Uncovering Robust Patterns of MicroRNA Co-Expression across Cancers Using Bayesian Relevance Networks

Parameswaran Ramachandran[1☯¤], Daniel Sánchez-Taltavull[1☯], Theodore J. Perkins[1,2*]

1 Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada K1H8L6
2 Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada K1H8M5

☯These authors contributed equally to this work.
¤Current Address: The Campbell Family Institute for Breast Cancer Research, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada M5G2M9
* tperkins@ohri.ca

Abstract

Co-expression networks have long been used as a tool for investigating the molecular circuitry governing biological systems. However, most algorithms for constructing co-expression networks were developed in the microarray era, before high-throughput sequencing—with its unique statistical properties—became the norm for expression measurement. Here we develop Bayesian Relevance Networks, an algorithm that uses Bayesian reasoning about expression levels to account for the differing levels of uncertainty in expression measurements between highly- and lowly-expressed entities, and between samples with different sequencing depths. It combines data from groups of samples (e.g., replicates) to estimate group expression levels and confidence ranges. It then computes uncertainty-moderated estimates of cross-group correlations between entities, and uses permutation testing to assess their statistical significance. Using large scale miRNA data from The Cancer Genome Atlas, we show that our Bayesian update of the classical Relevance Networks algorithm provides improved reproducibility in co-expression estimates and lower false discovery rates in the resulting co-expression networks. Software is available at www.perkinslab.ca/Software.html.

Introduction

Co-expression of genes, microRNAs, long non-coding RNAs and other transcribed entities is a key biological property with multiple implications [7, 26, 37]. On the one hand, co-expression can indicate co-regulation at the transcriptional level thereby revealing how gene expression is controlled [4, 30, 37] while, on the other hand, co-expression can be the result of coordinated epigenetic mechanisms [28, 31]. In yet other instances, co-expression of certain genes can serve as biomarkers in diseases such as cancer [11, 20] and mental disorders [31], or aid in defining distinct cell populations and subpopulations [25].

One of the earliest, and still widely used, tools for estimating and exploring networks of co-expression is the Relevance Networks algorithm of Butte *et al.* [10]. The algorithm has four main steps. First, entities (e.g., genes) with low estimated entropy are removed, as correlations between them may result from one or a few outlier samples. Second, Pearson correlations are computed between all pairs of remaining entities. Third, permutation testing is used to establish a null distribution for the correlations. Fourth and finally, a co-expression network is created by connecting any pair of entities whose correlation exceeds a statistical significance threshold set by the user (in conjuction with the estimated null distribution). The Relevance Networks algorithm has been used successfully in numerous studies to uncover significant co-expression relationships (e.g., [13, 17, 23, 27]).

Many elaborations and alternatives to the original Relevance Networks algorithm have been proposed over the years [1, 8, 9, 14, 34, 42]. These include improvements aimed at detecting non-linear relationships between the expression of different entities by using mutual information criteria [8, 9], or discriminating co-expression more likely to result from direct rather than indirect interactions [8, 14]. As replicate data became more common, algorithms were developed to accomodate co-expression analysis of data with replicates [1, 42]. Other work has focussed on robustly estimating correlations when the number of samples is much smaller than the number of entities [34]—although interestingly, we are finally emerging from that conundrum. For instance, the dataset we analyze in this paper describes 2,456 miRNAs measured over 10,999 samples.

While these algorithms included important new ideas and methods, they were all developed in the era of microarray-based expression measurements. The recent past has seen a fundamental shift in the technology used for expression measurement from microarray-based to sequencing-based platforms [2, 16]. Sequencing-based approaches produce measurement values with very different error properties, dynamic ranges, and signal-to-noise ratios than microarrays. In particular, the relative precisions of low-expression measurements are much worse compared to those of high-expression measurements. Furthermore, precision differs between samples, at the very least due to differences in sequencing depth, if not other factors [3]. Owing to all these reasons, the established body of algorithms for co-expression network construction may not be optimal for sequencing-based expression measurements, and hence there is a strong need to adapt these methods to the realities of this new type of data.

Here, we develop a Bayesian version of the classical algorithm of Butte *et al.* [10], which we call the Bayesian Relevance Networks algorithm. It builds on our recent work where we proposed a Bayesian correlation scheme to analyze sequence count data [33]. We employ Bayesian statistics both for estimating the expression levels and for quantifying the uncertainties in those estimates. From those beliefs, we construct estimates of mean expression levels and their uncertainties in groups of samples. This allows us to study cross-group correlations in studies with replicates or other natural sample groups (e.g., patients with the same disease). We describe how to perform permutation testing to estimate a null distribution for grouped Bayesian correlations. This enables the computation of p-values for the statistical significance of observed correlations, and allows us to estimate rates of true and false positive links in a Bayesian Relevance Network.

Throughout the paper, we evaluate our approach on a large-scale public microRNA (miRNA) expression dataset from The Cancer Genome Atlas project (TCGA) [41]. In a series of cross-validation studies, we find that Bayesian co-expression estimates are more reproducible than the Pearson co-expression estimates used by the original Relevance Networks algorithm. We find that Bayesian Relevance Networks are less prone to false positive links and have lower false discovery rates than classical Relevance Networks. Finally, we find that entropy filtering to remove “spurious” correlations improves both classical and Bayesian Relevance Networks. At the end of the Results section, we present a Bayesian Relevance Network based on the full datasets, where we demonstrate several interesting cancer type-specific clusters of co-expressed miRNAs.

Materials and Methods

Problem Formulation

The algorithm we propose is for computing a co-expression network among m possible entities (genes, miRNAs, etc.) measured across a set of samples organized into n groups. The groups may represent replicates of a condition, patients with a common disease, etc. Group g has n_g samples in it.

We observe R_{igs} reads for entity i in group g sample s. The total number of reads for that sample is $R_{gs} = \sum_i R_{igs}$. We assume that the R_{igs}, $i = 1 \ldots m$, are multinomially distributed.

$$Pr(R_{1gs}, R_{2gs}, \ldots, R_{mgs}) = \text{Multinom}(R_{gs}, p_{1gs}, p_{2gs}, \ldots, p_{mgs}) \ , \tag{1}$$

where the p_{igs} are unknown. Each p_{igs} represents the idealized fraction of the sample s in group g that comes from entity i. We can also think of it as what R_{igs}/R_{gs} should converge to in the limit of infinite sequencing depth ($R_{gs} \to \infty$). We define the group mean idealized fractions as $p_{ig} = \frac{1}{n_g} \sum_{s=1}^{n_g} p_{igs}$, and the grand mean idealized fraction as $p_i = \frac{1}{n} \sum_{g=1}^{n} p_{ig}$.

We take the p_{igs} to be our definition of the expression level. Other common definitions include reads per million (RPM), or fragments per kilobase per million (FPKM). Both of these normalize for sequencing depth in a given sample and are proportional to p_{igs}. As correlations are independent of scale, working with the p_{igs} is equivalent to working with RPM or FPKM. Other normalization schemes could be accomodated, as long as the expression level can be written as an affine function of the p_{igs}. However, so as not to overly complicate our notation, we leave this to the reader.

For any two entities i and j the cross-group Pearson correlation of their expression values is

$$r_{ij}^P = \frac{\text{cov}_g(p_{ig}, p_{jg})}{\sqrt{\text{var}_g(p_{ig})\text{var}_g(p_{jg})}} \ . \tag{2}$$

Ideally, we would like to connect entities i and j in a co-expression network if their cross-group correlation is statistically significantly large. The problem, of course, is that the p_{ig} are unknown, so we must estimate them.

The Bayesian Relevance Networks Algorithm

In principle, one could construct a Bayesian belief about the unknown Pearson correlation itself. However, this is not computationally convenient. Instead, we use Bayesian methods to construct estimates of the expression levels, p_{igs}, and then estimate their correlations. The algorithm we propose has four steps, which are detailed in the following subsections.

1. Remove low entropy entities from consideration (optional).
2. Compute Bayesian estimates of the cross-group correlations of expression between every (remaining) pair of entities
3. Use permutation computations to estimate a null distribution for the Bayesian cross-group correlations
4. Create a network by linking entities whose Bayesian correlations are statistically significant

Entropy filtering

This step is optional. We include it for the same reason it was included in the original Relevance Networks algorithm—that correlations may arise spuriously due to outliers. For instance, suppose two entities are generally expressed at constant levels, but in one sample both of their levels are much higher or lower than normal. These two entities will thus appear to have highly correlated expression levels. In some cases this may be genuinely true, although we may not be comfortable about the robustness of a correlation that depends on a single sample being present in the dataset. The same phenomenon might also arise for more mundane reasons, such as sample mishandling, contamination, poor sequencing depth, etc. Thus, it may make sense to remove entities with such expression profiles from consideration.

