dispersion rate is between samples, e.i., human subjects, in the homogenous B-cell lymphoblastoid data set. When a gene is said to exhibit a low degree of fluctuation, in refers to between samples/subjects: replicates of a biological system/model. This is not implying that the same gene demonstrating low expression dispersion in this cellular state can have radically different expression values under other cellular conditions.

The hypothesis presented above describing a possible correlation between a pathway's expression range distribution and its degree of multi-pathway involvement, can be further elucidated by examining two pathways from either end of the range size spectrum: the ribosome and the spliceosome. The pathways describe complexes, including both genes that are a part of the complex itself and genes directly involved in its function. When talking about gene expression being calibrated against each other and a reason for such extra level of control by the cell, complexes might be the most easily comprehensible example. The common thinking is that genes that are part of the same complex would be expected to have calibrated expression levels as an over-expression of one of them would be an unnecessary cost for the cell and an under-expression would limit the number of complexes present that is fully functional. This has been shown repeatedly, through large transcriptome studies, for many genes in the ribosome, where the expression levels of its genes are co-fluctuating across biological events and calibrated at homeostasis (*16, 17*). Regarding the spliceosome, the importance of calibrated expression levels between certain splicing factors has been shown through mutation studies (*18-20*). There is growing evidence that the combination and relative concentrations of splicing factors do determine the exact splicing event (*21, 22*). In addition, the importance of correct proportions of certain splicing factors comes from their observed disruption in some cancers (*23-25*) and also experiments where forced dis-regulation of these expression level relationships activated cancerous mechanisms (*19, 26, 27*). The fact that RA detects both pathways and reports similar portions of their genes does not only generally account for the RA results but the fact that many of the individual splicing factors proven to

exhibit such a behavior is found as top-ranked by RA further attest to the findings. Examples of such splicing factors are HNRNPK (*18*), SRSF1 (*19*) and TRA2β (*20*).

PE and MI, successfully detects the ribosome as having parts of its genes being correlated in expression levels, by the expression spreads across the samples being, as the methods require, sufficiently large. For the same reason, they fail to discover the strong associations present among a large portion of the spliceosome genes: they have too narrow expression ranges and thus escape detection by PE and MI. Thus the situation is as followed: these two complexes both have a large portion of their genes highly intra-calibrated but one, the ribosome, varies in its overall complex concentration among samples, while the other, the spliceosome, have a much more fixed absolute complex concentration across the samples, Figure 12. It could be speculated that the massive changes in ribosome concentrations required when the cell goes through different cell processes, like cell growth or acting on certain stimuli, has made it more independent from the rest of the cell components. The capability of huge fluctuations in numbers and the relatively high degree of autonomy of the ribosome, could give some wiggle room for variation among individual samples. On the other hand, the splicing factors in the spliceosome are known for being involved in many different cell activities: nuclear retention of pre-mRNAs (*28*), mRNA export (*29-32*), and translation initiation (*30*) as examples. This multi-pathway gene occurrences by some of the spliceosomal genes, limits the overall spliceosomal concentration allowed as if it would fluctuate heavily it is possible it would produce a chain reaction where other cellular activities have to adjust, which could lead to the system breaking down.

Finally, the spliceosome and particular one study of it, published by Piekielko-Witkowska et al., presents possible supporting evidence for RA actually detecting biologically present gene pair relationships that a correlation study does fail to report (*33*). The authors observed stable ratios between a subgroup of splicing factors in both a control group of healthy subjects and a group of renal cancer subjects. The appealing observation was the absence of correlation among these splicing factors in the control group and the presence of such in the disease state. A plausible explanation of these findings, which would be in agreement with how a RA approach differs from PE (and also MI): for a cell to function, regardless if it is cancerous or if it is healthy, many of the genes in the spliceosome are kept at tight relative expression levels to each other, though the two states might call for different combinations. In healthy individuals there are little inter-subject expression fluctuations of absolute levels of the spliceosome as a unit, as a consequence of the high connectivity the spliceosome have to other cellular processes (see above sections for further explanation). The absolute expression fluctuations of the 'spliceosome-unit' are increased in cancer cells, keeping in mind that there are still expression level relationships that are present within the spliceosome. These differences in concentrations of the 'spliceosome-unit' could be causative of the cancer or a consequence of it. Regardless, the fluctuations are large enough to enable PE to detect the gene pair associations and thus deem a part of the spliceosome as being correlated. Thus, in conclusion, a reasonable hypothesis is that if the RA approach would be applied to such data, the spliceosome would be regarded as an intra-calibrated complex in healthy subjects, providing indications of gene pair associations possibly vital for the proper

function of a healthy cell. By analyzing the cancer samples, the changes in these 'healthy' gene pair associations can be observed giving valuable information of how cancerous cells have altered their core cellular processes. When, as in the published study, the data is analyzed with PE, there is a risk that the conclusions drawn would postulate that the spliceosome changes into a correlation 'unit' when a cell becomes cancerous, while the real event is a mere alteration in the combination of genes co-calibrated within the spliceosome.

To explore what type of gene pair relationships that RA excluded compared to PE and MI, the gene pairs top-ranked by PE and MI, but rejected by RA as not pertaining to a ratiometric pattern, were examined. Out of those almost 40% exhibited a two-regime expression pattern. Their expression data could be divided up into two samples groups by which the expression profiles were distinctively different from each other. The false positive reporting due to two-regime patterns has been reported before in literature (*13*). The authors also demonstrated that correlation did not report these gene pairs but instead was vulnerable to large outliers. It is this vulnerability combined with the large number of samples in the highly expressed subgroup that makes PE also report these false positives in this study. Two observations can be made here. The first is that RA rejects expression patterns that are not representative for the entire sample group, referring to large subdivisions of the samples and not the presence of single outliers. This was corroborated by the lack of such two-regime gene pairs among the top-rankings produced by RA. The second observation, which ties into the first one, is the possibility of utilizing a combination of RA and, for example, PE, to discover subdivisions within a sample

