Multivariate information maximization yields hierarchies of expression components in tumors that are both biologically meaningful and prognostic

Shirley Pepke[1] and Greg Ver Steeg[2]

1 Lyrid LLC (South Pasadena, California, USA); **2** University of Southern California, Information Sciences Institute (Marina del Rey, California, United States)

Abstract

De novo inference of clinically relevant gene function relationships from tumor RNA-seq remains a challenging task. In this work we show that Correlation Explanation (CorEx), a recently developed machine learning algorithm that optimizes over multivariate mutual information, achieves significant progress toward this goal. CorEx utilizes high dimensional correlations for a principled construction of relatively independent latent factors that "explain" dependencies in gene expression among samples. Using only ovarian tumor RNA-seq, CorEx infers gene cohorts with related function, recapitulating Gene Ontology annotation relationships. CorEx is able to identify latent factors that capture dependencies in groups of genes whose expression patterns correlate with patient survival in ovarian cancer. Some inferred pathways such as chemokine signaling and FGF signaling have been implicated previously in chemo responsiveness, but novel survival-associated groups are identified as well. These include a pathway connected with the epithelial-mesenchymal transition in breast cancer that is regulated by a potentially druggable microRNA. Further, it is seen that combinations of factors lead to a synergistic survival advantage in some cases. Comparison to normal ovarian tissue exhibits substantial differences between cancerous and non-cancerous samples related to a variety of cellular processes. In contrast to studies that attempt to partition patients into a small number of subtypes (typically 4 or fewer), our approach utilizes subgroup

information for combinatoric transcriptional phenotyping. Considering only the 66 gene expression groups that are found to have significant Gene Ontology enrichment and are also small enough to indicate specific drug targets implies a computational phenotype for ovarian cancer that allows for 3^{66} possible patient profiles, enabling truly personalized treatment. The findings here demonstrate a new technique that sheds light on the complexity of gene expression dependencies in tumors and could eventually enable the use of patient RNA-seq profiles for selection of personalized and effective cancer treatments.

Author Summary

Quantifying gene expression dependencies from high throughput data is a challenge given the complexity of interactions and the noisiness of even the best available data. We introduce a new method, Correlation Explanation, that uses information theoretic principles to understand gene expression data. The algorithm captures relationships among many genes by assuming they are related via connection to a common hidden explanatory factor. We apply correlation explanation to large scale expression data from 420 ovarian tumors to construct multiple layers of explanatory factors. The 5000+ genes analyzed in this study can be combined in an astronomical number of ways in principle. By searching this large space in an efficient way, the correlation explanation algorithm is able to gradually focus in on gene cohorts with strongly interacting expression patterns that are relatively independent between groups. We find that many of the groups of genes correspond to specific cellular biological processes on multiple levels, some of which can be implicated in tumor progression, metastasis, and patient survival. Putting these factors together allows each patient to have a unique summary profile of gene expression patterns that could someday be used to guide treatment decisions.

Introduction

In recent years, innovations in high-throughput sequence-based assays that allow for the rapid and cheap interrogation of the genomes and transcriptomes of individual tumors at an unprecedented level of detail have given rise to fresh hope for significant progress toward understanding of complex cancer biology and effective treatments. **[1]** Inference of clinically relevant gene expression networks based upon RNA-seq profiles could provide crucial information in this context. The tumor transcriptome has the potential to provide a view of the functional network alterations that result from a variety of causes, including genetic mutations, copy number variations, epigenetic states, the tumor microenvironment, immunophenotype variations and even transient dynamic effects. In principle, machine learning could be used to leverage the rich information that gene expression provides about each individual to infer tumor-specific features that may guide treatment decisions. Unfortunately, several challenges limit the usefulness of existing methods. The data is high dimensional and the number of samples is small. For this reason, many methods focus on simple linear, pairwise relationships, limiting their generality. Clustering techniques typically use too broad a brush, pigeon-holing patients into a small number of clusters when understanding the interplay of multiple independent factors is crucial. Finally, interpretability of results is essential in this context in order to provide insight into biological mechanisms that could motivate new treatments.

To overcome these limitations, in this work we apply a recently developed machine learning algorithm, CorEx **[2,3]** , to the analysis of RNA-seq transcriptomes. Multivariate and nonlinear dependencies in the data are captured by total correlation, a generalization of mutual information to multiple variables. Intuitively, CorEx uses an optimization scheme in order to efficiently infer latent variables that explain as much of the dependence in the data as possible. Each latent factor discovered by CorEx is a function of some (possibly overlapping) subset of the input variables (genes) with high total correlation, and reflects a relatively independent axis of variation in comparison to the other latent factors. Further, each latent factor defines a separate probabilistic

clustering of samples, and thus provides a factor-specific stratification of tumors according to expression patterns of subgroups of genes across the tumor samples. We can apply CorEx again to these inferred latent factors to discover a hierarchy of dependencies among genes.

Ovarian cancer is the deadliest gynecologic malignancy due to its typically late stage at diagnosis, however it is responsive to a wide variety of therapies. In addition to the standard platinum/taxane combination for frontline, active second line therapies include pegylated doxorubicin, topoisomerase inhibitors, angiogenesis inhibitors, anti-metabolites and parp inhibitors **[4,5]** . Early data indicates that ovarian cancer may also be an immunogenic tumor for which checkpoint inhibitors, vaccines, and other immunotherapies may provide meaningful benefit **[6–9]** . Thus ovarian cancer tumors can be expected to display a rich variety of gene expression patterns related to outcomes under various treatments. On the other hand, current second line treatments are effective for less than 30% of patients, are not curative, and come with often serious cumulative toxicities. With the exception of DNA repair pathway mutations for PARP inhibitors, there are no genetic markers currently in use to predict therapeutic response. Thus new ways to match ovarian cancer patients to effective therapies based upon individual tumor biology will likely yield substantial gains in survival.

In this work we show how the use of CorEx to analyze ovarian cancer RNA-seq profiles uncovers relatively small cohorts of genes that exhibit coherent patterns of differential expression among tumors and that, in many cases, these cohorts are associated with biological pathways as indicated by database annotations of the genes. The upper levels of the hierarchy of CorEx gene clusters align with high level GO annotations and in some cases relate clusters of regulatory genes including noncoding RNAs to protein coding elements. Patient stratification with respect to individual gene groups shows differential survival associations in groups associated with known factors in chemo-responsiveness. Even more significantly, stratifications based upon combinations of group factors display synergistic survival associations. We highlight discovery of a component containing genes

related to epigenetic control of stem cell characteristics that have not been previously observed in ovarian tumors. Additionally, when breast cancer data not used for training is stratified relative to this latent factor, a survival association is also observed. Finally, we compare the group results in ovarian tumors to expression profiles in normal ovarian tissue and demonstrate substantial differences in some cases. The strong results presented here demonstrate that CorEx is a powerful tool for dissecting therapeutically relevant gene expression relationships in tumor RNA-seq, and suggest it may be usefully applied to infer correlation structure from other types of complex biological data **[10]** .

Results

CorEx infers gene expression relationships from tumor RNA-seq profiles

CorEx can be used to construct a hierarchy of latent factors that "explain" correlated gene expression among samples with respect to subsets of genes **[2]** (**Figure 1**). It does this by optimizing sample and gene clustering with respect to total correlation (TC) **[11]** , an extension of mutual information to multiple variables. In particular, the algorithm seeks to minimize the total correlation remaining among $\boldsymbol{X_G}$ variables after conditioning upon the $\boldsymbol{Y}$ latent factors. The CorEx optimization has several desirable properties for high throughput gene expression data: (1) it scales linearly with the input data, (2) genes can participate in multiple groups to an extent determined by optimization with respect to a parameter $\boldsymbol{\alpha}$, (3) it is robust to missing data, (4) uncorrelated random noise is automatically ignored, (5) hidden units "compete" to explain correlations. Details of the algorithm and implementation can be found in the refs and the Methods section.

In the context of RNA-seq tumor profiles, the input to the algorithm is a training matrix consisting of z-scores for each gene among all the samples. In this work the gene set is taken to be a subset of about 5300 of those available for The Cancer Genome Atlas (TCGA) cohort (the top 3000 by variance plus those that are mutated in at least N samples). Extending the computation to encompass the whole exome should be easily achievable in a cloud setting. The CorEx output is a set of m gene groups with the genes

ordered within a group according to their estimated mutual information (MI) with respect to the group's latent factor and a probabilistic assignment of samples to the k factor labels. Thus for each sample j, group i, and latent factor label k (here $k \in \{0,1,2\}$), CorEx provides an estimate of $\boldsymbol{P(Y_i=k \mid X_1^j, X_2^j \ldots X_{n_i}^j)}$ where X_g^j denotes the expression of gene g in the jth sample. For simplicity, one can assign each sample to a single factor label according the the maximum over the label values of the conditional probabilities. In this work, the $\boldsymbol{P(Y=k \mid X_1, X_2 \ldots X_n)}$ for each sample is converted to a single score, f_{ji} that indicates sample j's congruence with the factor i label values representing either very high or very low gene expression. This score provides an ordering of the samples with respect to gene expression. Heat maps of genes within groups ordered by MI versus samples ordered by f_{ij} show how groups with the greatest TC display strong correlations among many genes, with the clearest relationships being between genes with high group MI, whereas lower TC groups show noisier correlations and/or few genes. (**Figure 1** and Supplementary Fig 1 of heat maps, and the Supplementary data files).

Sharing of genes among groups is an integral part of the algorithm and, in fact, CorEx optimizes each latent factor's group membership by updating a parameter, $\boldsymbol{\alpha}_{gi}$, as the optimization proceeds. This gene sharing is desirable for a couple of reasons. For one, it known that gene products often perform multiple roles with the cell and therefore may participate in multiple interaction networks. For another, given that TC increases with the number of genes (all other things being equal), it is easier for the algorithm to find very small coexpression cohorts that participate in larger networks when they can be manifested as similar large network groups with different sub-clusters ranked most highly. In this application, genes tend to be associated with several groups on average, however typically only one to four at relatively high MI values.

Due to the complexity of the search space, the same gene groupings are not guaranteed to be found in different runs. However, comparison of gene lists between runs demonstrates that groups with high TC per gene are very likely to be reproduced. Additionally, we find that a Bayes shrinkage prior over the factor probabilities increases

reproducibility of groups with low TC, at the cost of possibly suppressing some real but weak signals (Supplementary Figure 3). In this work we fix the number of latent factors in layer one to 200. As one moves down the list of groups according to total correlation, confidence in the clusters decreases though there appear to be still biologically meaningful clusters at the lower end, suggesting that training on a greater number of samples or taking into account additional biological information will be useful. The latter is investigated in the following sections.

CorEx gene groups are enriched for protein-protein interactions, GO, and KEGG annotations

A majority of the CorEx gene groups show enrichment for genes with specific biological function as indicated by the enrichment with respect to protein-protein interactions (PPI), Gene Ontology terms, and KEGG pathway annotations (summarized in **Figure 2**). Multiple groups are found related to the mitotic cell cycle and its regulation, ECM organization and interaction, chromatin modification and DNA packaging, electron transport, and regulation of map kinase and growth factor pathways, among others. Genes that code for proteins that participate in larger functional complexes such as ribosomes are also grouped together. Some of the particular PPI graphs are shown in **Figure 2** where it can be seen that CorEx finds networks of genes related to specific immune processes such as Type I interferon signaling, antigen processing and presentation, immune cell activation, migration, and aggregation. In fact, a plurality of found groups have predominantly immune system-related annotations. Overall, the number of groups enriched for annotations as well as their diversity and specificity support the usefulness of CorEx for parsing the information about differentially activated biological processes from the RNA-seq profiles at an unprecedented level of detail. (Suppl. Figure 2 and Suppl. data for all interaction graphs and listing of terms)

The CorEx groups often display especially clear PPI enrichments with relatively little noise, though all three types of annotations typically exceed threshold together. Groups with greater total correlation are more likely to yield significant database annotations.

However, because groups with high TC tend to be larger, groups with somewhat lower TC tend have more specific functional annotations at low p-values. It can be seen that many groups even with quite low TC sometimes have associated annotations. Capturing these groups with relatively weak signals that nonetheless appear biologically coherent was part of the motivation for specifying such a large number of factors. The database annotations offer some guidance as to which may merit closer examination.