To allow for direct comparison between our Bayesian approach and the classic Relevance Networks algorithm, we use the exact same entropy filtering procedure. For each entity i, we compute the maximum likelihood expression estimates, $\hat{p}_{igs} = R_{igs}/R_{gs}$. We then compute the minimum, $A = \min_{gs} \hat{p}_{igs}$, and maximum, $B = \max_{gs} \hat{p}_{igs}$, expression levels across all samples in all groups. If $A = B$ then we estimate the entropy of entity i's expression as $H_i = 0$. Otherwise, we divide the interval $[A, B]$ into 10 equal-sized bins. We determine the empirical fraction of the $\hat{p}_{igs}$ that fall into each of those 10 bins, calling them $f_{i1} \dots f_{i10}$. We then estimate the entropy of entity i's expression as $H_i = -\sum_{j=1}^{10} f_{ij} \log_2 f_{ij}$. Entities with estimated entropies in the lowest $H_{thresh}\%$ are discarded, where H_{thresh} is chosen by the user.

Bayesian estimation of pairwise correlations

The essence of our Bayesian approach is to first construct beliefs over the true expression levels of all the entities. We then propose that the Pearson correlation between two entities be replaced by what we call the Bayesian correlation. We compute variances and covariances across groups and also with respect to our uncertainty about the true expression levels. Using u to denote our uncertainty informally—and we will become formal very shortly—the Bayesian correlation can be written as

$$r^B_{ij} = \frac{\mathrm{cov}_{g,u}(p_{ig}, p_{jg})}{\sqrt{\mathrm{var}_{g,u}(p_{ig})\mathrm{var}_{g,u}(p_{jg})}} \tag{3}$$

Intuitively, high uncertainty in expression levels may influence the covariance term, but it will definitely inflate the variance terms in the denominator, leading to lower estimates of correlation. (More precisely, estimates moderated towards zero.)

We adopt a standard Bayesian approach to estimate the idealized fractions p_{igs}. For each group g and sample s, we employ a Dirichlet distribution to model our uncertainty about the p_{igs}. We assume the Dirichlet beliefs for different samples are independent. Thus, for sample s and group g we adopt a prior belief,

$$\begin{aligned} Pr(p_{1gs}, p_{2gs}, \dots, p_{mgs}) &= \mathrm{Dirichlet}(\alpha^0_{1gs}, \alpha^0_{2gs}, \dots, \alpha^0_{mgs}) \qquad (4)\\ &= \frac{\Gamma(\sum_{i=1}^m \alpha^0_{igs})}{\prod_{i=1}^m \Gamma(\alpha^0_{igs})} \prod_{i=1}^m p_{igs}^{\alpha^0_{igs}} \ . \qquad (5) \end{aligned}$$

The posterior distribution is

$$\begin{aligned} & Pr(p_{1gs}, p_{2gs}, \dots, p_{mgs} | R_{1gs}, R_{2gs}, \dots, R_{mgs}) \qquad (6)\\ = \ & \mathrm{Dirichlet}(\alpha_{1gs}, \alpha_{2gs}, \dots, \alpha_{mgs}) \qquad (7)\\ = \ & \mathrm{Dirichlet}(\alpha^0_{1gs} + R_{1gs}, \alpha^0_{2gs} + R_{2gs}, \dots, \alpha^0_{mgs} + R_{mgs}) \ . \qquad (8) \end{aligned}$$

The prior parameters α^0_{igs} may be chosen however one likes. We previously showed that poor choice of priors can lead to highly biased estimates of correlation [33], and thus some care should be taken with the choice. We employ $\alpha_{igs} = 1/m$, which has provably low bias for low expression entities represented by few read counts [33]. For entities with high read counts, the prior makes little difference, as the posterior is determined almost entirely by the data. With these assumptions, and defining $\alpha_{gs} = \sum_{i=1}^{m} \alpha_{igs}$, the mean of the marginal posterior distribution for p_{igs} with respect to our beliefs (which we denote by u for "uncertainty") is

$$E_u \; p_{igs} = \frac{\alpha_{igs}}{\alpha_{gs}} \; . \tag{9}$$

The variance of that marginal posterior is

$$\mathrm{var}_u \; p_{igs} = \frac{\alpha_{igs}(\alpha_{gs} - \alpha_{igs})}{\alpha_{gs}^2(\alpha_{gs} + 1)} \; . \tag{10}$$

The covariance of our beliefs about the expression of two different entities, i and $j \neq i$, within the same sample s of group g is

$$\mathrm{cov}_u(p_{igs}, p_{jgs}) = \frac{-\alpha_{igs}\alpha_{jgs}}{\alpha_{gs}^2(\alpha_{gs} + 1)} \; . \tag{11}$$

This covariance is nonzero because of the implicit requirement that $\sum_{i=1}^{m} p_{igs} = 1$. Intuitively, if we believe that i's expression is larger, we must believe that the expression of other entities is smaller.

From these, we can readily compute the within-group means, variances and covariances between entities, accounting for our uncertainty. Recalling that by definition, p_{ig} is the average of p_{igs} across samples s, we have the following.

$$\begin{aligned} E_u \; p_{ig} &= E_u \sum_{s=1}^{n_g} \frac{1}{n_g} p_{igs} &(12)\\ &= \sum_{s=1}^{n_g} \frac{1}{n_g} \frac{\alpha_{igs}}{\alpha_{gs}} \; . &(13) \end{aligned}$$

$$\begin{aligned} \mathrm{var}_u \; p_{ig} &= \mathrm{var}_u \sum_{s=1}^{n_g} \frac{1}{n_g} p_{igs} &(14)\\ &= \frac{1}{n_g^2} \sum_{s=1}^{n_g} \mathrm{var}_u \; p_{igs} &(15)\\ &= \frac{1}{n_g^2} \sum_{s=1}^{n_g} \frac{\alpha_{igs}(\alpha_{gs} - \alpha_{igs})}{\alpha_{gs}^2(\alpha_{gs} + 1)} \; . &(16) \end{aligned}$$

Eq 15 follows because our estimates for different samples are statistically independent, so the variance of the sum is the sum of the variances.

$$\begin{aligned}
\mathrm{cov}_u(p_{ig}, p_{jg}) &= \mathrm{cov}_u\left(\sum_{s=1}^{n_g}\frac{1}{n_g}p_{igs}\ ,\sum_{s'=1}^{n_g}\frac{1}{n_g}p_{jgs'}\right) && (17)\\
&= \frac{1}{n_g^2}\mathrm{cov}_u\left(\sum_{s=1}^{n_g}p_{igs}\ ,\sum_{s'=1}^{n_g}p_{jgs'}\right) && (18)\\
&= \frac{1}{n_g^2}\sum_{s=1}^{n_g}\sum_{s'=1}^{n_g}\mathrm{cov}_u(p_{igs}, p_{jgs'}) && (19)\\
&= \frac{1}{n_g^2}\sum_{s=1}^{n_g}\mathrm{cov}_u(p_{igs}, p_{jgs}) && (20)\\
&= \frac{1}{n_g^2}\sum_{s=1}^{n_g}\frac{-\alpha_{igs}\alpha_{jgs}}{\alpha_{gs}^2(\alpha_{gs}+1)}\ . && (21)
\end{aligned}$$

Eq 20 follows because our beliefs are independent for different samples, hence there is no covariance when $s \neq s'$. We can then define the total variance across groups and uncertainty, for entity i, as

$$\begin{aligned}
\mathrm{var}_{g,u}\ p_{ig} &= \mathrm{var}_g E_u\ p_{ig} + E_g \mathrm{var}_u\ p_{ig} && (22)\\
&= \sum_{g=1}^{n}\frac{1}{n}(E_u\ p_{ig} - E_u\ p_i)^2 + \sum_{g=1}^{n}\frac{1}{n}\mathrm{var}_u\ p_{ig}\ . && (23)
\end{aligned}$$