population. By analyzing a data set with both RA and PE separately and then examine the gene pairs highly ranked by PE and rejected by RA, possible subgroupings of the samples can be detected. More detailed discussion of this subject can be found in chapter 4. To further support the hypothesis of subgroupings of the samples regarding certain gene pair relationships, the GO term enrichment analysis performed on the involved genes, showed focus on cell death and apoptosis. This is in agreement with the few GO terms that PE and MI had a stronger enrichment in than RA, which also involved cell cycle processes. As RA rejected these gene pair relationships, it lacked the enrichment in these GO terms, which in hindsight could be interpreted as a good result. As, if the intent of the study is to find gene pair relationships that are present in all samples, this particular group of relationships should be rejected. In this particular data set, an incorrect conclusion could very well be that B-cell lymphoblastoid cells have an active apoptosis process present, which would incorrectly describe the homeostatic state of these cells as the majority of them do not. These relationships would thus have been seen as ‘universal’ for this data set if not a pair-by-pair examination of the expression plot is performed, which is highly unlikely if the study involves thousands of genes. This is not to say that they are un-interesting to discover, to the contrary, they can describe important variation in cell culture handling, for example. But they can only be useful if they are actively sought out and they can even be misleading if not paid attention to during semi-automated large-scale transcriptome analyses. As RA rejects these gene pairs and only reports gene pair relationships that are ‘universal’ for the data set, this dilemma does not arise in a RA analysis.

A recently explored field of homogenous sample groups is single cell RNA-Seq. These experiments are often designed to study multiple samples from just one cellular state. Therefore RA is a possible candidate for analyzing such data. As a pilot study of analyzing single cell RNA-Seq data, two batches of 10 individual B-cell lymphoblastoid cells were analyzed. Due to the early stage of optimization this sequencing technic it is in, the gene expression population that can rather accurately be measured is very small and limited towards more highly expressed genes. With these restraints in mind, the analysis indicates the results to be in general agreement with the larger data set. Due to the low number of genes included in the analysis only the ribosome was properly detected and only by RA. Even if these are positive results for RA, it needs to be more testing to confirm the appropriateness of using RA for single cell data.

In this thesis, the performance of RA was evaluated by exploring its results in the form of the ranking order of the genes and gene pairs. The ranking is preferred instead of a cluster analysis due to two main reasons. Firstly, as RA rejects a large portion of the gene pair combinations, since they are excluded by the ratiometric definition, the RA results are less ideal for a cluster comparison with other methods. RA's fullest potential thus cannot be explored by clustering (further discussed in chapter 4), but in contrast, it can by ranking. Secondly, a comparison of ranking orders gives a much more detailed and in-depth assessment of how RA differs from the other tested methods. In contrast, with a cluster analysis the assessment could only be done on larger gene subgroupings generated by each method. Ranking gives information about which gene pairs are considered

having the strongest relationships and exact how many relationships each gene participates in at any given stringency level.

Conclusion

When applying RA to a homogenous sample group consisting on 462 B-cell lymphoblastoid cell lines, a more extensive coverage of core cellular processes and pathways was observed compared to PE and MI. The limited discovering yield by PE and MI, but unimpaired for RA, coincided with biological processes containing a large portion of genes with low expression dispersion rates. Furthermore, the general trend was that genes with high multiple-pathway occupancy had low expression dispersion rate. Thus the conclusion is that genes involved in a high number of pathways and/or processes in the cell have a lower degree of expression freedom across samples. Pathways enriched in such genes, are poorly detected by PE and MI, due to the methods' mathematical approach. RA is not limited by the same requirements and thus is able to extract those gene pair relationships, giving a more comprehensive view of the cellular state.

Methods

Data processing, gene expression quantification

The data set used in this study was obtained from the 1000 Genomes Project Consortium (http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/samples.html) (*1*). It consisted of 462 samples of B-cell lymphoblastoid cell lines. The downloaded reads were

aligned, by Georgi Marinov, against the refSeq (*34*) transcriptome (created by applying custom-written scripts for version hg19 of the human genome) using Bowtie 0.12.7 (*35*). Using both ends of the paired end datasets, they were aligned together with the following settings: \verb|'-aS -X 800 e -200 --offrate 1 --best strata'| settings. The accepted alignments were quantified with eXpress, version 1.3.0 (*7*). If a gene had multiple isoforms, the gene's expression value was the sum of the expression values of the isoforms.

GO category enrichment analysis

DAVID (http://david.abcc.ncifcrf.gov) (*8, 9*) was used to calculate the GO term enrichments. These were slightly modified to decrease the amount of GO terms to a number that could be plotted. The modification was that the nodes in the GO term tree was merged upwards in the tree such that the final aggregated GO categories consisted of at least 200 genes (all present in the analysis). The p-values were Bonferroni-corrected and only those of 10^{-4} or lower were considered significant.

Tables

Table1: **Caption**. Vertex sizes |V(G)| with corresponding stringency cut offs and gene counts for each method.

	Ratiometric		**Pearson correlation**		**Mutual information**	
i	*|V(G)|*	*CV*	*|V(G)|*	R^2	*|V(G)|*	*I*
1	22	0.08	28	0.89	22	1.02
2	182	0.09	166	0.85	177	0.87
3	631	0.1	623	0.8	618	0.76
4	1359	0.11	1303	0.75	1363	0.66
5	2220	0.12	2217	0.7	2226	0.59
6	3116	0.13	3184	0.65	3124	0.53
7	4300	0.145	4234	0.6	4214	0.47
8	5567	0.165	5539	0.54	5517	0.41
9	6950	0.205	6863	0.47	6905	0.35
10	8012	0.265	7993	0.4	8022	0.3
11	9563	1.0	9565	0.22	9580	0.19

Table2: **Caption**. Edge sizes |E(G)| with corresponding stringency cut offs and gene pair relationship counts for each method.