Some of the power of CorEx in this context comes from lifting the restriction of considering initially only pairwise gene interactions, common in other approaches **[12–14]** . As a concrete example of the impact of this less restrictive inference method, one need only examine the pairwise relationships of the top ranked genes in one of the CorEx groups, pictured in **Figure 3** . The colors of the scatter points indicate the CorEx label value for each tumor sample. In the figure it is clearly evident that the top genes have only weak pairwise expression correlations, yet the clouds of similarly hued points remain coherent across the various plots showing their higher dimensional correlation. Examination of the expression heat map for group 106 (Supplementary Figure 1) also shows clearly that, while any individual gene appears somewhat noisy, the trend across the whole set is easily discernible. The full set of genes from this group exhibit PPI enrichment as well as significant enrichment for the KEGG pathways: ‘pathways in cancer’, ‘Wnt signaling pathway’, ‘proteoglycans in cancer’, ‘basal cell carcinoma’, and 'Hippo signaling pathway’. CorEx detects this important set of expression relationships that would almost certainly be overlooked based upon pairwise correlations. Additionally, the sensitivity for groups that exhibit stronger pairwise correlation for the top genes can be increased when those signals are reinforced with signals from lower-ranked genes exhibiting weak pairwise correlations but high TC relative to the group as a whole.

Higher level group clusters show coherent GO annotations related to biological function

Despite the fact that no information other than RNA-seq gene expression was used in constructing the CorEx latent factors, higher level factors aggregate gene groups with related Gene Ontology enrichments. Hence the hierarchy of CorEx latent factors appears to reflect biological organization on multiple levels. Some of these higher level functional relationships are highlighted in **Figure 4** . Clustering of Gene Ontology terms in the groups reveal layer one and higher clusters for immune signaling, extra-cellular matrix organization, immune cell activation, mitotic cell cycle, microtubule-related processes, and electron transport/chromosome organization. While in some cases this is due to redundancy of membership among the layer zero gene groups (though the genes are ranked differently within the groups, the GO analysis here did not account for relative importance), in others there is a clear distinction in gene membership and function on the finer scale. For example, both groups 31 and 110 are incorporated into the large immune-related cluster near the center of Figure 3A. However, these groups have both different associated GO annotations and zero genes in common. Group 31 has gene ontology terms associated with immune cell activation (e.g. GO:45321 for genes EBI3, Cd79A, BATF, FCER1G, CCL5, EOMES, CD3D, Cd7, LCP1, ZNF683, CD2, ITK, THEMIS, B2M), adhesion (GO:7159 for genes EBI3, RAC2, BATF, CCL5, EOMES, CD3D, Cd7, LCP1, CD2, ITK, THEMIS, B2M), and migration (GO:2687 for genes SELL, RAC2, CCL4, CXCL13, CCL5, CXCL11, CXCL9, and CCL8). In contrast, group 110 primarily has terms associated with inflammatory response (GO:6954 for genes IL4R, MEFV, TGFB1, NLRP3, SIGLEC1, C5AR1, CD163, PIK3CG, NLRC4, TLR4, TNFRSF1B), apoptotic process (GO: 6915 for genes TGFB1, HGF, ABR, C5AR1, MKL1, NICAL1, MAP3K5, PIK3CG, INPP5D, NLRC4, TRAF1, TLR4, TNFRSF1B), and cell death (GO:43067 for genes STK10, HGF, ABR, SIGLEC1, C5AR1, MKL1, MICAL1, MAP3K5, PIK3CG, INPP5D, NLRC4, NCF2, TRAF1, TLR4, JAK3). Thus higher level clusters such as immune response are parsed at the zero layer according to specific biological sub-networks.

A particularly striking example of the kind of information that can be gleaned from the layer one factors is shown in Figure 4b . One can see a layer one factor (node) that joins four clusters of ribosomal proteins. The fifth layer zero cluster is comprised of a set of non-coding RNAs related to cancer metastasis through regulation of growth arrest and protein folding. Thus a cluster of non-coding regulatory factors is linked to the downstream interactions of proteins they influence. While the particular information highlighted here is already known, the presence of these meaningful relationships suggest that novel relationships can, in principle, be discovered by CorEx. This is more likely to happen when many thousands of samples are available to allow less common relationships and perhaps even tumor-specific rewiring of interactions to be seen.

Patient stratification based upon combinations of latent factors yields differential survival associated with tumor-activated pathways

The continuous labels for the CorEx latent factors can be used as predictors of survival under a Cox proportional hazard model **[15]** . A subset of the CorEx small to intermediate-sized gene groups was selected according to known biological significance (Gene Ontology enrichment) and used to stratify patients according to relative risk of death under a Cox proportional hazard model for the factor scores. We first constructed models for each factor individually. When this is done and patients are stratified roughly into thirds according to low, intermediate, and high relative risk of death, several gene expression groups appear to be associated with differential survival (as judged by the p-value for the difference between the two extremal Kaplan-Meier survival curves). (Figure 5a and Supplementary Figure 4a) False discovery rates (FDRs) were also estimated by randomly permuting the sample scores for each latent factor (shuffling each column) in the score matrix. The top single factor groups appear somewhat weak by this measure as the values are strongly limited by the small number of patient samples versus the number of potentially relevant networks. Using this analysis, the top latent factor group, related to immune chemokine and interferon signaling, yields an FDR level close to .3. Genes with large MI with respect to this group factor include among others

CXCL10, CXCL11, CXCL9, CXCL13, ISG20, GBP (1-5), OASL, HERC5, STAT1, IFIT1, IFI35, TRIM22, and JAK2. Other groups that exhibit survival associations encompass additional immune system factors, fibroblast growth factor regulation, transforming growth factor beta signaling, map kinase regulation, Wnt signaling, Jak-STAT signaling, and cellular metabolism.

A perhaps even more compelling finding is that combinations of these statistically weak factors yield increased significance as pairs relative to a randomization test for paired factors, with 16/21 combinations yielding FDRs less than about .35. (Figure5b) This suggests synergy of individual factors in terms of impact on survival. The pair of factors that show the greatest association with survival is immune system cytokine signaling combined with regulation of the fibroblast growth factor pathway. These two general features are hypothesized to be associated with long term survival in ovarian cancer from other studies, the latter due to its relationship to chemo resistance **[16,17]** . The advantage here with respect to these previously known prognostic features is that each patient is assigned a relative risk according to activation of related pathways and therefore shows how they may combine to influence survival, indicating patients who might benefit from either single or simultaneous treatment with immune system modulation or FGF pathway inhibitors. Additionally, the fine parsing of overall immune system function by the sub-networks present in different groups has the potential to discriminate among different immunotherapies in terms of likely efficacy on a personalized basis. Overall, most of the top survival-associated combinations involve immune-related factors, emphasizing the primacy of the immunological milieu in influencing ovarian cancer survival rates in this cohort under the standard treatment protocols. The combination of a factor related to Wnt signaling and one for FGF signaling also appears predictive of survival in a subset of patients. Wnt signaling has also been associated with clinical prognosis in ovarian cancer. **[18]** Other factors and combinations

of groups show narrower survival differentials. We expect that some of these play important roles for a relatively smaller proportion of patients and will gain significance once a greater number of tumor RNA-seq samples are available for analysis.

It should be emphasized that in this retrospective analysis of a patient cohort that was treated prior to 2010, we are able to analyze the survival impact only with respect to standard carbo/taxol therapies as no targeted therapies were widely available or approved. Cell line and xenograft experiments or analysis of tissues from other clinical trials will likely bring to the fore other factor groupings that are associated with response to recently approved or experimental therapies, e.g. angiogenesis inhibitors, parp inhibitors, and apoptosis signaling modulators to name a few. **[19] [20] [21]**

CorEx discovers differential expression in ovarian tumors of EMT-related genes previously implicated in breast cancer stemness and metastasis

A survival differential at longer times only is seen for one CorEx gene expression factor (denoted G159). This suggests that this factor may influence survival in a way that is not related to initial carbo/taxol response per se. Detailed examination of the Gene Ontology annotations for this latent factor cluster revealed genes associated with microRNAs in cancer (6 genes), chromatin modification (8 genes), regulation of transcription from RNA pol II promoter (10 genes), stem cell differentiation (5 genes), and regulation of Wnt signaling (5 genes). Further, when this factor is combined in a survival model along with the top single factor within the same CorEx run (G103, immune cytokine signaling), the resulting stratification yields a more significant survival association than any other pair of factors from that run. (Figure 6a)

Intrigued by the possibility of a factor related to something more general than response to a specific chemotherapy, we sought to validate its significance using the TCGA breast cancer data. The reasoning behind this was that while breast tumors exhibit some similarity on a molecular basis to ovarian tumors, there are also significant differences. Earlier detection and a greater variety of chemotherapeutic options means survival will

depend largely upon factors not directly related to carbo/taxol response. We calculated factor scores for TCGA breast cancer tumor RNA-seq samples across the factors determined by the OvCa training samples (i.e. we did not train a new set of latent factors, but mapped the breast cancer samples to the factors learned from ovarian tumors). Using these scores to fit a stratified model with respect to the G159 latent factor yields a survival differential for BRCA patients as well.(Figure 6b) . Subsequently we discovered that the genes in this group have important functions within a network that was recently studied by Song et al. **[22]** in preclinical models of breast cancer, where it was shown that TET3 controls expression of DNMT3A, a DNA methyl transferase that exerts specific influence on the chromatin state related to stemness. (Figure 6c) Further, the network is implicated in the progression of breast cancer via the epithelial-mesenchymal transition, invasion, and metastasis. Mesenchymal-like properties of tumor cells and EMT-associated features in general have been associated with poor prognoses in ovarian cancer **[23]** , though this particular network has not been previously implicated. Importantly, this network was shown to be controlled by a microRNA acting on TET3, thus providing a novel druggable target. Observation of this mechanism in ovarian cancer tumors suggests not only a new biological driver for recurrence but also the possibility for selection of ovarian cancer patients who might benefit from experimental miR-22 modulators being developed for breast cancer.

Some latent factors can be used to distinguish between normal and cancerous ovarian tissue

We mapped RNA-seq profiles from normal ovarian tissue samples against the factors learned from the tumor samples and identified those that showed the greatest differences between the two types of samples. Many factors that exhibited large differences were found related to the mitotic cell cycle and replication, as would be expected when comparing cancerous to non-cancerous tissues. There were also several other factors that showed large distinctions and appeared related to a variety of other biological processes including ones for development and regulation of cell differentiation, neurogenesis,

respiratory electron transport chain, sex differentiation, cell migration, and inflammatory response. Histograms of the factor scores for representatives of these groups are shown in **Figure 7** . These findings highlight some commonalities of pathway dysregulation among ovarian tumors. Further, several of the groups contain genes that can be targeted with existing drugs **[24]** . This presents the possibility that repurposing of existing drugs based upon this sort of data may provide new therapeutic options for a great many ovarian cancer patients, though the implications within these particular gene expression cohorts need to be further elucidated.

Discussion

In cancer and other diseases, high-throughput data combined with computational phenotyping i.e. the algorithmic distillation of observables that explain and allow prediction of the interaction of the course of disease and treatment, is poised to transform medicine as we know it. In this work we have presented a novel approach toward the goal of an RNA-seq -based computational phenotype. The method of correlation explanation leverages multivariate mutual information to infer complex hierarchical gene expression relationships directly from RNA-seq transcription levels, and the groups of genes thus obtained correspond to distinct cellular subnetworks as indicated by Gene Ontology annotations and known regulatory relationships. CorEx does this efficiently and without requiring any form of prior knowledge. The hierarchy of factors is shown to facilitate understanding and interpretability by relating networks at multiple levels. Though CorEx allows sharing of genes among groups, it also attempts to maximize overall independence between the groups in terms of expression patterns, with factors at higher levels being progressively less dependent. This leads to the ability to pick out groups from different upper level factors that may combine synergistically, and this is seen in the survival associations for some combinations of factors. In addition, tumor-specific assessment of differential transcriptional activation of the gene cohorts is

intrinsic to the method. These results show how the observation of differential activity in these gene networks can have great clinical impact by enabling selection of patients most likely to benefit from particular therapies and/or combinations of therapies.