Similarly, we define the total covariance across groups and uncertainty, for entities i and j, as

$$\begin{aligned}
\mathrm{cov}_{g,u}(p_{ig}, p_{jg}) &= \mathrm{cov}_g(E_u\ p_{ig}, E_u\ p_{jg}) + E_g\ \mathrm{cov}_u(p_{ig}, p_{jg}) && (24)\\
&= \sum_{g=1}^{n}\frac{1}{n}(E_u\ p_{ig} - E_u\ p_i)(E_u\ p_{jg} - E_u\ p_j) + \sum_{g=1}^{n}\frac{1}{n}\mathrm{cov}_u(p_{ig}, p_{jg}) && (25)
\end{aligned}$$

Eqs 23 and 25 can be substituted back into Eq 3 to completely specify the definition and computation of the Bayesian correlation. One step of this substitution and expansion is displayed below, as it will be relevant to our discussion of permutations in the next section.

$$\begin{aligned}
r_{ij}^B &= \frac{\mathrm{cov}_{g,u}(p_{ig}, p_{jg})}{\sqrt{\mathrm{var}_{g,u}(p_{ig})\mathrm{var}_{g,u}(p_{jg})}}\\
&= \frac{\mathrm{cov}_g(E_u\ p_{ig}, E_u\ p_{jg}) + E_g\ \mathrm{cov}_u(p_{ig}, p_{jg})}{\sqrt{(\mathrm{var}_g E_u\ p_{ig} + E_g \mathrm{var}_u\ p_{ig})(\mathrm{var}_g E_u\ p_{jg} + E_g \mathrm{var}_u\ p_{jg})}}\ . && (26)
\end{aligned}$$

A Permutation Scheme for Assessing Statistical Significance

Permutation testing is a common approach to assessing significance of associations between variables. However, in our context, this is not entirely straightforward. It is not sufficient to simply permute the read counts R_{igs} for each entity i and recompute Bayesian correlations. Recall that the estimated expression levels of entity i depend not only on R_{igs} but also on the total reads in the samples, R_{gs}. Permuting the read counts would change the R_{gs}, and therefore change the estimated expression levels. Permutation testing should "break" associations between different entities by reassigning their values to different samples, but it should not change the values themselves. It is also not sufficient to permute the estimated expression levels, $E_u\ p_{igs}$, as that could change estimated group expression levels, $E_u\ p_{ig}$.

With the null hypothesis being that there is no cross-group correlation between entities, we suggest that a proper way to estimate a null distribution between entities i and j is to compute many different permutations $\rho : \{1 \ldots n\} \mapsto \{1 \ldots n\}$ of the group numbers (all permutations, if possible). For each permutation ρ we evaluate the following formula.

$$r^{\rho}_{ij} = \frac{\mathrm{cov}_g(E_u\ p_{ig}, E_u\ p_{\rho(j)g}) + E_g\ \mathrm{cov}_u(p_{ig}, p_{jg})}{\sqrt{\mathrm{var}_{g,u}(p_{ig})\mathrm{var}_{g,u}(p_{jg})}} \tag{27}$$

The distribution of that value for many different permutations ρ is taken to be the null distribution of the Bayesian correlation.

In comparison with the formula for the Bayesian correlation (Eq 26), the permuted values of j's group-level expression are used in the first covariance term. This is the part of the formula where the hypothesis of no cross-group correlation would have its effect. We do not use the permuted j's in the second covariance term. That term represents the covariance of our beliefs within a sample, which results from the necessity that expression levels within a sample add up to one. This is not affected by the null hypothesis, so we leave it unchanged. The permutations also do not appear in the variance terms of the denominator, although it would not matter if they did, as the variances of i's and j's expression are independent.

Statistical Significance and Constructing the Bayesian Relevance Network

In the classical Relevance Networks algorithm, a single null distribution for correlations under the null hypothesis is constructed by combining the permuted correlations across all pairs of entities. Although it is technically more sound to maintain a separately estimated null distribution for each pair of entities (i, j), in order to maximize our ability to compare the results of Bayesian Relevance Networks to the classical algorithm, we do the same here. Thus, suppose that K times we have permuted the group idealized fractions, $E_u\ p_{ig}$, of every entity i, and recomputed the cross-group Bayesian correlations as in Eq 27. Let r^{ρ}_{ijk} represent the permuted Bayesian correlation between entities i and j in the k^{th} permutation. We estimate the overall probability of a correlation of at least t, under the null hypothesis, as

$$P(r \geq t) = \frac{|\{(i, j, k) :\ i < j \text{ and } r^{\rho}_{ijk} \geq t\}|}{Km(m-1)/2} \tag{28}$$

Suppose we construct a Bayesian Relevance Network by connecting any pair of entities i and j if their Bayesian correlation is at least t, obtaining N_t such pairs. Given that there are $m(m-1)/2$ possible pairs of entities, we can estimate the expected number of false positives at that threshold as $FP_t = P(r \geq t)m(m-1)/2$. The number of true positives can be estimated as $\max(N_t - FP_t, 0)$. The false discovery rate can be estimated as $\min(FP_t/N_t, 1)$, as long as $N_t > 0$. Together, these quantities—estimated numbers of true positives, numbers of false positives, and the false discovery rate—can be employed by the user to make a rational choice for the threshold t used to construct the network.

Data

To demonstrate and evaluate our approach, and potentially to generate some biological insights in an important area, we decided to analyze miRNA expression data from The Cancer Genome Atlas (TCGA) [41]. We used the Genomic Data Commons data portal [19] to download all available "isoforms.quantification.txt" files on November 10, 2016. These files report counts of miRNA-seq reads mapped to a large number of genomic intervals. Those intervals are also annotated for whether they represent a certain pre-miRNA, a mature miRNA, or several other types of objects. From each file, we collected all lines corresponding to a mature miRNA (specified by a unique miRBase [18] MIMAT identifier), and then added up all counts corresponding to the same mature miRNA. This includes reads mapped to slightly different genomic intervals within the same mature miRNA, as well as entirely different genomic regions that happen to code for the same mature miRNA. In the end, this left us with read counts for 2456 distinct mature miRNAs, across 10,999 patient samples.

While this gave us a wealth of data on miRNA expression in cancer, the isoform files do not specify which types of cancer each patient had (nor any other patient characteristics). To establish this information, we constructed a json query that, through the Genomic Data Commons API, returned a list of all isoform quantification files, along with their project IDs. The project IDs are synonymous with the types of cancer profiled. In this way, we assigned one of 33 unique cancer types to each miRNA-seq dataset.

In order to better inform our co-expression assessments, we downloaded from miRbase [18] their version 21 miR definitions in the file "hsa.gff3". This file specifies the IDs and genomic coordinates of both stem-loop pre-cursors and mature miRNAs. It also specifies which mature miRNAs are to be found in which stem-loop precursors. Multiple genomic occurrences of the same mature miRNA have IDs ending in _1, _2, etc., to discriminate them. However, the "Alias" field omits these IDs, which could then be matched to the MIMAT IDs in the TCGA isoforms file. Similarly, we downloaded from ENSEMBL their latest gene definitions in the file "Homo_sapiens.GRCh38.86.gtf". This file describes many types of transcribed entities, including protein-coding genes, pseudogenes, long non-coding RNAs, miRNAs, etc. Importantly, it includes their genomic locations. Using these sources of information, we were able to categorize every pair of mature miRNAs into one of the following categories: (1) "stem-loop" if the two mature miRNAs occur within the same stem-loop precursor miRNA anywhere in the genome; (2) "transcript" if the two mature miRNAs occur within the same transcribed entity (according to ENSEMBL) but not the same stem-loop precursor; (3) "near" if the two mature miRNAs occur within 10kb on the genome; (4) "cluster" if the two mature miRNAs occur within the same equivalence class in the transitive closure of the "near" relation, but are not themselves "near". For example, if i is near j and j is near k, but i and k are not near, then i and k are still in the same cluster; (5) "non-local" if none of the previous categories apply.