	Ratiometric		**Pearson correlation**		**Mutual information**	
i	*|E(G)|*	*CV*	*|E(G)|*	R^2	*|E(G)|*	*I*
1	56	0.085	68	0.87	61	0.95
2	411	0.095	484	0.83	462	0.85
3	4581	0.110	4,919	0.76	4545	0.72
4	8618	0.115	8,043	0.74	8384	0.68
5	42,843	0.130	45,923	0.65	43,474	0.56
6	99,112	0.140	95,861	0.60	104,570	0.49
7	467,497	0.165	476,435	0.46	505,308	0.36
8	1,088,031	0.185	1,089,421	0.37	1,069,168	0.30
9	2,006,347	0.205	1,963,503	0.30	2,072,871	0.25
10	4,943,161	0.255	4,838,371	0.19	4,997,296	0.19
11	1.01×10^7	0.395	1.07×10^7	0.10	1.01×10^7	0.15

Table 3: **Correlation between Measures**.

	I	$\mathbf{R^2}$	***r***	Δ_{CV}	**CV(A/B)**	**CV(B/A)**
I	1	0.9125	0.451	-0.0833	-0.152	-0.139
$\mathbf{R^2}$	0.9125	1	0.631	-0.0587	-0.121	-0.117
r	0.4509	0.6308	1	-0.053	-0.144	-0.136
Δ_{CV}	-0.0833	-0.0587	-0.053	1	0.705	0.723
CV(A/B)	-0.1519	-0.1212	-0.144	0.7052	1	0.126
CV(B/A)	-0.1394	-0.1171	-0.136	0.7230	0.126	1

References

1. C. Genomes Project *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (Nov 1, 2012).
2. E. C. Baechler *et al.*, Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2610 (Mar 4, 2003).
3. S. D. Der, A. Zhou, B. R. Williams, R. H. Silverman, Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 15623 (Dec 22, 1998).
4. G. St Laurent *et al.*, On the importance of small changes in RNA expression. *Methods* **63**, 18 (Sep 1, 2013).
5. L. Goentoro, M. W. Kirschner, Evidence that fold-change, and not absolute level, of beta-catenin dictates Wnt signaling. *Molecular cell* **36**, 872 (Dec 11, 2009).
6. L. Goentoro, O. Shoval, M. W. Kirschner, U. Alon, The incoherent feedforward loop can provide fold-change detection in gene regulation. *Molecular cell* **36**, 894 (Dec 11, 2009).
7. A. Roberts, L. Pachter, Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods* **10**, 71 (Jan, 2013).
8. W. Huang da, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44 (2009).
9. W. Huang da, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1 (Jan, 2009).
10. A. J. Butte, I. S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418 (2000).
11. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome. *Nucleic acids research* **32**, D277 (Jan 1, 2004).
12. G. K. Marinov *et al.*, From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome research*, (Jan 31, 2014).
13. C. O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC bioinformatics* **5**, 118 (Aug 31, 2004).
14. T. R. Hughes *et al.*, Functional discovery via a compendium of expression profiles. *Cell* **102**, 109 (Jul 7, 2000).
15. R. Steuer, J. Kurths, C. O. Daub, J. Weise, J. Selbig, The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18 Suppl 2**, S231 (2002).
16. W. H. Mager, R. J. Planta, Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate. *Molecular and cellular biochemistry* **104**, 181 (May 29-Jun 12, 1991).
17. R. P. Perry, Balanced production of ribosomal proteins. *Gene* **401**, 1 (Oct 15, 2007).

18. A. Expert-Bezancon, J. P. Le Caer, J. Marie, Heterogeneous nuclear ribonucleoprotein (hnRNP) K is a component of an intronic splicing enhancer complex that activates the splicing of the alternative exon 6A from chicken beta-tropomyosin pre-mRNA. *The Journal of biological chemistry* **277**, 16614 (May 10, 2002).
19. S. Das, O. Anczukow, M. Akerman, A. R. Krainer, Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell reports* **1**, 110 (Feb 23, 2012).
20. J. P. Venables *et al.*, Up-regulation of the ubiquitous alternative splicing factor Tra2beta causes inclusion of a germ cell-specific exon. *Human molecular genetics* **14**, 2289 (Aug 15, 2005).
21. C. W. Smith, J. Valcarcel, Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in biochemical sciences* **25**, 381 (Aug, 2000).
22. J. C. Long, J. F. Caceres, The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical journal* **417**, 15 (Jan 1, 2009).
23. G. S. Wang, T. A. Cooper, Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature reviews. Genetics* **8**, 749 (Oct, 2007).
24. C. Ghigna, C. Valacca, G. Biamonti, Alternative splicing and tumor progression. *Current genomics* **9**, 556 (Dec, 2008).
25. A. Srebrow, A. R. Kornblihtt, The connection between splicing and cancer. *Journal of cell science* **119**, 2635 (Jul 1, 2006).
26. S. Gout *et al.*, Abnormal expression of the pre-mRNA splicing regulators SRSF1, SRSF2, SRPK1 and SRPK2 in non small cell lung carcinoma. *PloS one* **7**, e46539 (2012).
27. E. Stickeler, F. Kittrell, D. Medina, S. M. Berget, Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene* **18**, 3574 (Jun 17, 1999).
28. R. Takemura, T. Takeiwa, I. Taniguchi, A. McCloskey, M. Ohno, Multiple factors in the early splicing complex are involved in the nuclear retention of pre-mRNAs in mammalian cells. *Genes to cells : devoted to molecular & cellular mechanisms* **16**, 1035 (Oct, 2011).
29. Y. Huang, R. Gattoni, J. Stevenin, J. A. Steitz, SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Molecular cell* **11**, 837 (Mar, 2003).
30. G. Michlewski, J. R. Sanford, J. F. Caceres, The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E-BP1. *Molecular cell* **30**, 179 (Apr 25, 2008).
31. K. Strasser, E. Hurt, Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p. *Nature* **413**, 648 (Oct 11, 2001).
32. M. L. Luo *et al.*, Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. *Nature* **413**, 644 (Oct 11, 2001).
33. A. Piekielko-Witkowska *et al.*, Disturbed expression of splicing factors in renal cancer affects alternative splicing of apoptosis regulators, oncogenes, and tumor suppressors. *PloS one* **5**, e13690 (2010).