The use of RNA-seq, rather than microarray data, is essential for detailed inference of the transcriptional computational phenotype due to its superior dynamic range, sensitivity, and inherent whole transcriptome measurement capability compared to microarrays. Many past studies have focused on the use of microarray data, however RNA-seq should have substantial advantages for discovery of cancer-specific network interactions and this is supported by studies comparing the two **[25,26]** . Some genes in the expression cohorts found here are not even measured in microarray experiments, often because their function is inadequately characterized and thus don't justify inclusion of corresponding microarray probes. Also, uncommon transcript variants that arise from unstable cancer genomes cannot be detected with microarrays, leading to misleading expression values. RNA-Seq data captures the effects of such alterations in ways that allow for their discovery and more accurate quantification **[27–29]** . The results presented here from such a relatively small patient cohort are extremely encouraging and clearly show that it is possible to extract a great deal of information from just a few hundred whole exome tumor transcriptomes and makes a strong case for obtaining larger numbers of samples in this context.

Early work on the clinical implications of large scale gene expression data in cancer included direct correlation of gene expression and responses of cell lines to cancer agents **[30]** . More recently, gene co-expression analyses based upon pairwise correlation or mutual information have been applied to infer networks and regulatory modules **[13,31–33]** , some of which are associated with cancer progression and prognosis. The number of modules found e.g. in breast cancer is markedly fewer (10-20 in the cited studies) and this is likely due to both the algorithms' decreased sensitivity and the use of microarray data rather than RNA-seq **[34,35]** . Regulatory module inference in these cases often relies upon the application of graph theoretic techniques such as maximal clique

identification to identify densely connected groups of genes, with special considerations to allow overlaps. CorEx, in contrast, constructs a graph with edges connecting multiple genes to latent factors rather than to one another and overlaps are intrinsic to the inference process. The approach in **[36]** is to define attractor metagenes using mutual information and, though relatively more restrictive in many ways and heuristic, is the most similar in spirit to CorEx. As far as we are aware, our work with CorEx demonstrates the first successful use of multivariate mutual information in the context of high throughput gene expression data. The principled approach of CorEx and also the addition of data-guided smoothing via the Bayes shrinkage prior allows for continuous improvement as more data is generated. These properties combined with favorable scaling suggest that CorEx can be widely adopted to analyze and interpret complex gene expression data in many contexts.

CorEx both provides a high level summary of the content of groups of differentially expressed genes and parses them out into particular groups for potential experimental and therapeutic purposes. Therefore it is an umbrella method that can encompass the particular findings of other studies on specific prognostic features of patient tumors. For instance, the original TCGA analysis of the ovarian tumor expression data used non-negative matrix factorization to identify a partitioning of genes and patients into a few subtypes **[1]** . Features of the four subtypes are clearly reflected in some of the higher level CorEx nodes such as that linking immune response factors. In contrast, here the functions broadly related to the putative subtypes have been parsed more finely on the lower level to reveal specific tumor-activated pathways and druggable targets. This allows one to make hypotheses regarding the interaction of functional gene modules in tumor development and progression. Further, the impact of expression modulation of such relatively smaller groups of genes may be more easily tested in cell lines and xenograft models.

There are many possible extensions and further applications of the methods described here. While mRNA provides a functional readout of many different causal processes, other types of salient information, e.g. somatic mutations, copy number variation, or chromatin marks related to gene activation or repression can be easily correlated with the gene expression groups at the higher level to directly link known biological causes to gene expression cohorts. This information can then point to specific mechanisms active in individual tumors or strengthen the RNA insights with the integration of complementary information. Though beyond the scope of this initial work, pan-cancer analysis is obviously of great interest, and CorEx's favorable scaling behavior should allow for straightforward application in that context. Additionally, CorEx can be applied to other TCGA cancer types alone where one is likely to see different networks implicated.

Methods

CorEx Algorithm

Multivariate mutual information, or total correlation (TC), quantifies the amount of dependence among a set of variables **[11]** . TC is zero if and only if the variables are all independent. If we could enumerate all of the underlying causes of dependence in the data, then the TC conditioned on these factors would be exactly zero. In other words, these latent factors ``explain" the dependence in the data. The recently introduced method of Total Correlation Explanation (CorEx) is an information-theoretic optimization that reconstructs latent factors that explain as much of the dependence in the data as possible **[2,3]** .

The random variables, $X_1, \ldots, X_n$, written as X for short, represent gene expression values. We assume that the probability of a sample, x depends on some unknown distribution over X which we write as $p(X = x)$. Total correlation can be defined in terms of the Kullback-Leibler divergence, D_{KL}, or in terms of H, the Shannon entropy . Intuitively, it represents the distance between the true distribution, and the distribution under the null hypothesis that all variables are independent.

$$TC(X) \equiv \sum_i H(X_i) - H(X) = D_{KL}\left(p(X = x) || \prod_i p(X_i = x_i)\right) \quad (1)$$

The principle behind CorEx is to search for a small set of latent factors, $Y = Y_1, \ldots, Y_m$, so that the TC between the original variables is minimized when conditioning on those factors. The conditional TC can be defined as $TC(X|Y) = \sum_i H(X_i|Y) - H(X|Y)$. While our goal is to search for latent factors that minimize this expression, for convenience, we actually focus on the equivalent task of maximizing reduction in TC after conditioning on Y. We refer to $TC(X; Y) = TC(X) - TC(X|Y)$ as the correlation in X that is explained by Y and we seek to maximize this expression.

We let each learned latent factor take different discrete values $Y_j = 1, \ldots, c$. Each discrete value in this context corresponds to a pattern of correlated expression among some subset of genes that are strongly connected to the latent factor, e.g. $Y_j = 0$, 1, and 2 might correspond to relatively low, neutral, and high expression. We found that stratifying into three groups for each latent factor was adequate for capturing relationships without causing problems with estimation. For a given sample of data, $X = x$, we should get a probability distribution over possible values for each latent factor j, $p(Y_j = y_j|X = x)$. At the beginning, we start with a random solution, for each sample of data, we set a random probability distribution over latent factors. What makes our optimization tractable is that we can write our solution for $p(Y_j = y_j|X = x)$ in a closed form that defines an iterative update scheme that is guaranteed to get us progressively closer to the optimal solution. The solution depends on two types of parameters. First, $a_{i,j}$ are weights between zero and one that determine how much Y_j depends on X_i. Second, $p(X_i = x_i|Y_j = c)$ is the marginal probability of observing $X_i = x_i$ given an observed label $Y_j = c$. We take this distribution to be a normal distribution with mean, $\mu_{i,j,c}$ and variance, $\sigma_{i,j,c}$. The parameters, $a_{i,j}$, $\mu_{i,j,c}$, $\sigma_{i,j,c}$, along with the class weights, can be updated at each time t according to the following rules:

$$
\begin{aligned}
\alpha_{i,j}^{t+1} &= I^t(X_i;Y_j|Y_1,\ldots,Y_{j-1})/I^t(X_i;Y_j) \\
p^{t+1}(Y_j=c) &= \frac{1}{N}\sum_l p^t(Y_j=c|X=x^l) \\
\mu_{i,j,c}^{t+1} &= \frac{\frac{1}{N}\sum_l p^t(Y_j=c|X=x^l)x_i^l}{p^{t+1}(Y_j=c)} \\
\sigma_{i,j,c}^{t+1} &= \sqrt{\frac{\frac{1}{N}\sum_l p^t(Y_j=c|X=x^l)(x_i^l-\mu_{i,j,c}^{t+1})^2}{p^{t+1}(Y_j=c)}}
\end{aligned}
\tag{2}
$$

In practice, we alter the first rule to speed up computation and we devised a novel Bayesian estimate for the last two update rules that improves results on datasets with a small number of samples. Details of these revised updates are described in Supplementary Methods.

Once we have estimates at time step t of $a_{i,j}{}^t$ and $p^t(X_i = x_i | Y_j = y_j)$, then we can update the distribution of latent factors to get a new solution that is guaranteed to increase our objective. The solution has the following form.

$$
p^{t+1}(Y_j=y_j|X=x) = \frac{p^t(Y_j=y_j)}{Z_j(x)}\prod_{i=1}^{n}\left(\frac{p^t(X_i=x_i|Y_j=y_j)}{p(X_i=x_i)}\right)^{\alpha_{i,j}^t} \tag{3}
$$

In this expression, $Z_j(x)$ is a normalization constant. Now that we have update rules for the probabilistic labels, and for the parameters, we simply iterate between them until we achieve convergence.

We have just described a scheme to find a set of latent factors, Y, that explain as much of the dependence in X as possible. These latent factors, Y, are guaranteed to be less correlated than the original variables, X, as measured by TC **[3]** . Therefore, we can view the latent factors as an approximate expression of independent components. In results below, we consider how combinations of these quasi-independent factors can improve our ability to predict long-term health outcomes.

The quality of the solution that is found through this procedure is quantified through the expression $TC(X;Y)$. As we add more latent factors this objective increases before achieving a peak and then decreasing. At the first layer, we used 200 latent factors even though more would have resulted in higher TC because of computational limitations. At the higher levels, we tried different numbers of latent factors to optimize $TC(X;Y)$ leading us to use 30 factors at layer 2 and 8 at layer 3. With this number of factors, each CorEx run took about two days on a single Amazon "r3.xlarge" node with 30 GB of RAM.

Data Sources

Gene level RNA-seq values (normalized to reads per kilobase per million or RPKM) for 420 high grade serous ovarian tumor samples and 780 breast cancer samples were downloaded from the TCGA data portal (The TCGA Research Network: http://cancergenome.nih.gov), along with available clinical metadata for the samples. Somatic mutation data for 316 samples was obtained separately from the cBioPortal for Cancer Genomics **[37,38]** . A training set of genes was selected by combining the RPKM data for the top 3000 genes by variance in expression with the RPKM values for the 2371 genes that were mutated in more than one sample. The gene RPKM values for the selected genes were then converted to z-scores to form the 420x5371 training matrix. 376 samples were associated with survival times in the clinical Biotab file and used for the Cox proportional hazard analyses. GTEx normal tissue data **[39]** for 39 samples was downloaded from the GTEx portal (www.gtexportal.org) and normalized similarly to the tumor data.

Database annotations

KEGG, GO, and PPI enrichments for the gene groups were obtained using the stringdb package in R **[40,41]** . A maximum of 400 genes were used from each list. For groups from CorEx runs that utilized the Bayes shrinkage prior, only genes with a weighted mutual information greater than a threshold of 0.002 were retained for annotation enrichment purposes. FDR values with respect to the 200 groups (corresponding to the p-

value cutoffs used in the Circos **[42]** plot in **Figure 2**) were calculated by random selection from the set of 5371 genes used for training. The genes were alloted to groups with a size distribution identical to that of the CorEx discovered groups. In order to obtain summary GO terms for upper layer nodes in the CorEx factor hierarchy, the lists of GO terms from stringdb were submitted to the Revigo web server **[43]** , which clusters GO terms to select a subset of distinct representative terms.

Survival Analysis

For survival analysis, the gene groups were filtered by requiring less than 100 correlated genes each group (above threshold significance when Bayes prior used) and gene ontology terms with adjusted p-values less than the 1% FDR threshold. Additionally, fourteen groups were eliminated that were presumed to represent chromosomal location correlations rather than network interactions. This filtering resulted in 66 groups that were then analyzed for single factor survival associations under a coxph model in R. The top individual factors with differences in survival (determined by p-values between the two extremal strata as calculated by the survdiff function less than 0.05) were retained for combination analysis. The false discovery rate for this procedure was estimated by repeatedly selecting 66 factor groups (200 times), randomly shuffling their factor labels with respect to patient time of death and calculating p-values for both single and combination factors in the same fashion as before. False discovery rates were then calculated as the number of randomized groups or combinations exhibiting p-values less than a given threshold divided by the observed number of groups or combinations with p-values less than that threshold.

Bibliography

1.

Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615. Available: http://www.nature.com/doifinder/10.1038/nature10166.

2.

Ver Steeg G, Galstyan A (2014) Discovering structure in high-dimensional data through correlation explanation. Advances in Neural Information Processing Systems NIPS.

3.

Ver Steeg G, Galstyan A (2015) Maximally informative hierarchical representations of high-dimensional data. Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics AISTATS.

4.

Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, et al. (n.d.) SEER Cancer Statistics Review. Available: http://seer.cancer.gov/csr/1975_2012/.

5.