Results

TCGA miRNA expression data spans many orders of magnitude across miRNAs and samples

As described in the Methods section, we obtained miRNA-seq expression data from the TCGA project through the Genomic Data Commons, resulting in read counts for 2456 miRNAs in 10,999 patient samples, representing 33 cancer types. The data is shown in Fig 1A. Each row corresponds to a miRNA, and each column corresponds to a patient sample. The most-represented cancer was breast cancer, with 1207 samples, while the least-represented was glioblastoma multiforme, with 5 samples. There are clearly miRNAs with cancer-specific, or at least tissue-specific, expression profiles. Fig 1B shows the average expression in units of RPM for the top 20 most highly expressed miRNAs. The most highly expressed miRNA is mir-21-5p, with an RPM over 200,000,

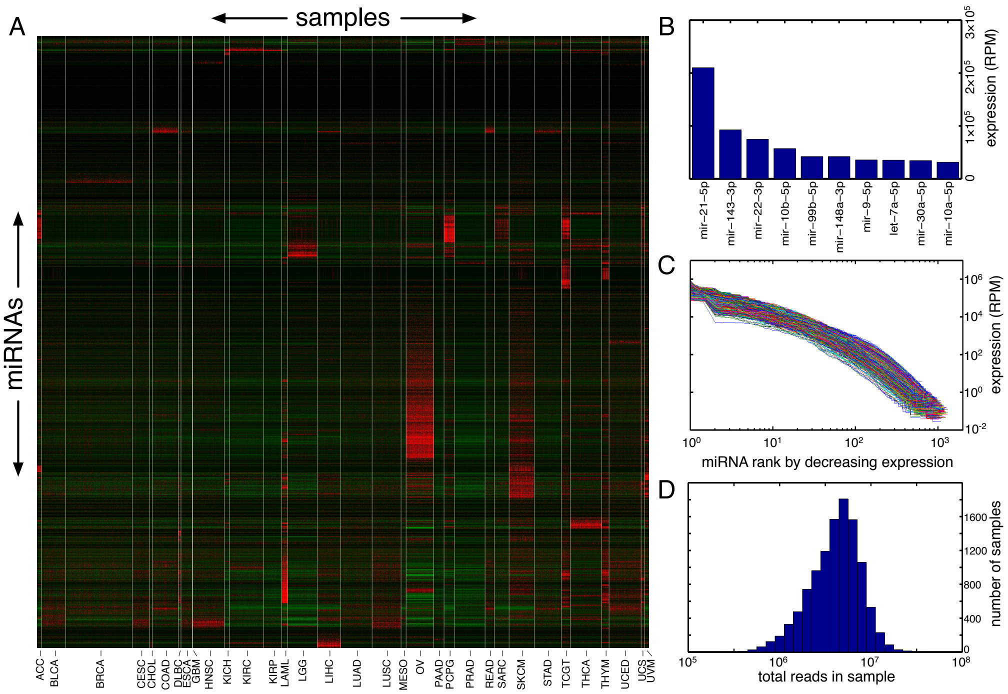

Fig 1. Mature miRNA expression data for 10,999 cancer patients from the TCGA project. (A) Heatmap of expression, with red indicating high and green indicating low, relative to the mean for each miRNA across samples. miRNAs are ordered based on a hierarchical average-linkage Euclidean-distance clustering of the reads per million across samples. Samples are grouped by cancer type, indicated by labels along the bottom. (B) Average expression across samples of the 20 highest-expressed miRNAs. (C) Curves showing expression of all miRNAs within each sample, sorted from highest to lowest expression. (D) Histogram of the numbers of reads (i.e., sequencing depth) in each sample.

meaning it comprises more than 20% of the total miRNA pool on average. This miRNA is well known for its role in oncogenesis and metastasis [5, 15, 36].

Fig 1C shows the expression of every miRNA in every sample, sorted by decreasing order within the sample. Expression values range from around 10^5 RPM to below 1 RPM. Because all miRNAs are measured in the same units—reads—this means that relative to their expression levels, the miRNAs with lowest expression are measured with approximately 1/100,000 the precision of the miRNAs with highest expression. There are also great differences in sequencing depth between samples, as shown in Fig 1D. The sample with the greatest sequencing depth has over 36 million reads, while the sample with the shallowest sequencing depth has under a quarter million. There is approximately a 150-fold difference in resolution between these two samples. Given these statistics, it is clear that our uncertainties about the true expression levels of the miRNAs must vary widely by miRNA and by sample.

Bayesian correlations are more reproducible than Pearson correlations

We expected that Bayesian correlation estimates would suppress correlations between low expression miRNAs. By contrast, we expected that Pearson correlations would be more subject to falsely high or low correlations, due to spurious correlations between miRNAs with low read counts. To test this, we computed the Bayesian and Pearson correlations across cancer types for all miRNA pairs. For the Pearson correlations, this was the correlation across cancers of the with-cancer average expression in units of RPM. Fig 2A shows a density scatterplot of the Pearson and Bayesian correlations. Points along the $y = x$ diagonal line correspond to miRNA pairs where Pearson and Bayesian estimates agreed. We note that there are some miRNA pairs correlated at essentially +1 by both Pearson and Bayesian estimates, but no miRNA pairs with such strong anticorrelations. At the same time, there are many miRNA pairs that have high correlations according to the Pearson estimate, but that are relegated to much lower correleation levels—including essentially zero—by the Bayesian estimate. These involve miRNAs that, by our approach, have too much uncertainty in their expression levels to be able to confidently assert a strong correlation. As a rather extreme example, there was a strong disagreement in the estimated correlations between miR-4459 and miR-5692b. The former shows expression in 70 different samples across 12 cancer types, but is primarily seen in thyroid cancers, albeit at low levels (53 samples, 139 total reads). The latter is expressed at only 2 reads in a single thyroid cancer sample, and nowhere else. The Pearson correlation between these two is a near perfect 0.9731, whereas the Bayesian correlation is 0.0512.

To test the reproducibility of Pearson and Bayesian correlations, we randomly assigned each sample to one of two data folds, keeping the numbers of samples representing each cancer type as even as possible. We then computed cross-cancer Pearson correlations on each half of the data separately (Fig 2B), and likewise for the Bayesian correlations (Fig 2C). For the Pearson correlations, there is broad agreement between correlations computed based on each fold of the data—the estimates from each half are themselves correlated. But there are also many miRNA pairs where correlations from the two folds disagree dramatically. For a substantial number of pairs, one fold of the data produces a Pearson correlation near 1, while the other fold produces a Pearson correlation near zero. The two “lines” visible along the x- and y-axes of the density scatterplot arise from miRNAs that have absolutely zero reads in one fold of the data (hence no correlation to anything), but some reads in the other fold (and in some cases strong correlations, although they may be spurious). In comparison, the Bayesian correlation estimates from each fold of the data tend to be closer to each other. There are no “lines” of exceptional behaviour for zero-count miRNAs, and no miRNA pairs with near zero Bayesian correlation in one fold and near +1 Bayesian correlation in the other fold (although there are a very few near 0.9).

To quantify the reproducibility of the two approaches more carefully, and also to study the relationship between expression level and correlations, we divided miRNAs into 21 bins of increasing average RPM expression. Let X denote the set of miRNAs in one expression bin, and Y denote the set of miRNAs in another expression bin. From data fold 1, we computed all pairwise Pearson correlations between miRNAs in bin X with those miRNAs in bin Y, namely, $\{r_{xy}^{P1} : x \in X, y \in Y\}$. We did the same for data fold 2, compute the correlations $\{r_{xy}^{P2} : x \in X, y \in Y\}$. Finally, we computed the mean absolute deviation between these two sets of correlations, $\mathrm{MAD}(X, Y) = \sum_{x \in X, y \in Y} |r_{xy}^{P1} - r_{xy}^{P2}|/|X||Y|$. This gives the average disagreement of Pearson correlations computed from the two data folds, as a function of binned expression level. Then, we did the same for the Bayesian correlations. Fig 2D shows those mean absolute deviations. Generally, as the expression of both miRNAs trends higher, the disagreement between the two halves of the data decreases, and the error in the Pearson and Bayesian estimates is essentially identical. For these miRNAs, low signal-to-noise ratio is not an issue, and Pearson and Bayesian estimates are nearly the same. Error is worst when both miRNAs have low but nonzero expression, and it is