34. K. D. Pruitt, T. Tatusova, W. Klimke, D. R. Maglott, NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic acids research* **37**, D32 (Jan, 2009).
35. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

Chapter IV

FUTURE DIRECTIONS

Abstract

RA was shown in this thesis, to produce a gene pair ranking that is distinctive to PE and MI. Furthermore RA had a higher enrichment of core biological processes and pathways than the latter. This thesis has not fully explored the possibilities of using RA in additional biological settings, therefore, there are experiments were hypothetically RA could be advantageous for extracting biological information. These include, but are not limited, to the application to data sets containing multi-cellular states, disease and control groups, and single cell data is discussed. As clustering is a standard analytical tool in gene expression analysis, the possibility of using RA-calculations as the input for clustering is examined elucidating the limitations and the cautions are expressed. Finally, the potential of using RA combined with PE or MI for detecting sub groupings in sample data sets, is addressed.

Introduction

RA applies very simple criteria to a gene pair relationship, based on the ratio profile. This approach circumvents the limitations exhibited by more classical methods, such as PE and MI, arising from their prerequisite of a certain degree of fluctuations among the samples' absolute expression values. This thesis shows that the advantages from such an alternative methodology are both the decrease in false positives reported and the more

extensive identification of known biological relevant relationships among the high-ranked gene pairs. Thus RA achieves the set expectations of producing biologically pertinent gene pair relationships under the given premises of using a homogeneous data set. But the vast majority of studies done in biology include multi-dimensional data sets. They often involve several cellular states, be it different tissues, time courses and/or perturbed cellular conditions. These are gene expression avenues not explored in this thesis but it is worth discussing possible strategies of implementing RA in such experimental set ups.

An aspect of RA, vital to consider when applying the method to any data set and/or when combining it with additional methods, is its non-all-inclusive property. RA performs a pre-selection from all the possible gene pair combinations, so that only the gene pair relationships conforming to a ratiometric definition are included in the final analysis. This will have implications when RA, for example, is combined with clustering algorithms, further addressed below. Moreover, it should also be considered during multi-cellular state analyses as it provides an extra dimension to possible changes in gene pair relationship regime between sample conditions. As addressed below, RA reports both if a gene pair relationship, comparing two sample conditions, has a weakened ratio, thus becoming more fluctuating in the second sample group, or if the relationship has become non-ratiometric and thus is rejected completely in the second sample group. This will extend into an application to disease studies, where such distinction could have beneficial utilization. Finally, proposing a strategy to detect subgroups of deviating samples in large data set, capitalizing on the differential results between RA and PE/MI.

Further directions

Multi-cellular states datasets

Analyzing multiple-cellular states data sets include studies such as those involved with more than one tissue type, time lines, and/or controls versus unhealthy samples. RA can easily and time-efficiently be applied to this type of data and by its different analytical approach, extract information about gene pair relationships previously missed. The proposed steps of analysis, using RA, are to first calculate the gene pair relationships accepted and scored in each cellular state separately. Then, both the sets of accepted gene pair relationships and their ratiometric stability scoring, CV, are compared between the cellular conditions. By comparing the sets of accepted gene pair relationships, information about which relationships are potentially 'universally' (among the studied conditions) valuable and which are uniquely valuable in each cellular state can be determined. Among the 'universally' accepted gene pair relationships, the scoring can be compared to elucidate changes in the potential importance of keeping each relationship very stable or less. Finally, the ranking of the accepted gene pair relationships, can display changes in relative priority changes across the cellular states. Absolute numbers, such as the actual scorings, can always be potentially affected by measuring noise and experimental variation present between, for example, sequencing runs. Therefore, ranking order, in that aspect can be less influenced as its nature is such that each gene pair relationships score is in relation to the rest of the accepted gene pair relationships in the sample group.

Two things differentiate RA analysis from commonly applied PE and MI. First, RA gives a two-dimensional comparison between cellular states, as it not only reports changes in stability (and thus changes in CV), but also whether the relationship has become non-compliant to the ratiometric definition. PE and MI can only report the change in their score of each gene pair relationship; they do not differentiate between a gene pair expression pattern that complies with a certain criteria and how strong/stable that relationship is. Second, there are scenarios where PE and MI would fail to detect informative changes in gene pair relationships (see figure 1).