Foley OW, Rauh-Hain JA, del Carmen MG (2013) Recurrent epithelial ovarian cancer: an update on treatment. Oncology (Williston Park, NY) 27: 288. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23781692&retmode=ref&cmd=prlinks.

6.

Matsuzaki J, Gnjatic S, Mhawech-Fauceglia P, Beck A, Miller A, et al. (2010) Tumor-infiltrating NY-ESO-1–specific CD8+ T cells are negatively regulated by LAG-3 and PD-1 in human ovarian cancer. Proceedings of the National Academy of Sciences 107: 7875–7880. Available: http://www.pnas.org/content/107/17/7875.full.

7.

Homicsko K, Coukos G (2015) Targeting Programmed Cell Death 1 in Ovarian Cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology 33: 3987–3989. Available: http://jco.ascopubs.org/content/33/34/3987.full.

8.

Preston CC, Goode EL, Hartmann LC, Kalli KR, Knutson KL (2011) Immunity and immune suppression in human ovarian cancer. Immunotherapy 3: 539–556. Available: http://www.futuremedicine.com/doi/abs/10.2217/imt.11.20.

9.

Cubillos-Ruiz JR, Silberman PC, Rutkowski MR, Chopra S, Perales-Puchalt A, et al. (2015) ER Stress Sensor XBP1 Controls Anti-tumor Immunity by Disrupting Dendritic Cell Homeostasis. Cell 161: 1527–1538. Available: http://www.cell.com/article/S0092867415005759/fulltext.

10.

madsen sarah, Ver Steeg G, Mezher A, Jahanshad N, Nir TM, et al. (n.d.) Information-theoretic characterization of blood panel predictors for brain atrophy and cognitive decline in the elderly. IEEE th International Symposium on Biomedical Imaging ISBI: 980–984.

11.

Watanabe S (1960) Information Theoretical Analysis of Multivariate Correlation. IBM Journal of research and development.

12.

Van Dam S, Craig T, de Magalhães JP (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic acids … 43: NaN – NaN. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1042.

13.

Yang Y, Han L, Yuan Y, Li J, Hei N, et al. (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications 5: 3231. Available: http://www.nature.com/doifinder/10.1038/ncomms4231.

14.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science (New York, NY) 334: 1518–1524. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1205438.

15.

R Core Team (2015) R: A language and environment for statistical computing. Available: www.R-project.org.

16.

Cole C, Lau S, Backen A, Clamp A, Rushton G, et al. (2010) Inhibition of FGFR2 and FGFR1 increases cisplatin sensitivity in ovarian cancer. Cancer biology & therapy 10: 495–504. Available: http://www.tandfonline.com/doi/abs/10.4161/cbt.10.5.12585.

17.

Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, et al. (2003) Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. New England Journal of Medicine 348: 203–213. Available: http://www.nejm.org/doi/abs/10.1056/NEJMoa020177.

18.

Bodnar L, Stanczak A, Cierniak S, Smoter M, Cichowicz M, et al. (2014) Wnt/β-catenin pathway as a potential prognostic and predictive marker in patients with advanced ovarian cancer. Journal of ovarian research 7: 16. Available: http://ovarianresearch.biomedcentral.com/articles/10.1186/1757-2215-7-16.

19.

Janzen DM, Tiourin E, Salehi JA, Paik DY, Lu J, et al. (2015) An apoptosis-enhancing drug overcomes platinum resistance in a tumour-initiating subpopulation of ovarian cancer. Nature communications 6: 7956. Available: http://www.nature.com/doifinder/10.1038/ncomms8956.

20.

Glass K, Quackenbush J, Spentzos D, Haibe-Kains B, Yuan G-C (2015) A network model for angiogenesis in ovarian cancer. BMC bioinformatics 16: 115. Available: http://www.biomedcentral.com/1471-2105/16/115.

21.

D'Andrea AD (2015) Targeting DNA Repair in Cancer Therapy. ONCOLOGIST. Available: http://sto-online.org/sites/default/files/2015/STO_Article_CC_DAndrea.pdf.

22.

Song SJ, Poliseno L, Song MS, Ala U, Webster K, et al. (2013) MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling. Cell 154: 311–324. Available:

http://linkinghub.elsevier.com/retrieve/pii/S0092867413007654.

23.

Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. (2008) Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. Clinical Cancer Research 14: 5198–5208. Available: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-08-0196.

24.

Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, et al. (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. Nucleic acids … 44: NaN – NaN. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1165.

25.

Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, et al. (2012) Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics (Oxford, England) 28: 1592–1597. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts245.

26.

Cai Y, Fendler B, Atwal GS (2012) Utilizing RNA-Seq Data for Cancer Network Inference. arXiv.org q-bio.GN. Available: http://arxiv.org/abs/1211.4543v2.

27.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics 10: 57–63. Available: http://www.nature.com/doifinder/10.1038/nrg2484.

28.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5: 621–628. Available: http://www.nature.com/doifinder/10.1038/nmeth.1226.

29.

Peng L, Bian XW, Li DK, Xu C, Wang GM, et al. (2015) Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA

Cancer Types. Scientific reports 5: 13413. Available: http://www.nature.com/articles/srep13413.

30.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences of the United States of America 97: 12182–12186. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.220392197.

31.

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC bioinformatics 13: 328. Available: http://www.biomedcentral.com/1471-2105/13/328.

32.

Shi Z, Derow CK, Zhang B (2010) Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC systems biology 4: 74. Available: http://www.biomedcentral.com/1752-0509/4/74.

33.

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4: Article17. Available: http://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1128/sagmb.2005.4.1.1128.xml.

34.

Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, et al. (2012) Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics (Oxford, England) 28: 1592–1597. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts245.

35.

Cai Y, Fendler B, Atwal GS (2012) Utilizing RNA-Seq Data for Cancer Network Inference.

arXiv.org q-bio.GN. Available: http://arxiv.org/abs/1211.4543v2.

36.

Cheng W-Y, Ou Yang T-H, Anastassiou D (2013) Biomolecular events in cancer revealed by attractor metagenes. PLoS computational biology 9: e1002920. Available: http://dx.plos.org/10.1371/journal.pcbi.1002920.

37.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 6: pl1. Available: http://stke.sciencemag.org/cgi/doi/10.1126/scisignal.2004088.

38.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery 2: 401–404. Available: http://cancerdiscovery.aacrjournals.org/cgi/doi/10.1158/2159-8290.CD-12-0095.

39.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, et al. (2013) The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45: 580–585. Available: http://www.nature.com/doifinder/10.1038/ng.2653.

40.

Franceschini A, Franceschini MA, RUnit S (2015) Package ‘STRINGdb’. Available: http://bioconductor.riken.jp/packages/3.1/bioc/manuals/STRINGdb/man/STRINGdb.pdf.

41.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic acids … 37: NaN – NaN. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn760.

42.

Krzywinski M, Schein J, Birol I, Connors J (2009) Circos: an information aesthetic for comparative genomics. Genome …. Available:

https://genome.cshlp.org/content/19/9/1639.full.

43.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one 6: e21800. Available: http://dx.plos.org/10.1371/journal.pone.0021800.

Figures

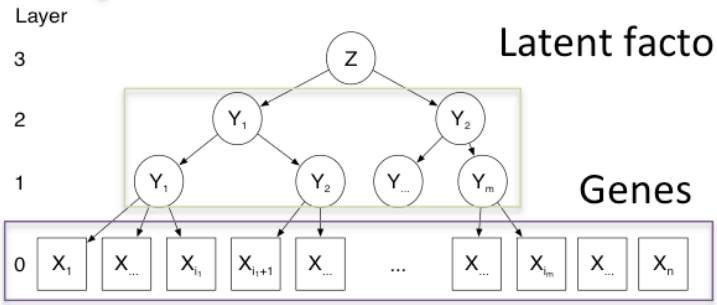
Genes

Expression table (normalized)	ABC1	BARD1	...
TCGA-0124-01A	10.2...		
TCGA...			

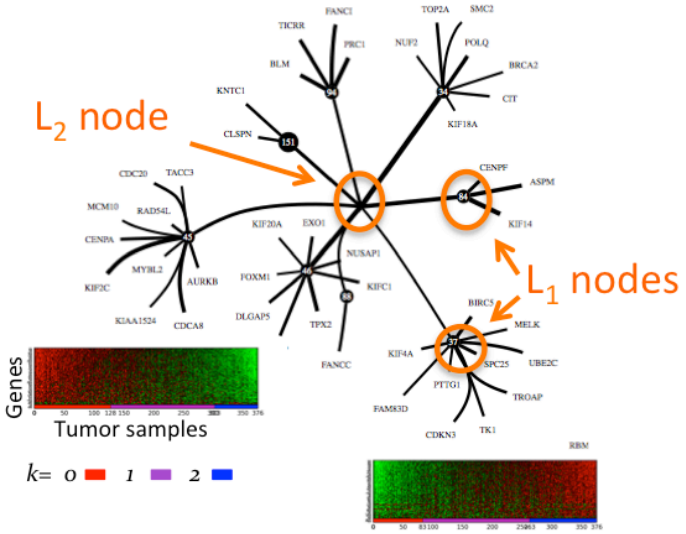
Tumor Samples

CorEx maximizes:

$$TC(X_G;Y) = TC(X_G) - TC(X_G|Y)$$



Gene Groups
with correlated expression



Factor Label Probabilities
(for each tumor sample *j*)

$$p^j_{i,k} = p(Y_i = k \mid X^j_1, \ldots X^j_m) \quad k \in \{0,1,2\}$$

$$f_{ji} = \log p^j_{i,0} - \log p^j_{i,2}$$

	Y0	Y1	...	Ym
TCGA-0124-01A	f_{00}=-52.1	24.0		-123.
TCGA-0123-01A	170.	-7.15		0.115
TCGA-1980-01A	47.3	10.2		37.9
...				

Figure 1:

Correlation Explanation algorithm applied to tumor RNA-seq. *For training, CorEx is provided only a matrix of normalized gene expression values for the available tumor samples. The number of possible labels for each latent factor is specified, here set to three. In this application, we also set the number of layer one latent factors to 200. CorEx finds probabilistic assignments of genes to latent factors by maximizing the total correlation of the genes in groups, simultaneously minimizing dependence between latent factor groups. The factor labels from lower layers are used as input to upper layers in order to generate a hierarchical model. The output from CorEx is thus a hierarchical model, specified as a set of probabilities characterizing the association of genes or factors with latent factors at the next highest layer as well as probabilities that a given tumor sample's expression pattern can be explained by each factor in a particular label state. The three probabilities for a given tumor sample and factor are can be usefully summarized by a single value that is the natural logarithm of the probability difference for the factor labels corresponding to extremal expression. The genes with high mutual information relative to latent factors show clear patterns of correlation when viewed on expression heat maps with tumor samples ordered by the summary latent factor score.*