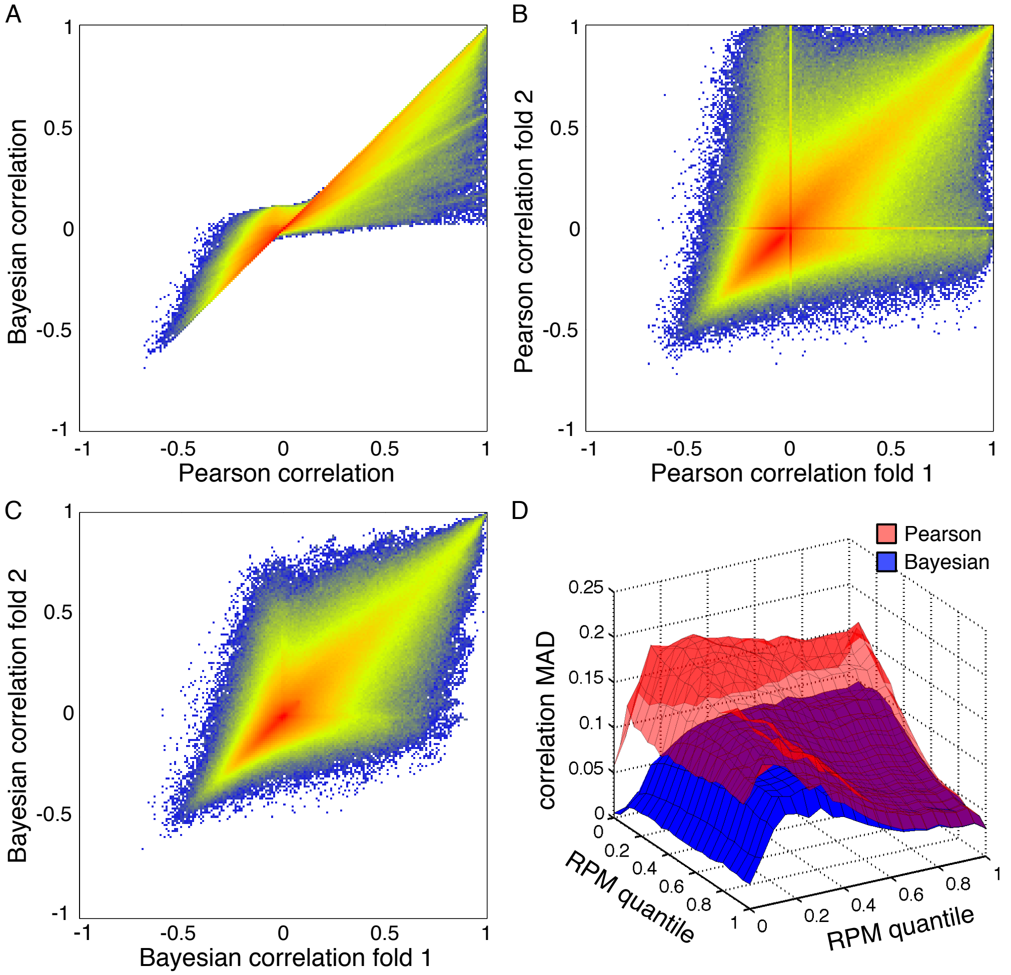

Fig 2. Comparison of Pearson and Bayesian grouped correlations across cancer types. (A) Density scatterplot of Bayesian versus Pearson correlations. Non-white points are where at least one pair of miRNAs has the specified Pearson (x-axis) and Bayesian (y-axis) correlations. Colored points, going from blue to yellow to red, indicate increasing numbers of miRNA pairs with the specified correlations. (B) Agreement of Pearson correlations when the data is divided in half and correlations computed for each half separately. (C) Agreement of Bayesian correlations when the data is divided in half and correlations computed for each half separately. (D) For each pair of miRNAs, organized by their expression quantiles across all samples, the average mean absolute deviation (MAD) between the two data halves of Pearson and Bayesian correlations.

nearly as bad when just one of the two miRNAs has low but nonzero expression. When one of the miRNAs is in the lowest expression bin, error tends not to be quite as bad, as both methods will tend to assign zero correlation (but Bayesian more so than Pearson). At all levels of expression, the average error of the Pearson estimates exceeds the error of the Bayesian estimates. Across all pairs of miRNAs, the Pearson MAD is 0.1304 between folds, and the Bayesian MAD is 0.0843, a difference that is statistically significant by a simple sign test at a p-value too small for machine precision (easily $p < 10^{-100}$).

Entropy filtering improves reproducibility of both Pearson and Bayesian correlations

As described in the Introduction, the classical Relevance Networks algorithm begins by filtering out entities whose expression demonstrates low entropy. The purpose of this step is to avoid correlations that arise from a single sample or small set of "outliers." Whether or not such an approach is appropriate is situation dependent. For example, if a subset of miRNAs were highly expressed only in glioblastoma multiforme tumours, and no others, such miRNAs would appear to have low entropy. (Remember, just five out of our 10,999 samples are for that disease.) We may not want to naively dismiss correlations among such miRNAs, as they arise from a clear disease relevance. Nevertheless, in the worst case, individual samples may be faulty and can create spurious correlations.

To test the effect of entropy filtering on both the Pearson and Bayesian correlations, we first computed the entropy of each miRNA's expression (Fig 3A). In the original paper [10], it was suggested to discard the 5% of entities with lowest entropy (dashed red line). However, the appearance of the empirical entropy distribution suggested to us cut off around 10% (solid red line) would better separate entities with a "normal range" of entropies from those that appear unusually low. Hence, we chose 10% as our cut off, and defined miRNAs with entropies below that to be "low entropy" and the remainder to be "high entropy." Fig 3B examines the relationship between miRNA expression and entropy. For the most part, the low entropy miRNAs also have very low expression. However, a small number of miRNAs with above average expression also have low entropy. The miRNA with the highest average expression that is still classified as low entropy is miR-205-3p, a miRNA with some known associations with cancer [12, 22, 40]. This miRNA is exceptionally high in two patient samples, one thymoma and one head or neck squamous cell carcinoma, where its expression levels of over 10,000 RPM are more than 100 times greater than in any other sample.

Restricting attention to the high-entropy genes, and we recomputed the density scatterplots of Pearson correlations from the two halves of our data (Fig 3C), we see that the lines of exceptional correlations along the x- and y-axis are gone. (Compare to Fig 2B.) However, the overall qualitative shape of the point cloud remains, as do numerous miRNA pairs that have near +1 correlation in one half of the data and near zero correlation in the other half. Fig 3D shows the Bayesian correlations of the high-entropy miRNAs from each half of the data. There is little apparent change compared to Fig 2C, which includes the low entropy miRNAs. Perhaps surprisingly, entropy filtering does not improve the mean absolute deviation between the two halves of the data. For Pearson correlations restricted to high-entropy miRNAs, the MAD is 0.1305 (versus 0.1304 for all miRNAs), and for Bayesian correlations the MAD is 0.0894 (versus 0.0843). Although filtering eliminates some spurious correlations, it also eliminates many (correctly) zero correlations between low- or non-expressed miRNAs, driving the average error up.