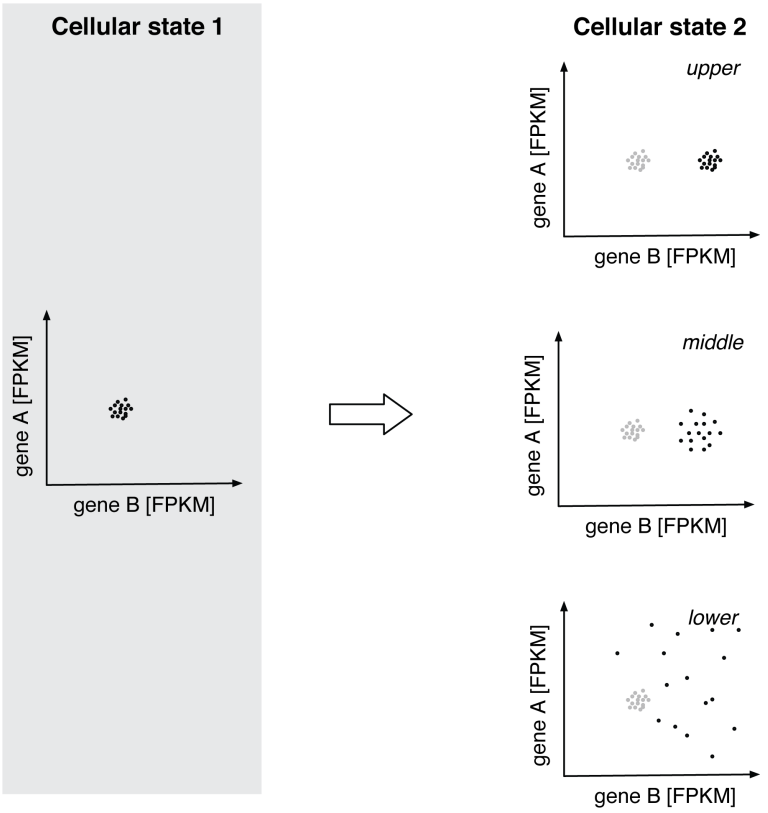


Figure 1: (Previous page) **Schematics of unique detection by RA.** The gene pair relationship in cellular state 1 is only reported by the RA method compared to PE and MI, as the latter cannot detect a gene pair combination with such low expression dispersion rate. When information from the second cellular state is added, here three different scenarios are given, the gene pair relationship remain undetected by PE and MI, but according to the RA method different outcomes would be reported. The upper graph indicates that the gene pair relationship has remained stable between the two cellular

states but it has shifted its ratiometric value from 1:1 to 1:2. The middle graph also exhibits the shift of ratiometric value but additionally it has loosened the stability of the ratio. The last, lower, graph, displays a scenario when the gene pair relationship has dissolved in the second cellular state. Note that these three scenarios show changes in a gene pair relationship that could be vital for understanding the molecular changes occurring in the two different states and moreover, they are undetected by the commonly used methods PE and MI.

The schematic example depicts the scenario where two cellular states are measured and in the first cellular state the gene pair relationship involves two genes with very low expression dispersion rates. Thus PE and MI do not detect it but RA reports it as a high-scoring relationship. When the second cellular state is analyzed, there are three possible scenarios where the gene pair relationship remains undetected by PE and MI. But the results indicate the information as of potential value, as it adds a possible gene pair relationship that is altered and thus could further the understanding of the differences between the two cellular states. The first scenario, upper graph, is depicting when the gene pair relationship has maintained the ratiometric stability, the CV is the same, but it has shifted the ratiometric value from 1:1 to 1:2. This can give indications of how the two different cellular states are in need of distinctive ratios between the two genes even if the relationship is equally strong in both conditions. The second scenario, middle graph, shows a similar situation where the ratiometric value has also changed from 1:1 to 1:2, but it has also decreased in ratiometric stability. This could possibly be an indication of changed cellular requirements between the two conditions, where not only the proportions between the two genes are distinct but also the importance of keeping the proportion, ratio, has decreased in the second state. The last scenario, illustrate a case where the gene pair relationship is not present at all in the second cellular state. Notice here that as PE and MI do not detect the gene pair relationship in the first cellular state

either, they would report this gene combination as not being in a relationship in neither state. In contrast, the RA would report the gene combination as being in a strict relationship in cellular state 1 and not in 2. At last, it should be mentioned, maybe the most obvious scenario but yet maybe also the easiest to miss, is the situation where no change is seen between the two cellular states. Meaning that the gene pair relationship is complying with the ratiometric definition and is highly stable with the same ratiometric value in both cellular states. Thus, it is a gene pair relationship possibly important in both states. Learning about these relationships in combination with those that become altered between the two states, can further the understanding of the differential of the two states.

Thus, by analyzing multiple-cellular state data sets, using RA, potentially new information about changes in gene pair relationships can be detected and enrich the understanding of the differential nature of the studied states.

Disease and control datasets

Studies involving disease-related data sets are worth discussing further as it is an area of gene expression profiling that might benefit from a RA analysis or at least being supplemented by one. Gene expression studies of a disease state are often designed as a comparison between a group of subjects affected by the disease and a complementary group of healthy subjects. The control group is desirably homogenous and thus RA could be a preferred analytical approach to determine the entire present cohort of gene pair relationships. Importantly, it most likely includes gene pair relationships containing genes with low expression dispersion rates. Subsequently, by comparing which ones of the

reported gene pair relationships that have been dis-regulated completely, thus rejected in the disease group as not complying to the ratiometric definition, or just decreased their ratiometric stability, it potentially could give a new revised assessment of the disease state. Specifically, which gene pair relationships and parts of cellular processes and pathways that have become dis-regulated and disintegrated, giving possible indications of either causes or effects of the disease condition. As PE and MI report gene pair relationships involving genes with low expression dispersion rates as non-correlated, this entire group of relationships would pass un-noticed with the consequence of reporting active cellular processes and pathways as only partially calibrated at the gene expression level.

An additional use of RA working with disease-related data is the possibility of detecting subdivision among the disease-affected subjects. There is increasing evidence that a disease can exhibit subdivision of affected subjects when the individual gene expression profiles are examined. See section below *Detecting subgroupings within a sample group* for further details.

Clustering

Many studies in biology currently involve clustering of some kind, be it, for example, hierarchical or k-mean clustering. Therefore, the application of RA to these kinds of analytical approaches will be discussed here, to both discern limitations and concerns about the fruitfulness and possible strategies.

Both PE and MI are well suited as the input matrix to clustering methods due to following reasons. First, the association strength for the gene pair relationships are determined within a set range, r^2 ranges between 1 and 0 for PE and I ranges from 2 and 0 for MI. Second, they are completely inclusive, which means that all gene pair relationships will be given an association strength. The same two reasons make RA less suitable. As RA does not have a fixed range of association strength, CV goes from 0 to infinity and therefore, the first approximation that has to be done is to set an upper bound, implying that above that threshold all gene pairs will be seen as not associated at all and automatically set to the cut-off value. This is most likely of a lesser problem compared to the second concern. RA only reports association strength for gene pairs that pass the stringency cut-off, Δ_{CV}, and thus can be described as using a ratiometric definition, which was shown in previous chapters to be a smaller portion (~1/4) of all possible gene pair combinations. This leaves the input matrix interspersed with a large number of empty values. As, notice that these are not of the same characteristics as the first group, where the CV if larger than the cut-off is simply put to the cut-off value. There is a difference between not passing the stringency level cut-off, Δ_{CV}, and passing it but exhibiting very pour stability, thus having a low strength (CV). Currently, there is no other available option then to put the rejected gene pairs values to the upper bound cut off value too, thus clumping the two categories together. This leads into the next concern, which is that RA is an approach that selects a sub group of interesting gene pairs and disregards the remaining gene pairs from the analysis. When the input matrix for a clustering study is constructed, the adding of one gene pair automatically adds all other gene pair combinations those two genes produce with all the genes already in the matrix. A

possible bias could occur as the rejected gene pair combinations will blur and maybe skew the results so that the true relationships will not have enough influence to make a proper clustering.