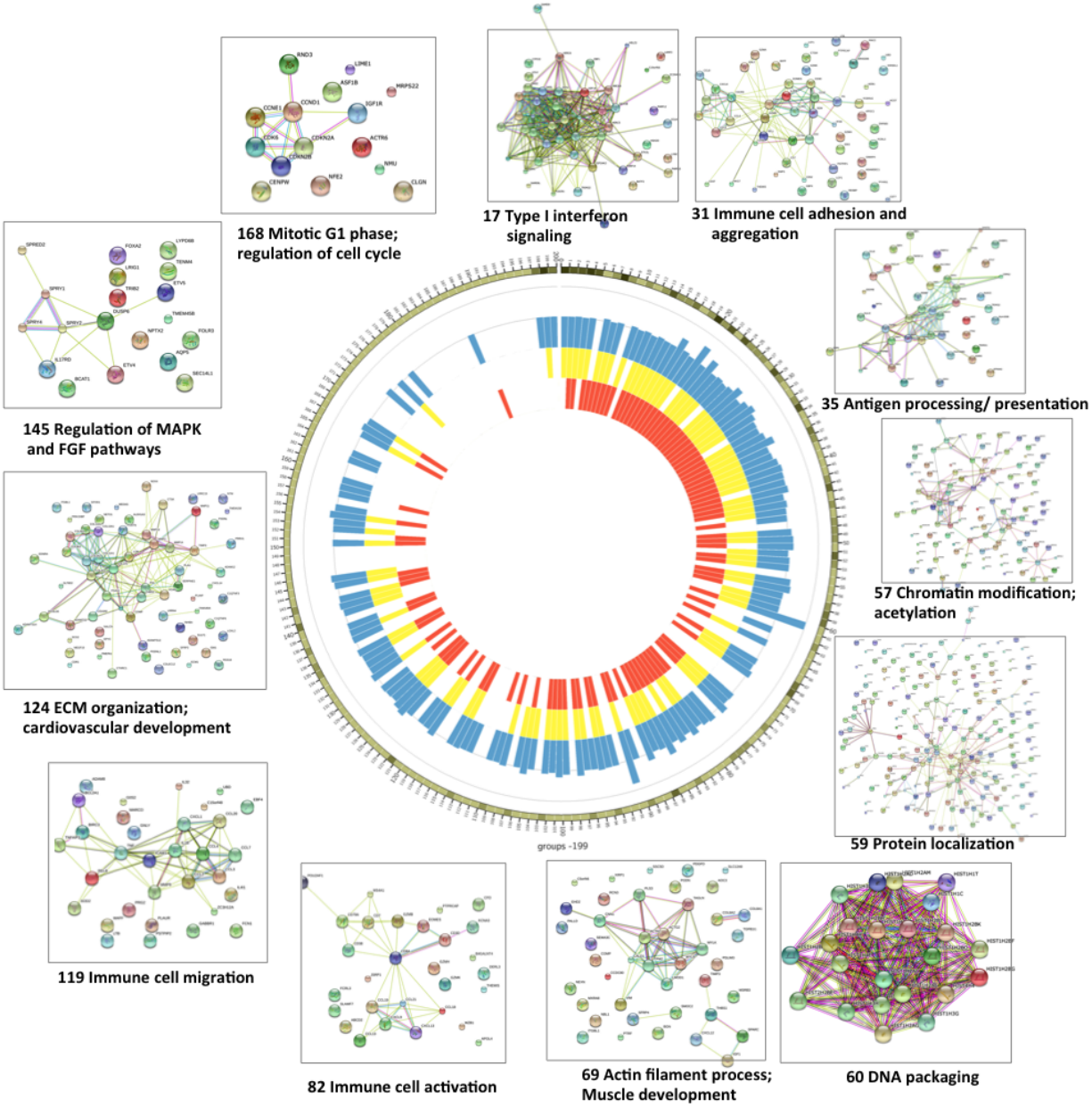

168 Mitotic G1 phase; regulation of cell cycle
17 Type I interferon signaling
31 Immune cell adhesion and aggregation
35 Antigen processing/ presentation
145 Regulation of MAPK and FGF pathways
57 Chromatin modification; acetylation
124 ECM organization; cardiovascular development
59 Protein localization
groups -199
119 Immune cell migration
82 Immune cell activation
69 Actin filament process; Muscle development
60 DNA packaging

Figure 2:

Gene Ontology, KEGG, and protein interaction enrichment for CorEx gene groups. *A circos plot indicates which latent factor groups of the 200 inferred by CorEx are enriched for genes on KEGG pathways (yellow), Gene Ontology terms (orange) and Protein-Protein interactions (blue). A tick mark of the corresponding color is drawn for groups exhibiting enrichment greater than a threshold. Further, the blue PPI tick marks have heights proportional to the relative enrichment (observed/expected). The group results appear in order of decreasing Total Correlation, with the number of genes in a group indicated by the green shading on the outer ring of the circos plot. (darker means a greater number of genes with a range from 4 to 1135) Cutoffs were initially chosen according to the presence of coherent biological signals, however randomization tests indicate corresponding FDR values of 1%, 5%, and 10% for GO, KEGG, and PPI, respectively. The vast majority of CorEx groups show annotation enrichment, typically for all three types. A sample of string network connections are shown on the outside, showing both a diversity of network function and relatively few apparently spuriously associated genes within the groups.*

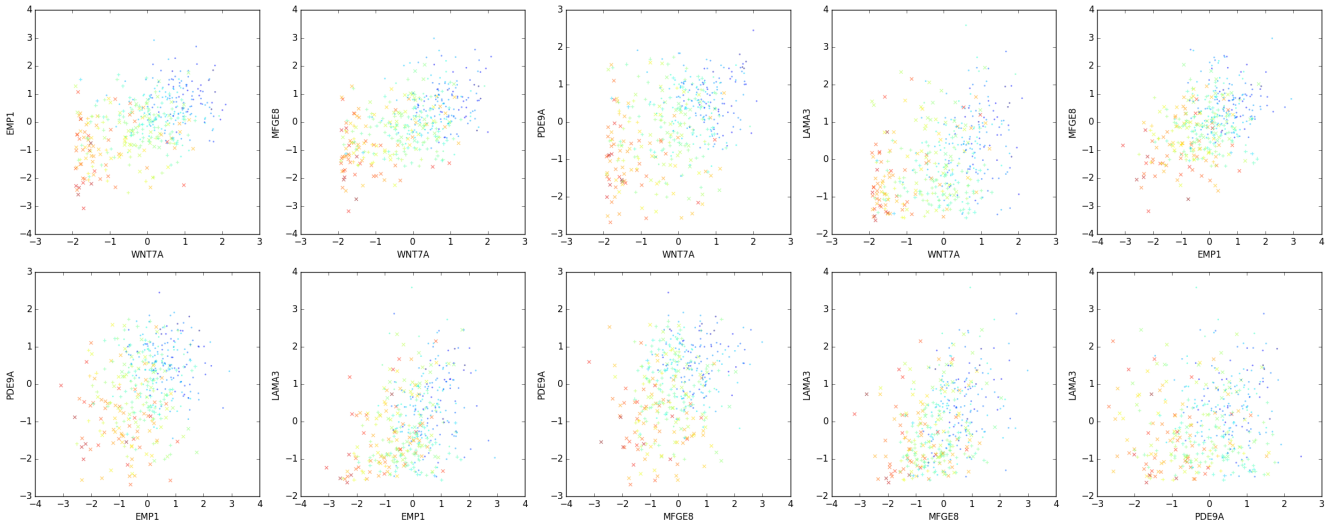
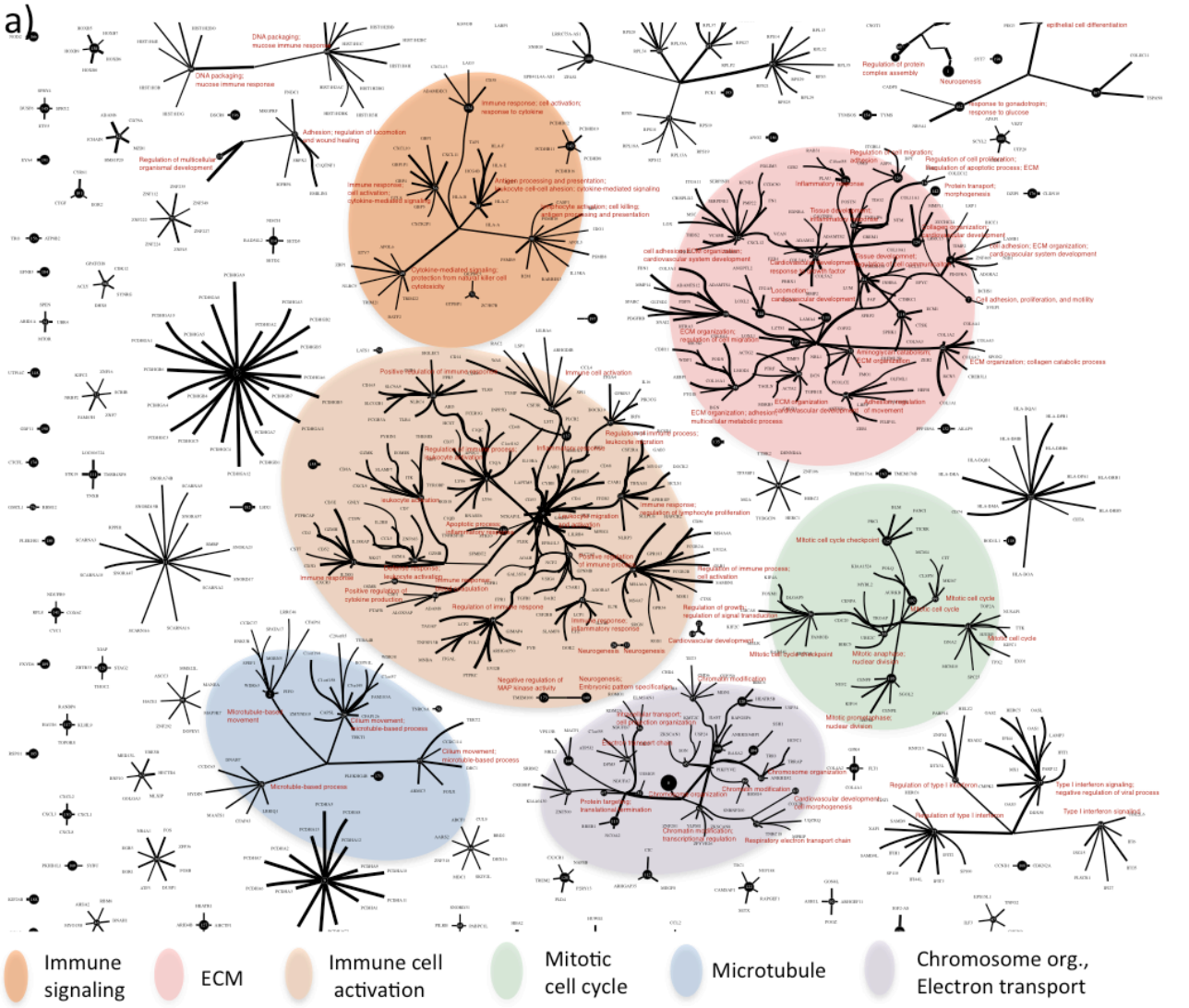


Figure 3:

The use of multivariate mutual information increases the clustering power of CorEx. *Pairwise correlations of normalized expression of individual samples for the top genes in group 106 appear very weak. The factor label values (indicated by color) show strong coherence across the different plots indicating how samples cluster more strongly in terms of total correlation.*


a)
Immune signaling
ECM
Immune cell activation
Mitotic cell cycle
Microtubule
Chromosome org., Electron transport

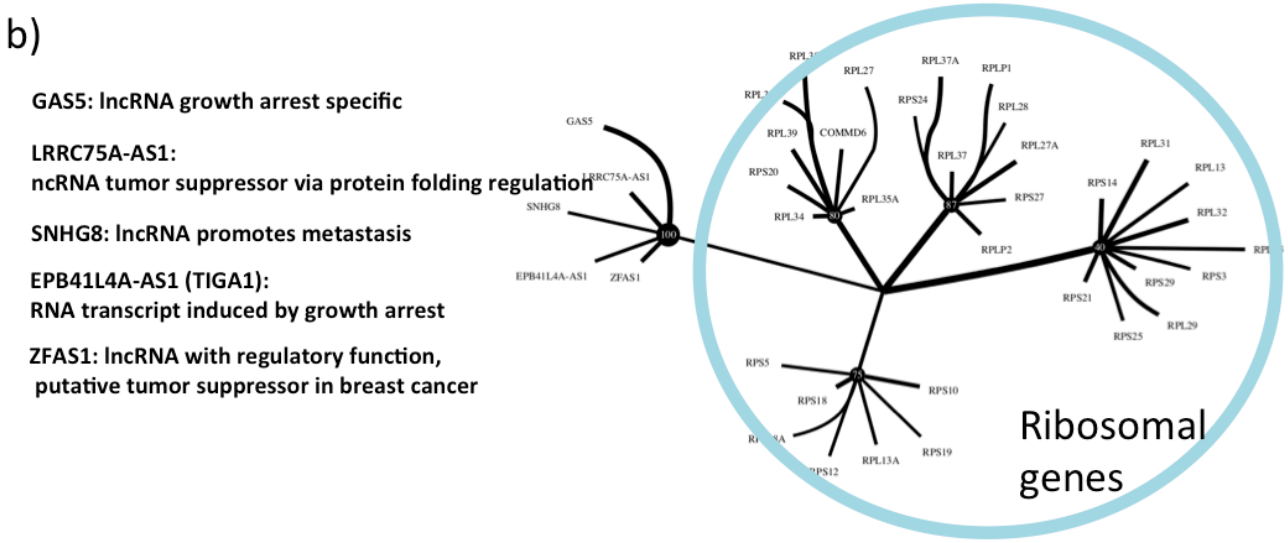

b)
GAS5: lncRNA growth arrest specific
LRRC75A-AS1:
ncRNA tumor suppressor via protein folding regulation
SNHG8: lncRNA promotes metastasis
EPB41L4A-AS1 (TIGA1):
RNA transcript induced by growth arrest
ZFAS1: lncRNA with regulatory function,
putative tumor suppressor in breast cancer
GAS5
LRRC75A-AS1
SNHG8
EPB41L4A-AS1
ZFAS1
RPL27
RPL37A
RPLP1
RPS24
RPL28
RPL39
COMMD6
RPL27A
RPL37
RPS20
RPL35A
RPS27
RPL34
RPL31
RPL13
RPS14
RPL32
RPLP2
RPS3
RPS29
RPS21
RPL29
RPS25
RPS5
RPS10
RPS18
RPS19
RPL13A
Ribosomal genes