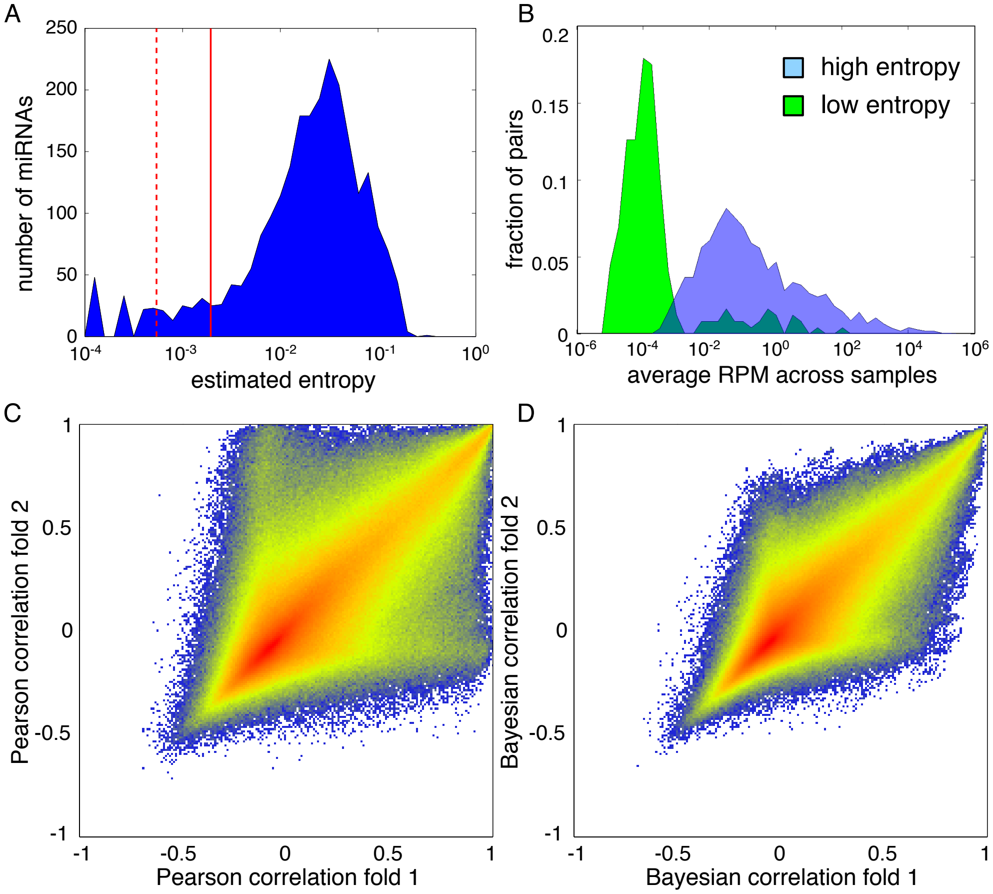

Fig 3. The effects of entropy filtering on Pearson and Bayesian correlations. (A) Empirical distribution of entropies of miRNAs' expression across samples. Dashed red line indicates 5^{th} percentile and solid red line indicates 10^{th} percentile. (B) Empirical distribution of expression levels (average RPM across samples) for low entropy and high entropy miRNAs. (C) Comparison of Pearson grouped correlations from two halves of the data, when restricting attention to the high entropy miRNAs. (D) Comparison of Bayesian grouped correlations from two halves of the data, when restricting attention to the high entropy miRNAs.

Bayesian Relevance Networks have lower false discovery rates

As Bayesian correlations between miRNAs match better between data folds than do Pearson correlations, we predicted that Bayesian Relevance Networks built based on each half of the data would agree better than classical Relevance Networks would. To test this hypothesis, we performed permutation testing on each half of the data, estimating null distributions for both the Pearson correlations and the Bayesian correlations based on all miRNAs (not just the high-entropy ones). The results are shown in Fig 4A,B. The blue curves indicate the observed distributions of correlations on each half of the data, while the red curves indicate the estimated null distributions. For both Pearson and Bayesian correlations, there appear to be stronger positive correlations than would be predicted based on the null hypothesis of no statistical association between miRNAs. The shapes of the distributions estimated from each half of the data are in close agreement. There are more Pearson correlations at the highest levels (near 1) than there are Bayesian correlations—because of the tendency of the Bayesian approach to discount apparent correlations between low expression miRNAs.

Next, we constructed Relevance Networks at different correlation thresholds. At each threshold, we determined the number of miRNA pairs above threshold, as well as the expected number of such pairs under the null hypothesis. Based on these, we estimated the false discovery rate (FDR) for links in the Relevance Networks as a function of correlation threshold. At the same time, we compared the specific links constructed from each half of the data to the links in the other half. Links appearing in one half but not the other were labeled as putative false positives, and from these we constructed a second estimate of the FDR as a function of correlation threshold. The results are shown in Fig 4C,D and are radically different for Pearson and Bayesian approaches. Firstly, the Bayesian FDRs are uniformly better than the Pearson FDRs, especially at higher correlation thresholds. The estimated Pearson FDRs from permutation testing hover around 0.2 for most correlation thresholds, whereas estimated Bayesian FDRs are smaller than 0.15. The empirical Pearson FDRs, based on comparing the networks obtained from each half of the data, are worse than 0.4 at all thresholds. The empirical Bayesian FDRs are somewhat different between the two folds of the data, but average to around 0.3 at most thresholds. The Bayesian FDR estimates either improve (drop) with increasing correlation threshold (permutation-based) or are relatively constant (based on data folds). This is a reasonable behaviour, as increasing the threshold intuitively means increasing stringency. However, Pearson FDRs actually get worse at the highest thresholds, as the relative number of spurious correlations from low-expression entities grows.

We then repeated the entire experiment while restricting attention to the high-entropy miRNAs only. Fig 4E,F shows the percentage improvement in empirical FDR for Pearson and Bayesian approaches—as quantified by $(FDR_{all} - FDR_{he})/FDR_{all}$, where FDR_{all} is the false discovery rate when analyzing all miRNAs, and FDR_{he} is the false discovery rate when analyzing only the high-entropy miRNAs. The Pearson approach benefits modestly from the entropy filtering, especially at the higher correlation thresholds, where improvements of up to 20% can be seen. However, its performance still does not reach that of the Bayesian approach. The false discovery rate of Bayesian Relevance Networks seems almost entirely immune to entropy filtering (Fig 4F).

A Bayesian Relevance Network describing co-expression of miRNAs across 10,999 patients with 33 types of cancer

Having established the soundness of the Bayesian Relevance Networks algorithm in the previous sections, we conclude the Results section by presenting the Bayesian Relevance

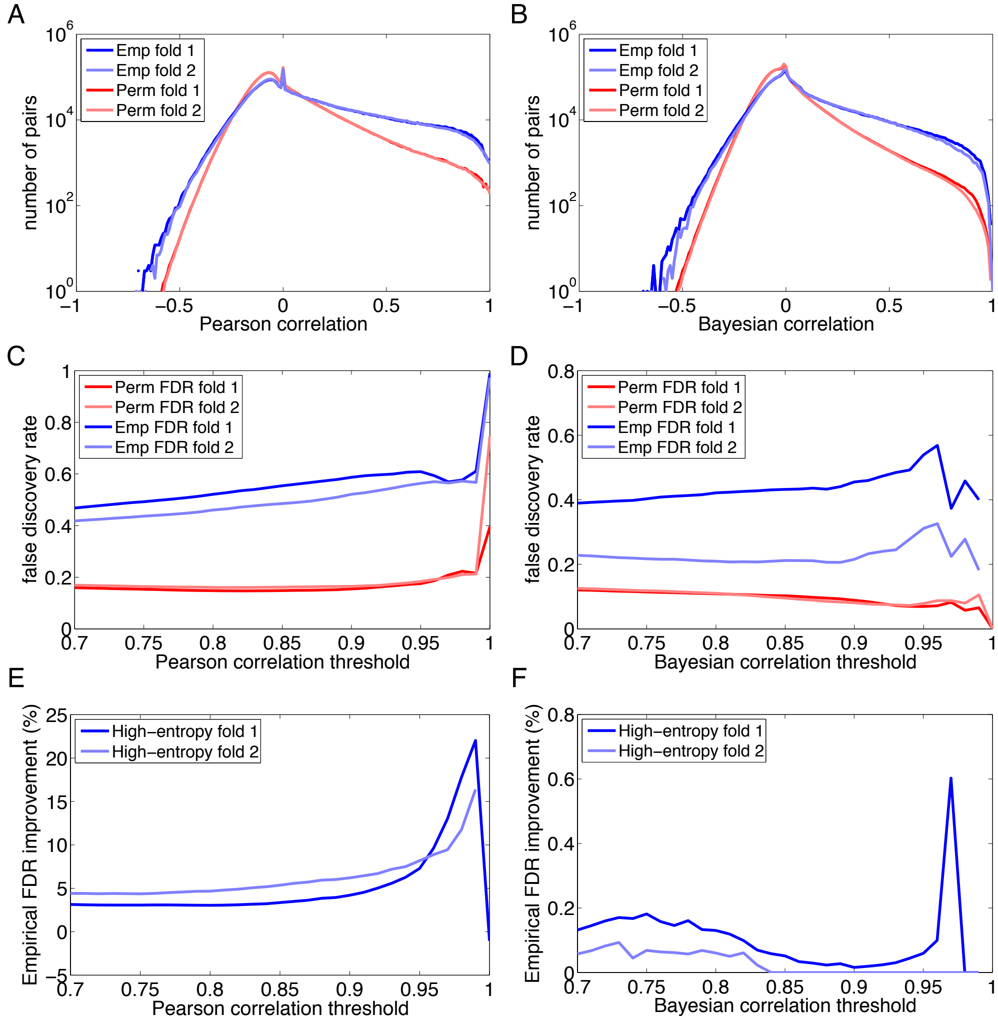

Fig 4. Permutation testing and agreement of Relevance Networks constructed based on Pearson or Bayesian correlations. (A) Empirical (blue) and permutation-based (red) distributions of Pearson correlations from each half of the data. (B) Empirical (blue) and permutation-based (red) distributions of Bayesian correlations from each half of the data. (C) Estimated false discovery rates (blue, based on permutations) and empirical false discovery rates (red, taking other half of the data as gold standard) at varying Pearson correlation thresholds. (D) Estimated false discovery rates (blue, based on permutations) and empirical false discovery rates (red, taking other half of the data as gold standard) at varying Bayesian correlation thresholds. (E) Percent improvement in empirical FDR when restricting attention to high-entropy genes, as a function of Pearson correlation threshold. (F) Percent improvement in empirical FDR when restricting attention to high-entropy genes, as a function of Bayesian correlation threshold.