Clustering is a powerful tool to visualize large groupings of genes according to similar expression patterns. RA associations strengths could possibly be a valid input for such studies as long as the meaning of the RA results are kept in mind and the clustering results are interpreted based on how the adjustments of the input matrix was done.

Detecting subgroupings within a sample group

From investigating the highly ranked gene pairs from PE and MI, which were rejected by the stringency cut-off level, Δ_{CV}, used by RA, a new possible application for RA was revealed. The fact that a large portion of these gene pairs were showing expression patterns having a two-regime expression pattern, which furthermore were absent among the first couple of hundred examined top ranked RA gene pairs, demonstrates the possibility of detecting subgroupings within the sample group by combining one of PE or MI with a RA analysis. This type of analysis can be fruitful when studying large datasets where, for example, a certain degree of stochasticity among the samples are expected or suspected. Single cell data sets is one type of studies where there are indications of natural 'gene expression based' subdivisions among large cohorts of cells from one cell type (*1*). By applying PE or MI and RA separately and we can examine the gene pairs that RA rejects but PE (or MI) ranks highly. There is a probability that these gene pairs' expression patterns display the present subgroupings of the individual cells, and these

could then be isolated and analyzed by method of choice to determine their expression profile without the noise of the remaining cell samples.

Single cell studies

In this thesis a preliminary analysis of single cell data was conducted as a first test of RA for processing such type of data. Despite the current limitations of the single cell RNA-seq method, the results from the KEGG pathway analysis indicated a success of detecting at least the ribosome pathway for RA while PE and MI did not, see chapter 3 for details. Besides further the improvements of the single cell RNA-seq method itself, it is worth mentioning the possible requirements of the raw data imposed by RA on such type of data. The requirement most probably being imperative for a correct RA analysis is a sufficiently large sample group. This is likely even more important for single cell data sets as they have been shown to exhibit a certain degree of stochasticity among the cells' individual expression profiles (*1*). Basing a recommendation for sample group size, on a preliminary study run on an older B-cell lymphoblastoid cell line data set containing 70 samples (*2*), data not shown, the lower limit for a stable reporting of gene pair ranking by RA is around 15 samples. This is to say that, in a homogenous sample group using 15 samples is likely sufficient for proper results. Applying this to single cell data, it is recommended that a rough estimate or guess of the stochasticity is made, so that the sample group size would be preferable a sum of the hypothesized sub groupings frequency, such that the rarest sub grouping would be represented by at least 15 samples. For example, if there is reason to believe that a cell population, from which the single cells are randomly selected from, consists of 3 different expression profiles with the

concurrency frequency of 3, 5, and 2 out of 10. Then the minimum sample group size recommended, based on the pilot run performed in this study, would be ~80 samples.

Conclusion

The potential of RA was shown in this thesis by demonstrating how it produced a unique and biologically interesting ranking of gene pairs compared to PE and MI. Future uses of RA has been hypothesized here based on the reported findings in earlier chapters. The application of RA in cluster studies has its limitations due to RA's selective nature. The nature of how RA judges gene pair expression patterns renders it potentially powerful to discern previously undetected expression changes between multi-cellular states, be it tissue types, time lines or disease vs. control studies. In addition, the capacity of rejecting gene pairs exhibiting two-regime expression patterns makes RA, combined with either PE or MI, a promising approach to determine sub groupings within a sample group.

References

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183 (Aug 16, 2002).
2. J. K. Pickrell *et al.*, Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768 (Apr 1, 2010).

Supplemental Methods

Analytical Analysis of the CV

We explore several models to show the behavior of the cv's of the ratios. Essentially, under certain assumptions and conditions, it can be shown the that cv(x/y) and cv(y/x) approach each other. In particular, this equivalancy can be independent of the variation of x and y, as is shown in some simulations at the end of this supplement.

Given two genes g and h, let $\tilde{x}$ be the expression level of gene g and $\tilde{y}$ be the expression level of gene h. Then g and h are described as being in a ratio relationship if

$$\frac{\tilde{y}}{\tilde{x}} = c + u, \tag{1}$$

where c is the ratio of the expression levels of the two genes and u is an error term which is uncorrelated with any other variable. Then if x and y are the observed expression levels for g and h, one has that

$$x = \tilde{x} + u_x \quad \text{and} \quad y = \tilde{y} + u_y. \tag{2}$$

Since the sampling scheme in any single RNA-seq run can be presumed to be multinomial, and the accumulation of reads at any single gene is miniscule with respect to the total number of reads, a normal approximation to the distribution of this error term can be made under a Poisson sampling scheme, so that $\text{Var}(u_x) \approx x$ and $\text{Var}(u_y) \approx y$, with both of course having zero expectation.