Figure 4:

Gene Function Relationships in the Latent Factor Hierarchy. *a) Hierarchical factor relationships are shown exhibiting a tendency to cluster layer one groups with similar general Gene Ontology enrichments. Lines are shown connecting the top X strongest relationships and the top 1000 genes. Thicker lines indicate higher CorEx relationship probabilities. Representative GO terms for hierarchically clustered groups only are shown in red. Though this figure highlights clustering at layer one and above, many nonclustered layer one factors have very significant annotations as well, detailed in Fig. 2 and the supplementary information. b) In addition to general relationships, specific causal relationships are also detected at the upper layers. Here an example is shown in which noncoding RNAs that are implicated in tumor growth and metastasis via regulation of ribosome activity are isolated in a layer zero group and joined at the higher level to ribosomal structure genes. Noncoding RNA functional annotations are taken from the NCBI Gene database (http://www.ncbi.nlm.nih.gov/gene).*

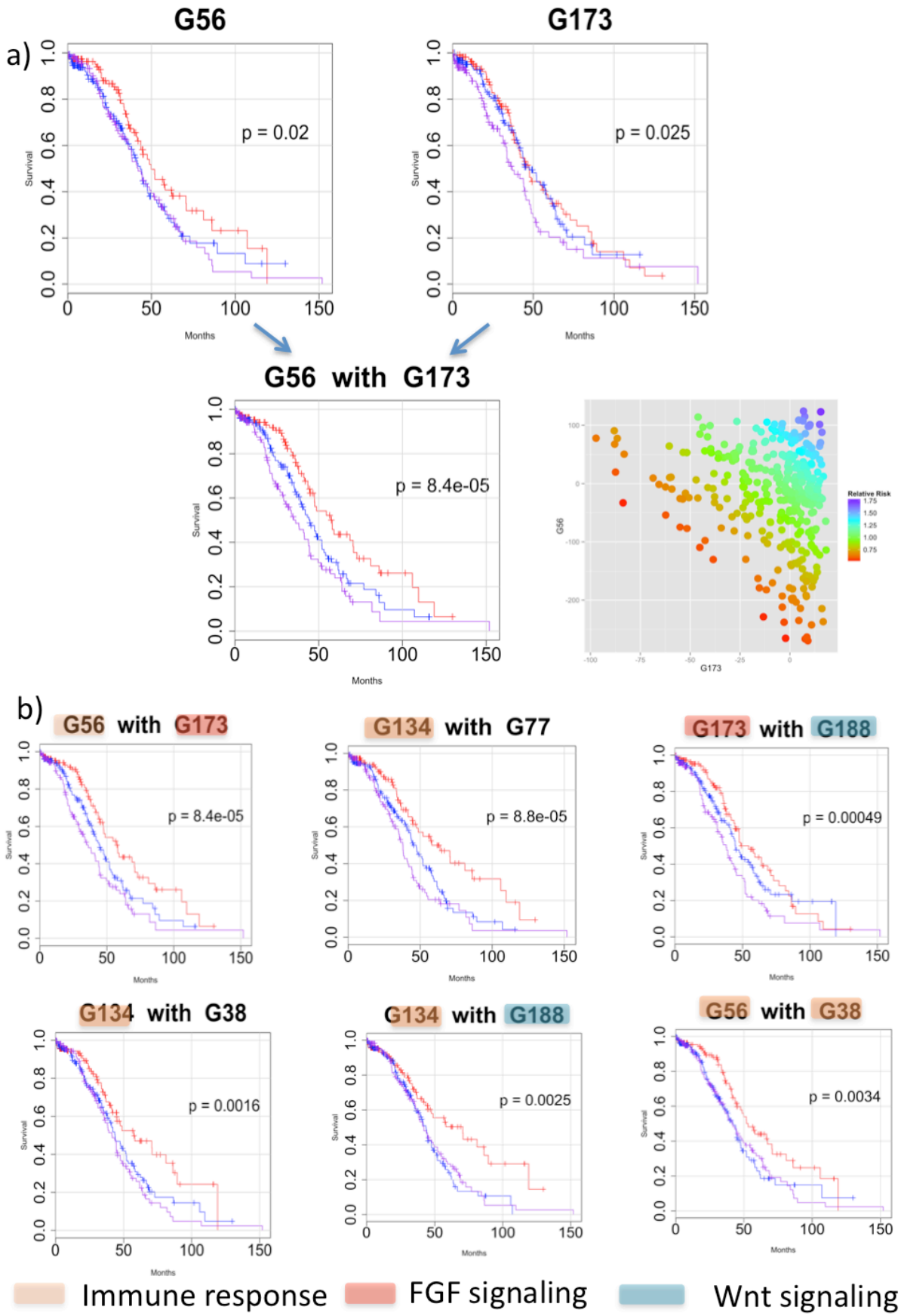

a)
G56
G173
p = 0.02
p = 0.025
Survival
Months
G56 with G173
p = 8.4e-05
Relative Risk
G56
G173
b)
G56 with G173
p = 8.4e-05
G134 with G77
p = 8.8e-05
G173 with G188
p = 0.00049
G134 with G38
p = 0.0016
G134 with G188
p = 0.0025
G56 with G38
p = 0.0034
Immune response
FGF signaling
Wnt signaling

Figure 5:

Group Survival Associations and Combinations *a) A Cox proportional hazard model is used to estimate relative risk for patients according to individual factor label values. Using this model, a relative risk of death can be calculated for each patient based upon her group label value. When patients are stratified into thirds according to relative risk, some factors show an association with survival indicated by the difference in Kaplan-Meier survival curves for the extremal groups (p-values shown in the graphs). The figure shows survival curves for two such factors, G56 and G173. The bottom 30%, middle 40%, and top 30% in terms of predicted survival under the coxph model are plotted in magenta, blue, and red, respectively. When these two factors are combined into a single risk model, the survival differential increases dramatically. b) Single factor combinations with various Gene Ontology annotations can be combined and show synergy relative to survival association. The false discovery rate under a randomization test for six such pairs is below about 1/3 compared to only one single factor achieving that level. Note that it is not suggested that factors be selected based solely upon survival associations but rather factors that are interesting for other reasons, e.g. they contain druggable targets, can be investigated for an impact on survival both alone and in combinations.*

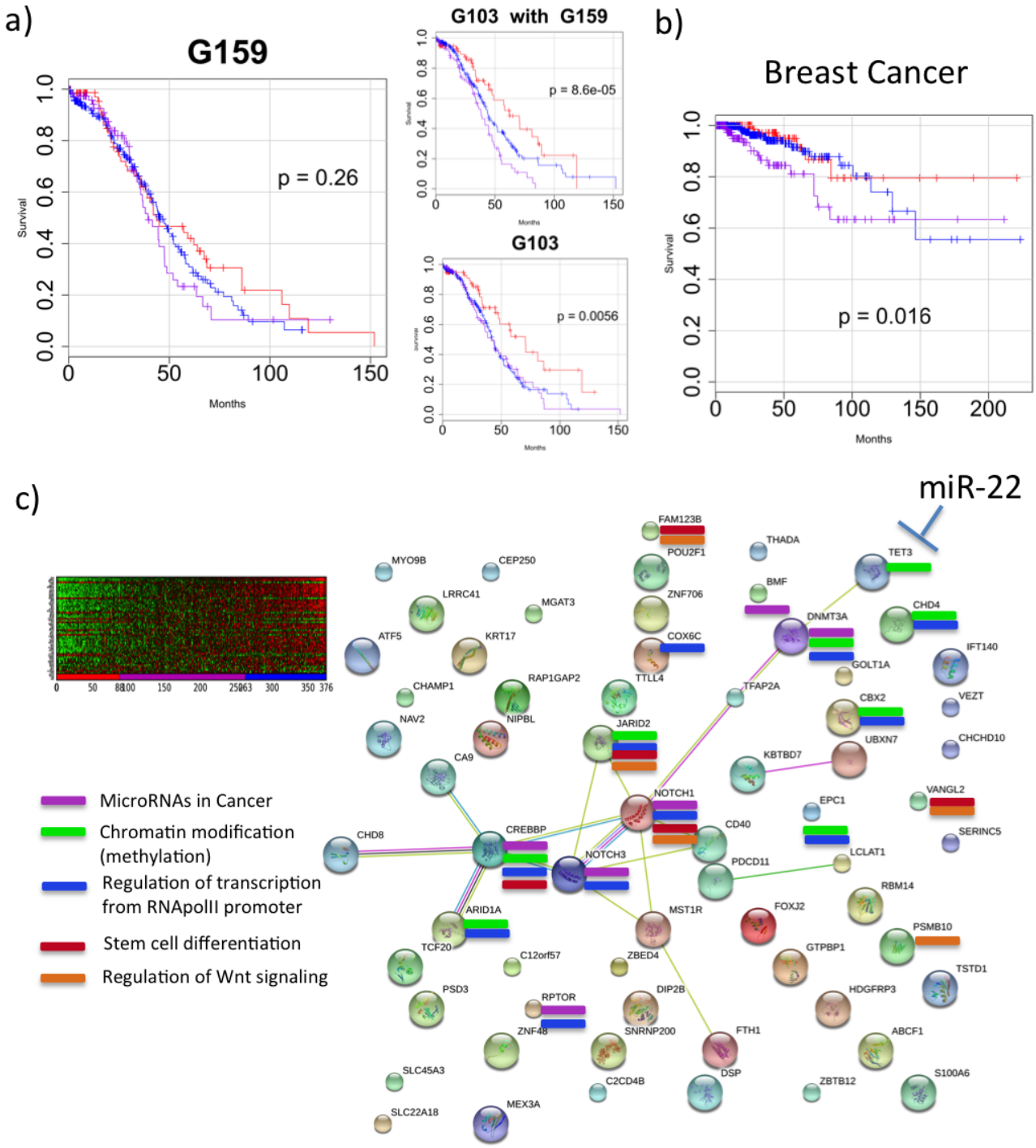
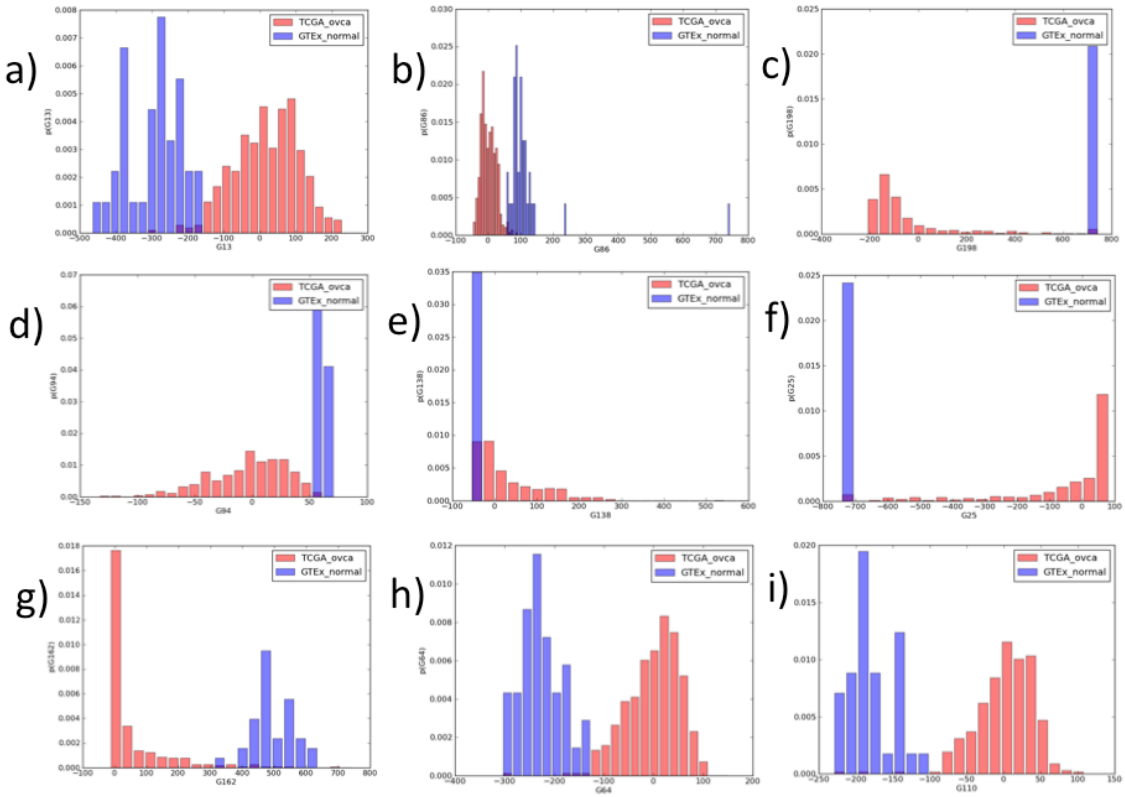