Network obtained by analyzing the full dataset. We chose not to filter out miRNAs based on low entropy, so that we would not overlook potentially interesting connections, and because our results above suggest there would be little benefit. Accordingly, we computed all pairwise Bayesian correlations, and we performed 100 permutation computations to assess statistical significance. The empirical distributions of actual and permuted Bayesian correlations are shown in Fig 5A. As expected, we see many miRNA pairs that are highly correlated. However, high correlation can also be obtained by chance, as shown by the permutation testing. Even at a threshold of $r = 0.99$, which links just 60 miRNA pairs, our permutation testing suggests that four of those would be false positives.

We decided to construct the relevance network at the threshold $r = 0.96$. This gave us 1479 links between 338 distinct miRNAs, with an estimated 95 false positive links, or an empirical false discovery rate of 6.5%. We chose this level because it produced a large enough relevance network to see some interesting results, without letting the FDR grow too far out of control. The network is depicted in Fig 5B. We used Cytoscape [35] to construct the layout of the network. Links are colored by their locality: blue for miRNAs in the same pre-miRNA stem-loop, red for miRNAs in the same transcript, light green for miRNAs nearby on the genome, dark green for miRNAs in the same genomic cluster, and black for those not having any of those locality properties. As is typical for relevance networks, and indeed many types of biological networks, we observe connected components of widely varying sizes. Several major components have tens of miRNAs each, heavily cross-connected, while there are also many isolated pairs of miRNAs connected by a single link. The majority of the links do not represent any locality relationship (Fig 5C).

A typical cluster is indicated by *(i)* in Fig 5B. Only a few links are related to genomic locale; most of the miRNAs are spread throughout the genome. miRNAs in this subnetwork are highly expressed in acute myeloid leukemia (TCGA code LAML) (Fig 5D). We found that many of the other connected subnetworks are also highly expressed in just one or a few cancer (or tissue) types.

A notable subnetwork is the “C”-shaped one in the upper left of the layout. This includes many miRNAs that are nearby on the genome (within 10kb) or at least within the same genomic cluster. However, the most densely connected part of the subnetwork, towards the bottom of the “C”, contains a mixture of stem-loop, transcript, local and non-local links. When we analyze miRNAs in three different parts of that network, we see different expression patterns (Fig 5E) The mostly-back cluster at the bottom is expressed almost exclusively in testicular germ cell tumors. At the opposite end of the “C”, the dense genomic cluster in green is expressed somewhat in testicular tumors but primarily in thymomas. miRNAs in between those two ends display a mixture of testicular tumor and thymoma expression. These miRNAs comprise the primate-specific C19MC miRNA cluster, which has normal functions in the placenta [29, 39]. This cluster’s roles in various cancers are still being worked out [6, 24, 32, 38].

Although one must zoom in on the figure to see clearly, the vast majority of the links between isolated pairs of miRNAs do have some kind of locality relationship—unlike the majority of links in the network. Nearly half of the isolated miRNAs pairs are in the same stem-loop (11 of 23), five are in the same transcript, and five are nearby on the genome. Only two links are non-local, between miR-1180-3p and miR-6511b-3p, and between miR-548d-3p and miR-3613-5p (Fig 5F). These pairs show some evidence of cancer/tissue-specificity, with the first pair largely expressed in glioblastoma multiforme and ovarian cancer samples, and the latter pair largely expressed in acute myeloid leukemia samples and thymomas.

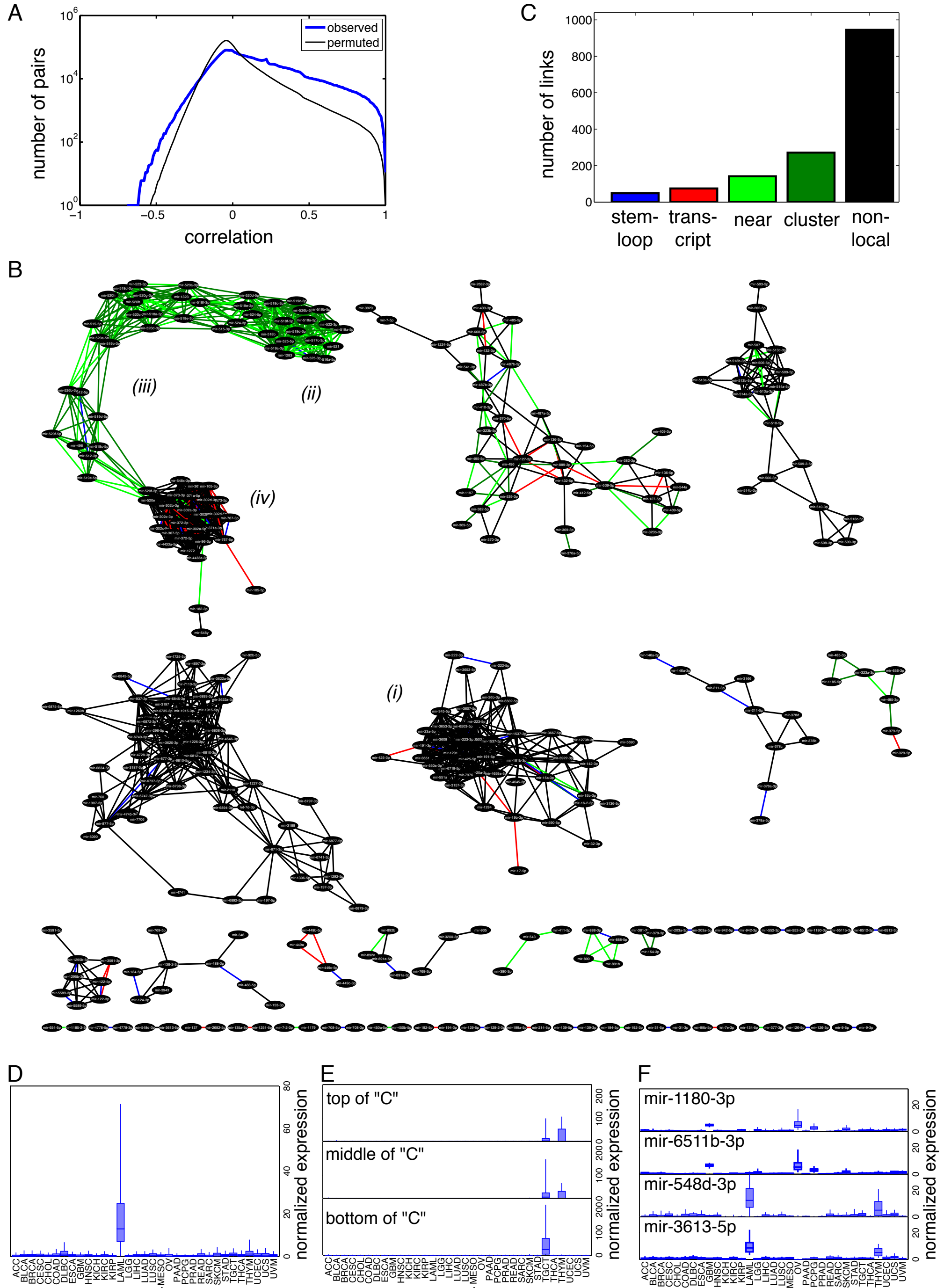

Fig 5. A Bayesian Relevance Network describing cross-cancer correlations between miRNAs. (A) Empirical distributions of Bayesian and permuted correlations. (B) The network obtained at a correlation threshold of $r = 0.96$. (C) Numbers of links with different locality relationships. (D) Normalized expression of miRNAs in the mostly-black subnetwork *(i)* near the center of the diagram in panel B. (E) Normalized expression of miRNAs at the top *(ii)*, middle *(iii)*, and bottom *(iv)* of the "C"-shaped subnetwork in the top left of panel A. (F) Normalized expression of four miRNAs participating in the only two non-local miRNA pairs in the relevance network.