Rewriting (1) as

$$\tilde{y} = \tilde{x}c + \tilde{x}u,$$

one can use the equalities in (2) to obtain

$$\begin{aligned} \tilde{y} &= \tilde{x}c + \tilde{x}u, \\ y + u_y &= (x - u_x)c + (x - u_x)u, \\ \frac{y}{x} &= c + u + \frac{1}{x}\left(u_y - cu_x - u_x u\right). \end{aligned}$$

Only the variance of the u (call it σ_u^2) is unknown so it is relatively simple to estimate this. Note that the above gives

$$\begin{aligned} \tilde{y} &= \tilde{x}c + \tilde{x}u, \\ 1 &= c\frac{\tilde{x}}{\tilde{y}} + u\frac{\tilde{x}}{\tilde{y}} \\ \frac{\tilde{x}}{\tilde{y}} &= \frac{1}{c} - \frac{u}{c}\frac{\tilde{x}}{\tilde{y}}, \\ \frac{\tilde{x}}{\tilde{y}} &\approx \frac{1}{c} - \frac{u}{c^2}, \\ \frac{\tilde{x}}{\tilde{y}} &= d + v \end{aligned}$$

the next to last step coming since $\tilde{x}/\tilde{y} \approx 1/c$ and setting $d = 1/c$ and $v = u/c^2$.[1] Estimation can then proceed as outlined above for $\tilde{y}/\tilde{x}$.

So $\tilde{x}/\tilde{y}$ will have a coefficient for the ratio which is the reciprical of the $\tilde{y}/\tilde{x}$ expression and has the same error term scaled by $1/c^2$. Thus if estimation is performed when this is the true model, one would expect that

$$\begin{aligned} \frac{\hat{\sigma}_v}{\hat{d}} &\approx \frac{\sigma_u}{c^2}\frac{1}{1/c} \\ &= \frac{\sigma_u}{c}. \end{aligned}$$

The above ratios are actually the coefficient of variations of the ratios $\tilde{y}/\tilde{x}$ and $\tilde{x}/\tilde{y}$. Thus the two coefficients should be approximately the same when the ratio regime describes the relationship between the data.

[1] In actual fact $\tilde{x}/\tilde{y} = 1/(c+u)$, not $1/c$, but since there is an estimate for the distribution of u from the regression of $\tilde{y}/\tilde{x}$, this approximation can be made more accurate. If the magnitude of the errors u/c are small, as is typically the case, this approximation will be good, as the variance term can be rewritten as $(u/c^2)(1/(1+u/c))$. Otherwise the actual variation may need to be modelled for accurate results.

We now use the Delta method (based on Slutsky's theorem (30)) to demonstrate asymptotic equivalence of estimators of the cv's under a ratiometric regime. This demonstration depends on $\mathrm{E}[y/x]$ being close to $\mathrm{E}[y]/\mathrm{E}[x]$, a condition that would seem to be satisfied by most ratiometric specifications, since $y/x = c$ implies $\mathrm{E}[x/y] = c$ but also that $y = cx$ so that $\mathrm{E}[y] = c\mathrm{E}[x]$ or $\mathrm{E}[y]/\mathrm{E}[x] = c = \mathrm{E}[y/x]$.

Define μ_x and μ_y ($(\mu_x, \mu_y) = \mu$) as the means of x and y, so that μ is asymptotically normal and has covariance matrix Σ. Then by the usual corollary to Slutsky's theorem, any function $f(x, y)$ is also asymptotically normal with mean $f(\mu_x, \mu_y)$ and variance $\nabla f(\mu)^t \Sigma \nabla f(\mu)$. Setting $r(a, b) = a/b$ and $s(b, a) = b/a$, the asymptotic means of these two functions are $r(\mu_x, \mu_y) = \mu_x/\mu_y$ and $s(\mu_x, \mu_y) = \mu_y/\mu_x$. Variances are given by

$$\begin{aligned}
\text{Var } (r(x,y)) &= \left[\begin{array}{cc} \frac{1}{\mu_y} & -\frac{\mu_x}{\mu_y^2} \end{array}\right] \Sigma \left[\begin{array}{c} \frac{1}{\mu_y} - \frac{\mu_x}{\mu_y^2} \end{array}\right]^t \\
&= \frac{1}{\mu_y^2}\left[\begin{array}{cc} 1 & -\frac{\mu_x}{\mu_y} \end{array}\right] \Sigma \left[\begin{array}{c} \frac{1}{\mu_y} - \frac{\mu_x}{\mu_y} \end{array}\right]^t \\
&= \frac{\mu_x^2}{\mu_y^4}\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t \\
\text{Var } (s(x,y)) &= \frac{1}{\mu_y}\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t
\end{aligned}$$

so

$$\begin{aligned}
\frac{\text{SD } (r(x,y))}{r(\mu_x,\mu_y)} &= \frac{\mu_x}{\mu_y^2}\left\{\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t\right\}^{1/2} \left[\frac{\mu_x}{\mu_y}\right]^{-1} \\
&= \frac{1}{\mu_y}\left\{\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t\right\}^{1/2} \\
&= \left\{\left[\begin{array}{cc} -\frac{1}{\mu_x} & \frac{1}{\mu_y} \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{1}{\mu_x} \quad \frac{1}{\mu_y} \end{array}\right]^t\right\}^{1/2}
\end{aligned}$$

$$\begin{aligned}
\frac{\text{SD } (s(x,y))}{s(\mu_x,\mu_y)} &= \frac{1}{\mu_x}\left\{\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t\right\}^{1/2} \left[\frac{\mu_y}{\mu_x}\right]^{-1} \\
&= \frac{1}{\mu_y}\left\{\left[\begin{array}{cc} -\frac{\mu_y}{\mu_x} & 1 \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{\mu_y}{\mu_x} \quad 1 \end{array}\right]^t\right\}^{1/2} \\
&= \left\{\left[\begin{array}{cc} -\frac{1}{\mu_x} & \frac{1}{\mu_y} \end{array}\right] \Sigma \left[\begin{array}{c} -\frac{1}{\mu_x} \quad \frac{1}{\mu_y} \end{array}\right]^t\right\}^{1/2}
\end{aligned}$$

This shows that the coefficients of variation for the ratio and inverse ratio are asymptotically the same. So any observed difference in the coefficients of variation are due to speed of convergence to the asymptotic approximation and the validity of the assumption $\mathrm{E}[y/x]$ being close to $\mathrm{E}[y]/\mathrm{E}[x]$. To illustrate how this might work in practice, we turn to some simple simulations.