a)
G159
p = 0.26
Survival
Months
G103 with G159
p = 8.6e-05
G103
p = 0.0056
b)
Breast Cancer
p = 0.016
Survival
Months
c)
miR-22
MicroRNAs in Cancer
Chromatin modification (methylation)
Regulation of transcription from RNApolII promoter
Stem cell differentiation
Regulation of Wnt signaling
FAM123B
POU2F1
THADA
TET3
MYO9B
CEP250
BMF
LRRC41
MGAT3
ZNF706
DNMT3A
CHD4
ATF5
KRT17
COX6C
IFT140
GOLT1A
TTLL4
RAP1GAP2
TFAP2A
CHAMP1
CBX2
VEZT
NAV2
NIPBL
JARID2
UBXN7
CHCHD10
KBTBD7
CA9
NOTCH1
EPC1
VANGL2
CD40
SERINC5
CHD8
CREBBP
NOTCH3
PDCD11
LCLAT1
RBM14
ARID1A
MST1R
FOXJ2
PSMB10
TCF20
C12orf57
ZBED4
GTPBP1
TSTD1
PSD3
RPTOR
DIP2B
HDGFRP3
ZNF48
SNRNP200
FTH1
ABCF1
SLC45A3
C2CD4B
DSP
ZBTB12
S100A6
SLC22A18
MEX3A

Figure 6:

Discovery of new targets: regulatory microRNAs and EMT *a) One factor was observed to show a bias toward a difference in long-term survival. Though its survival differential was not significant alone, when put into a model combined with the factor for immune chemokine signaling, the resulting survival association was substantially greater and exceeded that of every other tested pair for that run. b) Breast cancer samples were mapped to CorEx factors learned from ovarian tumors and demonstrate a survival difference for this factor. c) The group contains a small network related to microRNAs in cancer, chromatin modification, and stem cell differentiation. Work in breast cancer associates the network with a cellular EMT phenotype and increased aggressiveness and metastasis of breast tumors. EMT features have been shown to be influenced through regulation by TET3 of the demethylase DNMT3A and represent a druggable target via the regulatory microRNA miR-22*

.



	Genes (number of distinct drugs in Dgldb)	Annotation
a	PDGFRA (27), CCND1 (4), LRP1 (2),CYP26B1(1)	Development, regulation of cell differentiation
b	LGALS1(2),SMO(9),TRPC1(3),SCN5A(39)	Neurogenesis
c	KIT(41),EPOR(1), RB1(1)	Tissue development
d	COX7B(1),NDUFC2(1), UQCRQ(5)	Respiratory electron transport chain
e	---	Embryonic development
f	CYP11A1(2),FSHR(6),ICAM1(3),NR5A1(1)	Sex differentiation
g	GJB2(4),KCNK3(8)	Response to gonadotropin, cell-cell signaling
h	F2R(5),MMP14(1),S1PR1(11),SGK1(3),TEK(9), THBS1(1)	Regulation of cell migration
i	C5AR1(4),CD163(1),IL4R(1),PIK3CG(34),SIGLEC1(2),TGFB1(2),TLR4(7),TNFRSF1B(1),JAK3(17),TLR4(7)	Inflammatory response, regulation of programmed cell death

Figure 7:

Comparison to normal ovarian tissue suggests druggable targets. *When normal ovarian samples are assigned label values for the factors found from ovarian tumors, many exhibit striking differences. The histograms show factor scores in cases with large differences that correspond to novel drug targets in ovarian cancer (examples listed in Table 1).*

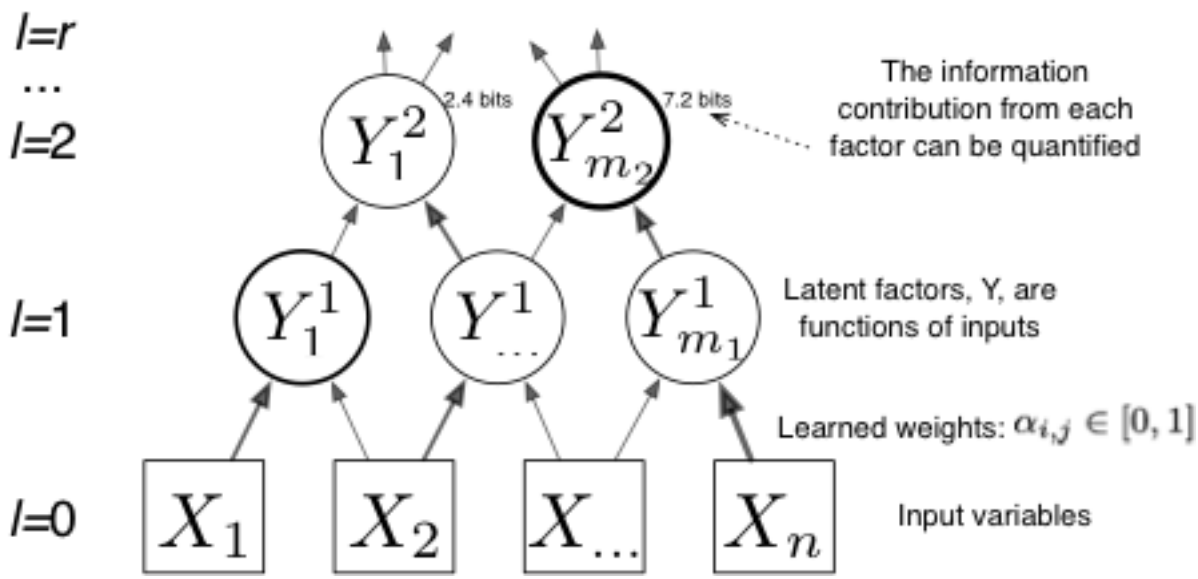


Figure 8:

A simple hierarchical model constructed by CorEx. *Variables in each layer are associated with latent factors in the next higher layer with a weight that is determined by optimizing over its relative contribution to the explanatory latent factors. Visually we set the thickness of links based on mutual information between variables and the thickness of nodes is proportional to the information contributed by that factor toward the objective, TC(X; Y).*

Supplementary Figures and Text for Multivariate information maximization yields hierarchies of expression components in tumors that are both biologically meaningful and prognostic

Shirley Pepke[1] and Greg Ver Steeg[2]

1 University of Southern California, Information Sciences Institute (Marina del Rey, CA, USA); **2** Lyric LLC (South Pasadena, CA, USA)

1. Supplementary Figures

see in separate file

Figure 1:

All heat maps located at:

https://app.box.com/s/r2wxxfj89obz47q4qxrcdpf9lj4czu26/SuppFig1.heats.pdf

see in separate file

Figure 2:

All string networks located at:

https://app.box.com/s/r2wxxfj89obz47q4qxrcdpf9lj4czu26/SuppFig2.stringnetworks.pdf

Consider groups X_A^m and X_B^n (mth group from training A and nth group from training B.

If RBO_extn > $RBO_{threshold}$:

$$\{X_A^m, X_B^n\} \rightarrow \begin{cases} \{X_A^m\} & \text{if } TC_N(X_A^m) > TC(X_B^n) \\ \{X_B^m\} & \text{otherwise} \end{cases}$$

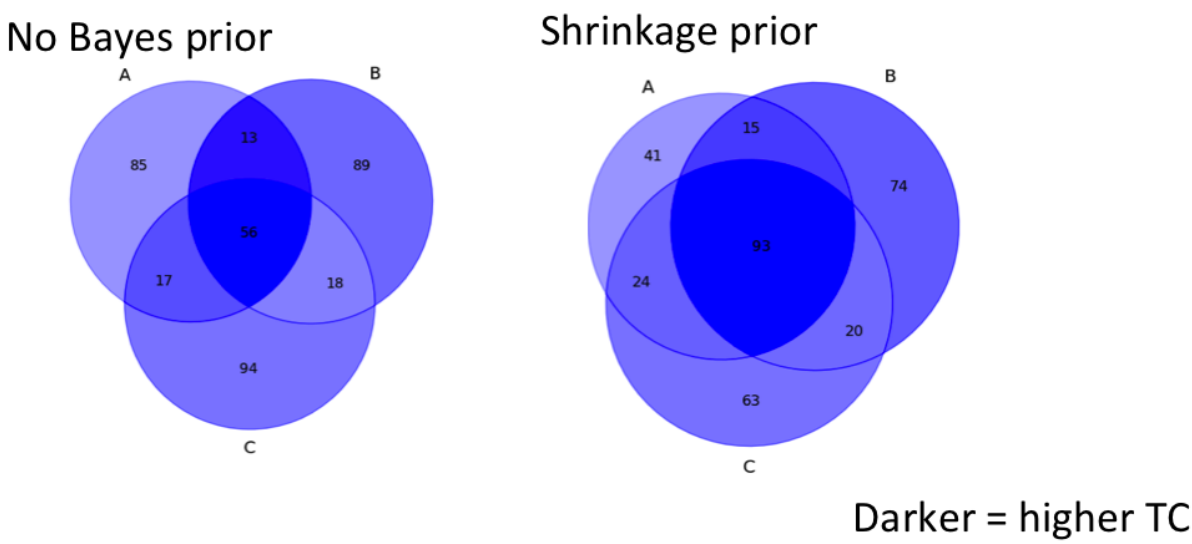


Figure 3:

The rank biased overlap can be used to identify equivalent groups from different runs (to within some tolerance) Run results can be aggregated by declaring groups equal if their RBO exceeds some threshold and then retaining only the group with highest TC from among a set of equivalent groups. Reproducibility between runs was measured using the rank-biased overlap (after excluding very large groups). The Venn diagrams show how utilizing a Bayes shrinkage prior as a smoothing technique increases the fraction of groups that are identified as equivalent according to the RBO measure. In both cases, groups with higher TC or more likely to appear relatively unchanged across runs.