We can use the following model (a variant of a common errors in variables setup) to illustrate some of the convergence issues. Define

$$\begin{aligned}
y &= (\tilde{y} + v) \\
x &= (\tilde{x} + u) \\
\text{cv(y/x)} &= \text{sd (y/x)/mean (y/x)} \\
\text{cv(x/y)} &= \text{sd (x/y)/mean (x/y)},
\end{aligned}$$

where $\tilde{y}$ and $\tilde{x}$ are assumed to have normal distributions with means 300 and 200, respectively, and v and u are error terms with zero means. We run simulations of 10,000 observations, perturbing the standard deviations of these and display the results in Table S1. Increased variation in u and v increases the cv's of x and y, obviously. But increased variation that follows a ratiometric pattern does not necessarily increase the cv's of the ratios, as is now shown.

We model explicitly ratiometric behavior (in the sense of Schnute), by simulating 10,000 observations where the mean of x is 200, with standard deviation 20, and 10,000 observations where the mean of x is 300, also with standard deviation 20. Similarly for y, 10,000 observations are simulated with a mean of 400 and 10,000 observations with a mean of 600.

Table S1: cv's as function of input variation

sd(x)	sd(y)	sd(u)	sd(v)	cv1	cv2
20	20	.01	.01	0.1225757	0.120245
20	20	.1	.1	0.1244791	0.1219522
20	20	1	1	0.1243477	0.1223783
20	20	10	10	0.1379601	0.1351505
20	20	20	20	0.1756124	0.1696806
20	20	40	40	0.3236766	0.2829277
40	40	.1	.1	0.2754183	0.2467305
40	40	.01	.01	0.2759232	0.249387

Thus the ratio of y/x is equal to 1.5 for each set of observations, but the standard deviations of x and y will be very large, since the difference of the two means of x is 100 and the difference of the two means of y is 150. The results are display in Table S2. The standard deviations of x and y are as expected and equal to what was described above. As we would

Table S2: cv's as function of input variation, Schnute ratio

sd(x)	sd(y)	sd(u)	sd(v)	cv1	cv2
100.99	150.79	10	10	0.1097509	0.1076288

expect from the Schnute model (which these simulation parameters fit), though, the cv's are low, lower, actually, than any in Table S1. These simulations would thus seem to fit the analytic results derived above–a ratiometric relationship implies cv's closer to one another than would be expected notwithstanding large variation in the marginal distributions.

Statistical test of CV(FPKM) groupings

In the paper an analysis of gene variation (as defined by the CV(FPKM) of the KEGG pathway genes is made. This analysis is performed by dividing the pathways into four groups: 1) pathways recovered by the RA approach with a wide margin compared to PE or MI; 2) pathways recovered similarly by all methods, but slightly better by the RA analysis; 3) pathways for which the three methods identify an approximately equal number of genes; and 4) pathways better recovered by PE and MI. It is concluded that the values of CV(FPKM) are lowest for pathways in the first group, and highest for pathways that are in the fourth group. We now provide evidence that the first and fourth groups are different than all the other groups and that the second and third groups are statistically indistinguishable.

To provide a statistical framework for this statement, we turn to linear regression and regress the CV(FPKM) of the gene on dummy variables, coded as follows: c1 is set to one if the gene is in group 1, 0 otherwise, c2 is set to one if the gene in group 2, 0 otherwise, c3 is set to one if the gene is in group 3, 0 otherwise, and c4 is set to one if the gene is in group 4, 0 otherwise (the constant term is of course not included in the estimation since $c1 + c2 + c3 + c4 = 1$, by definition). The regression is given in Table S3. All of these coefficients are highly significantly different than zero.

Table S3: Regression of CV(FPKM) on group membership dummy variables

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
c1	0.195424	0.004536	43.09	$< 2e-16$
c2	0.219751	0.003410	64.44	$< 2e-16$
c3	0.222762	0.004308	51.71	$< 2e-16$
c4	0.242154	0.005370	45.09	$< 2e-16$

The interpretation of this regression is that the average effect of being in a group is the coefficient of the group. Thus

the first column of Table S3 is the average CV(FPKM) of each group. We are interested in the null hypotheses that all groups have the same average CV(FPKM), and we do this by testing whether the coefficient of one group is statistically different from the coefficient of another group, for all six possible combinations (group 1 compared to 2, 3, and 4, group 2 compared to 3 and 4, and group 3 compared to 4).

To test whether two groups have the same CV(FPKM), we need to divide by the standard deviation of the difference of those two coefficients to obtain a t statistic. Since both coefficients are random quantities and have a covariance, that standard deviation is the square root of the sum of the individual variances for each coefficient minus twice the covariance of the two coefficients. All of these variance/covariance quantities can be obtain from the covariance matrix of the linear regression. These t-values can be displayed in matrix form, as they are in Table S4 (here Second Group is the t-value of the coefficient of that group being subtracted from the coefficient First Group). These t-values are all very high, except

Table S4: T values for difference of CV(FPKM) for groups

		First group			
		1	2	3	4
Second	1	.	4.28	4.36	6.65
Group	2	.	.	0.54	3.52
	3	.	.	.	2.81
	4	.	.	.	.

for the t-statistics representing the difference between groups 2 and 3. Since we have a sizeable number of observations for each group, a normal approximation to the t statistic is valid, the t-values for the difference between groups 1 and the remaining groups are all significant at well below the .001 level, and for between group 4 and the others groups, at below the .005 level. Groups 2 and 3, with a t-value of 0.54, are not significantly different at the .05 level (given the magnitude of these t-values, it is clear that these conclusions obtain even if a Bonferroni type procedure is applied). In particular, our statement that group 1 has a smaller CV(FPKM) than the other groups and group 4 has a larger CV(FPKM) than the other groups is statistically supported.

Appendix B: **The distributions of the Pearson correlation coefficient *r*, R^2, Δ_{CV}, and I**.