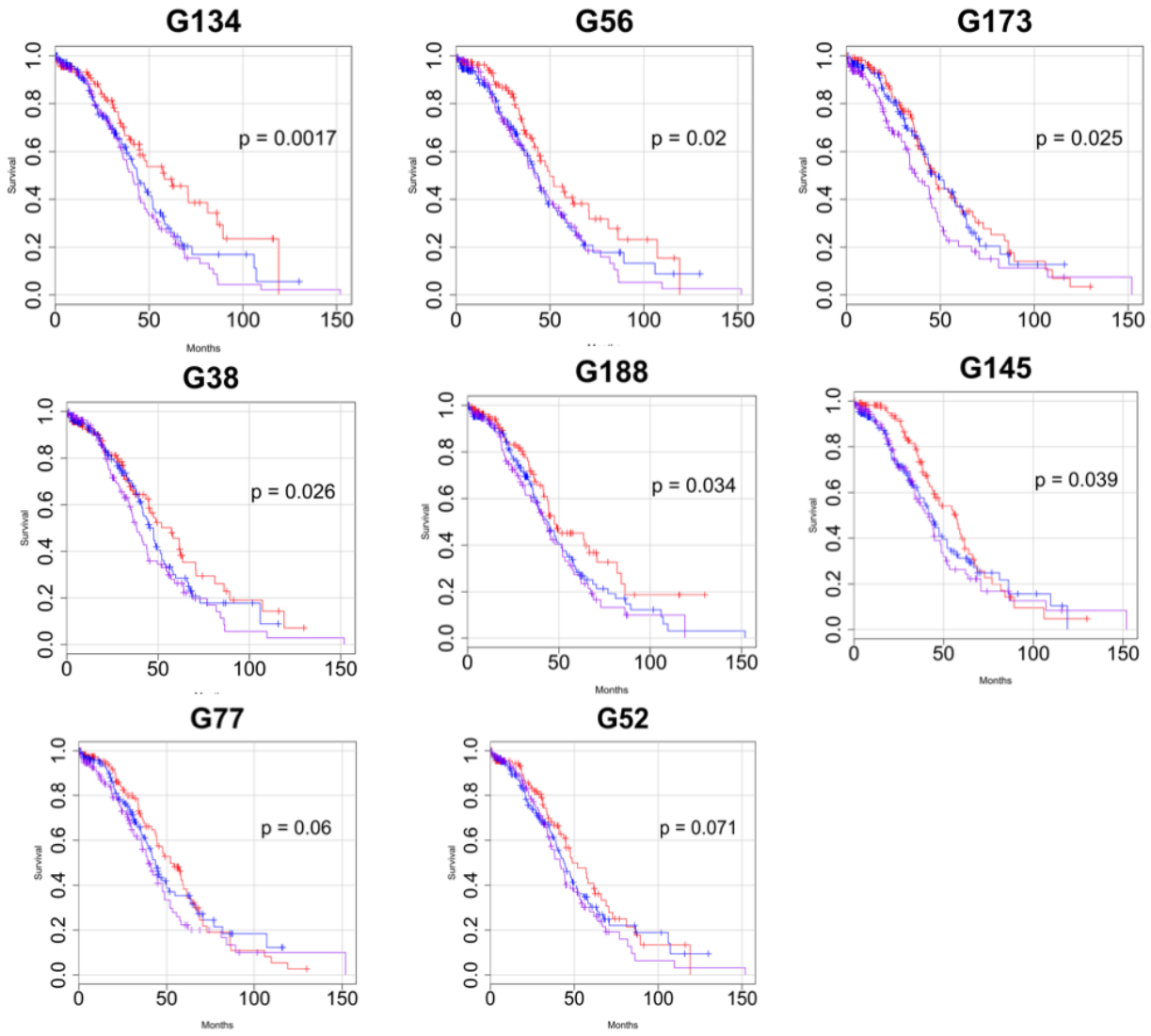


Figure 4: *Supplementary survival curves*

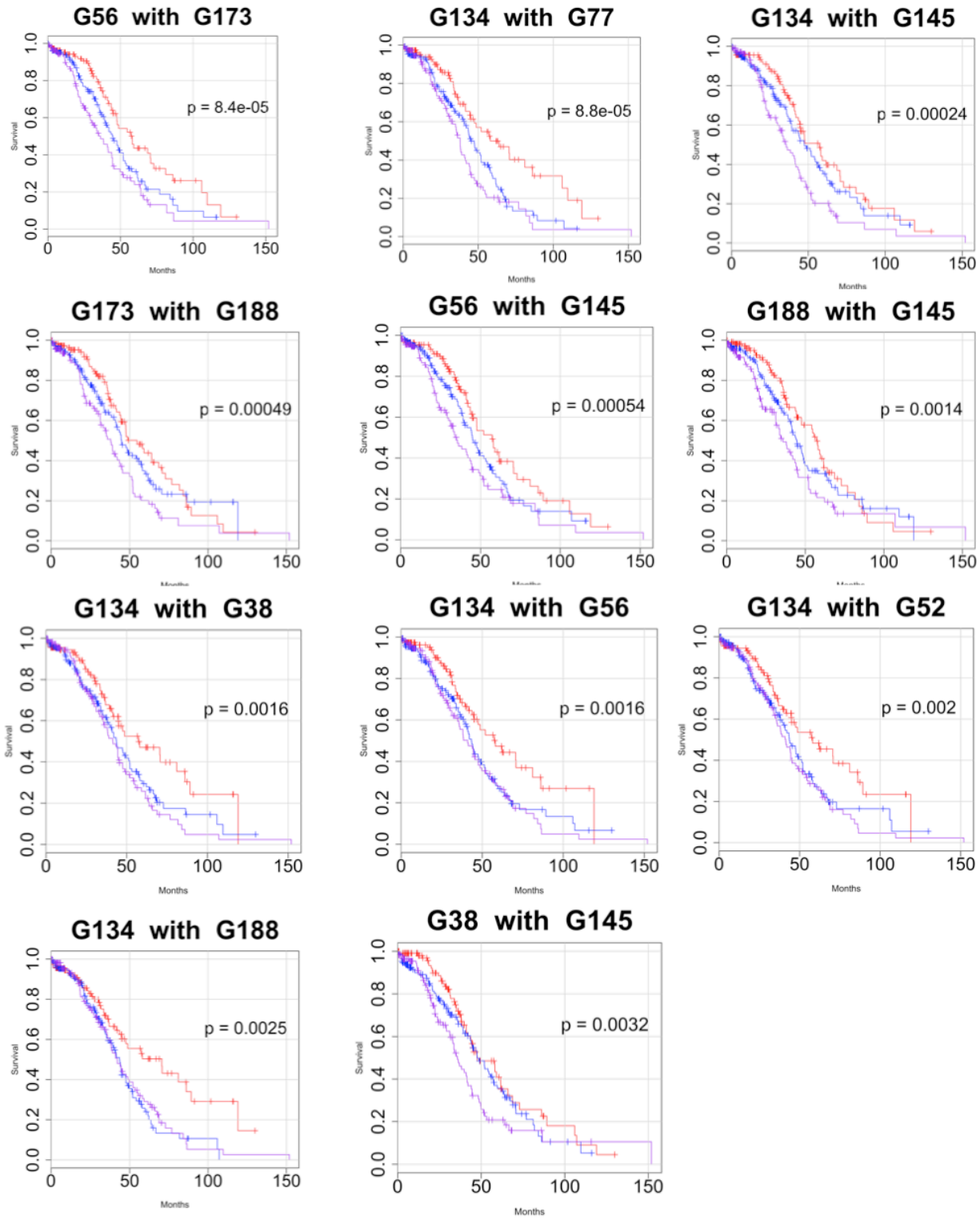


Figure 5: *Supplementary combination survival curves*

2. Supplementary Text

Given the variability between runs of CorEx in such a high dimensional space, some way to compare and reconcile the results of multiple runs is needed. We have found that comparison of the gene groupings using the rank-biased overlap (RBO) of **[1]** can be used to draw intuitively satisfying correspondences. The RBO is calculated as a single number that describes the similarity between two ranked lists. It is very flexible and has appealing asymptotic properties. It is possible to usefully aggregate results from various runs by declaring any two groups that exceed a threshold RBO to be equivalent. When this is the case, the group with greater TC is retained.

3. Supplementary Methods

3.1. Estimation of parameters

3.1.1. Bayesian shrinkage estimates for the marginals

The remaining details of the learning scheme concern the estimation of parameters in terms of our samples of data x, and the current estimate of probabilistic labels, $p(Y_j = y_j | X = x)$. First of all, as we mentioned, $p(X_i = x_i | Y_j = c)$ is estimated as a normal with mean $\mu_{i,j,c}$ and variance, $\sigma_{i,j,c}$. Let x^l denote the l-th sample of data. Then an empirical estimate for μ, would be the following.

$$\mu_{i,j,c}^{ML} = (\sum_l p(Y_j = c | X = x^l) x_i^l) / (\sum_l p(Y_j = c | X = x^l) \tag{1}$$

We could simply use this estimate and the corresponding one for variance. However, if there are a small number of samples with label $Y_j = c$, this could become quite noisy. Instead, we consider a Bayesian estimate based on James-Stein type shrinkage estimators **[2,3]** . We take as our Bayesian prior the hypothesis that the mean is $\mu_i{}^0 = 1/N \sum_i x_i{}^l$.

A Bayesian estimate for the mean of a normal distribution has a simple form.

$$\mu_{i,j,c} = \lambda_{i,j,c} \mu_{i,j,c}^{ML} + (1 - \lambda_{i,j,c}) \mu_i^0 \tag{2}$$

The main question is how to set the value of the ``shrinkage parameter", λ. The idea behind shrinkage estimators is to analytically estimate the value for λ to be the one with the minimum risk. We derive our estimate here since the setting differs slightly from previous attempts **[2,3]** .

For simplicity, we drop the subscripts and assume the true distribution has mean, μ, the prior is μ^0, and the empirical estimate is $\hat{\mu}$, and that the standard deviation is fixed and known σ. Let $p(x; \mu, \sigma)$ be a normal distribution parametrized by μ, σ. The risk, R is defined as the KL divergence between the true distribution and the estimated one.

$$R = \mathbb{E}[D(p(x; \mu, \sigma)||p(x; (1-\lambda)\mu_0 + \lambda\hat{\mu}, \sigma)] \tag{3}$$

We set λ to be the one that minimizes the risk by taking the derivative and solving as usual. It turns out that this leads to the following expression.

$$\lambda = \frac{(\hat{\mu} - \mu_0)^2}{(\hat{\mu} - \mu_0)^2 + z^2} \tag{4}$$

We have $z^2 = \sigma^2/N$ and we recognize this is just the standard error of the mean. Rather than estimate that, we use a shuffle test to estimate how far we expect $\hat{\mu}$ to be from μ_0 under the null hypothesis, and we call this $\hat{z}$. For some random permutation of the samples, π, we have the following.

$$\mu_{i,j,c}^{\text{shuffle}} = (\sum_l p(Y_j = c|X = x^l) x_i^{\pi(l)})/(\sum_l p(Y_j = c|X = x^l)) \tag{5}$$

Then we take an expectation over several shuffles, $\hat{z}_{i,j,c}^2 = \mathbb{E}_\pi[(\mu_{i,j,c}^{\text{shuffle}} - \mu_i^0)^2]$. For ``agressive" shrinkage, instead of shuffle x_i^l, we sample with replacement from the empirical distribution.

3.1.2. Setting the weights

The form of the weights, $a_{i,j}$ are directly determined by the optimization which requires that $a_{i,j} = I(X_i; Y_j|Y_1, \ldots, Y_{j-1})/I(X_i; Y_j)$, for some ordering of the Y_j. We use a more tractable estimate for this weight **[4]** .

For each sample, x^l, we first calculate the most likely label for each latent factor, $\bar{y}_j^l = \arg\max_y p(Y_j = y | X = x^l)$. We define the prediction of Y_j based on X_i for sample l as $P_{i,j,l} = \arg\max_y p(X_i = x_i^l | Y_j = y) / p(X_i = x_i^l)$. We define, $C_{i,j,l} = 1$ if $P_{i,j,l} = \bar{y}_j^l$ and 0 otherwise. $C_{i,j,l}$ shows whether X_i correctly predicts Y_j for sample l. Next, for each i, we sort j's in decreasing order of $C_{i,j} \equiv \sum_l C_{i,j,l}$. Then we set $a_{i,j}$ to be the fraction of samples for which $C_{i,j,l} = 1$ and $C_{i,k,l} = 0$, $\forall k < j$. In other words, it is the fraction of times that X_i correctly and uniquely predicted Y_j.

4. Bibliography

1.

Webber W, Moffat A, Zobel J (2010) A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS) 28: 20–38. Available: http://portal.acm.org/citation.cfm?doid=1852102.1852106.

2.

Hausser J, Strimmer K (2009) Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. The Journal of Machine Learning Research 10: 1469–1484. Available: http://dl.acm.org/citation.cfm?id=1577069.1755833.

3.

Schafer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology 4: Article32. Available: http://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1175/sagmb.2005.4.1.1175.xml.

4.

Ver Steeg G, Galstyan A (2015) Maximally informative hierarchical representations of high-dimensional data. Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics AISTATS.