Discussion

In this work, we have proposed Bayesian Relevance Networks as an update to the classical and widely-used Relevance Networks algorithm [10], with the aim of making it better suited to high-throughput sequencing data. Our approach accounts for the fact that sequence-based expression measurements can have widely varying precision, both for different entities (e.g., genes or miRNAs) and for different samples. It builds on our recent proposal for Bayesian correlation analysis [33], adding two main ingredients helpful for the construction of co-expression networks: 1) a method for estimating uncertainties in the expression levels in groups of samples; and 2) a permutation-testing scheme to assess statistical significance of Bayesian correlations. In testing on a large-scale miRNA expression dataset from The Cancer Genome Atlas [41], we found that Bayesian estimates of co-expression were more reproducible than the Pearson estimates used in the classical algorithm. As a consequence, we found that Bayesian Relevance Networks had lower false discovery rates than standard Relevance Networks. We also found that the entropy filtering step, with its additional and arbitrary cut off parameter, is unnecessary in the Bayesian approach, leading to a simpler algorithm over all. Although we focused on this single, large-scale dataset for demonstration and empirical evaluation, an important direction for future work is testing on other datasets. We suspect that one area where Bayesian Relevance Networks will be particularly helpful is in the analysis of single-cell RNA-seq data [21]. In such datasets, the average number of reads per gene are much smaller than for bulk RNA-seq data, and there can be great variability in the sequencing depths for each cell. This is exactly the situation where uncertainties in expression levels need to be considered, and where Bayesian approaches can provide a solution.

As mentioned in the introduction, since the publication of the original Relevance Networks algorithm, many other algorithms have been proposed for the construction of co-expression networks [1, 8, 9, 14, 34, 42]. All these algorithms contain important insights about the assessment of co-expression. Although we chose in this paper to develop a Bayesian version of the Relevance Networks algorithm—the “grandfather” of all co-expression algorithms—an important avenue for future research is incoporating similar notions of Bayesian reasoning about expression levels and uncertainty into other co-expression network construction algorithms.

References

1. Lipi R Acharya and Dongxiao Zhu. Estimating an optimal correlation structure from replicated molecular profiling data using finite mixture models. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 119–124. IEEE, 2009.

2. Peter AC't Hoen, Yavuz Ariyurek, Helene H Thygesen, Erno Vreugdenhil, Rolf HAM Vossen, Renee X de Menezes, Judith M Boer, Gert-Jan B van Ommen, and Johan T den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic acids research*, 36(21):e141–e141, 2008.

3. Peter AC't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen FJ Laros, Henk PJ Buermans, Olof Karlberg, Mathias Brännvall, et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*, 31(11):1015–1022, 2013.

4. Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5(1):1, 2004.

5. IA Asangani, SAK Rasheed, DA Nikolova, JH Leupold, NH Colburn, S Post, and H Allgayer. MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene*, 27(15):2128–2136, 2008.

6. Claudia Augello, Valentina Vaira, Luca Caruso, Annarita Destro, Marco Maggioni, Young Nyun Park, Marco Montorsi, Roberto Santambrogio, Massimo Roncalli, and Silvano Bosari. Microrna profiling of hepatocarcinogenesis identifies c19mc cluster as a novel prognostic biomarker in hepatocellular carcinoma. *Liver International*, 32(5):772–782, 2012.

7. Scott Baskerville and David P Bartel. Microarray profiling of micrornas reveals frequent coexpression with neighboring mirnas and host genes. *Rna*, 11(3):241–247, 2005.

8. Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

9. Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, page 26, 2000.

10. Atul J Butte, Pablo Tamayo, Donna Slonim, Todd R Golub, and Isaac S Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

11. Shih-Hwa Chiou, Mong-Lien Wang, Yu-Ting Chou, Chi-Jen Chen, Chun-Fu Hong, Wang-Ju Hsieh, Hsin-Tzu Chang, Ying-Shan Chen, Tzu-Wei Lin, Han-Sui Hsu, et al. Coexpression of oct4 and nanog enhances malignancy in lung adenocarcinoma by inducing cancer stem cell–like properties and epithelial–mesenchymal transdifferentiation. *Cancer research*, 70(24):10433–10444, 2010.

12. Indra N Dahmke, Christina Backes, Jeannette Rudzitis-Auth, Matthias W Laschke, Petra Leidinger, Michael D Menger, Eckart Meese, and Ulrich Mahlknecht. Curcumin intake affects mirna signature in murine melanoma with mmu-mir-205-5p most significantly altered. *PLoS One*, 8(12):e81122, 2013.

13. Laura L Elo, Henna Järvenpää, Matej Orešič, Riitta Lahesmaa, and Tero Aittokallio. Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process. *Bioinformatics*, 23(16):2096–2103, 2007.

14. Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biol*, 5(1):e8, 2007.

15. Lisa B Frankel, Nanna R Christoffersen, Anders Jacobsen, Morten Lindow, Anders Krogh, and Anders H Lund. Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *Journal of Biological Chemistry*, 283(2):1026–1033, 2008.

16. Anna Git, Heidi Dvinge, Mali Salmon-Divon, Michelle Osborne, Claudia Kutter, James Hadfield, Paul Bertone, and Carlos Caldas. Systematic comparison of microarray profiling, real-time pcr, and next-generation sequencing technologies for measuring differential microrna expression. *Rna*, 16(5):991–1006, 2010.

17. Luciana I Gomes, Gustavo H Esteves, Alex F Carvalho, Elier B Cristo, Roberto Hirata, Waleska K Martins, Sarah M Marques, Luiz P Camargo, Helena Brentani, Adriane Pelosof, et al. Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism. *Cancer research*, 65(16):7127–7136, 2005.

18. Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. mirbase: tools for microrna genomics. *Nucleic acids research*, 36(suppl 1):D154–D158, 2008.

19. Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

20. Shimin Hu, Zijun Y Xu-Monette, Alexander Tzankov, Tina Green, Lin Wu, Aarthi Balasubramanyam, Wei-min Liu, Carlo Visco, Yong Li, Roberto N Miranda, et al. Myc/bcl2 protein coexpression contributes to the inferior survival of activated b-cell subtype of diffuse large b-cell lymphoma and demonstrates high-risk gene expression signatures: a report from the international dlbcl rituximab-chop consortium program. *Blood*, 121(20):4021–4031, 2013.

21. Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

22. Min Jiang, Peng Zhang, Guozhu Hu, Zuke Xiao, Fanghua Xu, Ting Zhong, Fang Huang, Haibin Kuang, and Wei Zhang. Relative expressions of mir-205-5p, mir-205-3p, and mir-21 in tissues and serum of non-small cell lung cancer patients. *Molecular and cellular biochemistry*, 383(1-2):67–75, 2013.

23. Wei Jiang, Xia Li, Shaoqi Rao, Lihong Wang, Lei Du, Chuanxing Li, Chao Wu, Hongzhi Wang, Yadong Wang, and Baofeng Yang. Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. *BMC systems biology*, 2(1):1, 2008.

24. Claudia L Kleinman, Noha Gerges, Simon Papillon-Cavanagh, Patrick Sin-Chan, Albena Pramatarova, Dong-Anh Khuong Quang, Véronique Adoue, Stephan Busche, Maxime Caron, Haig Djambazian, et al. Fusion of ttyh1 with the c19mc microrna cluster drives expression of a brain-specific dnmt3b isoform in the embryonal brain tumor etmr. *Nature genetics*, 46(1):39–44, 2014.

25. Yoshiyuki Kubota, Naoki Shigematsu, Fuyuki Karube, Akio Sekigawa, Satoko Kato, Noboru Yamaguchi, Yasuharu Hirai, Mieko Morishima, and Yasuo Kawaguchi. Selective coexpression of multiple chemical markers defines discrete populations of neocortical gabaergic neurons. *Cerebral cortex*, 21(8):1803–1817, 2011.

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

27. Niall J Lennon, Alvin Kho, Brian J Bacskai, Sarah L Perlmutter, Bradley T Hyman, and Robert H Brown. Dysferlin interacts with annexins a1 and a2 and mediates sarcolemmal wound-healing. *Journal of Biological Chemistry*, 278(50):50466–50473, 2003.

28. Mathieu Lupien, Jérôme Eeckhoute, Clifford A Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S Carroll, X Shirley Liu, and Myles Brown. Foxa1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6):958–970, 2008.

29. Noguer-Dance Marie, Abu-Amero Sayeda, Al-Khtib Mohamed, Lefèvre Annick, Coullin Philippe, Moore E Gudrun, and Jérôme Cavaillé. The primate-specific microrna gene cluster (c19mc) is imprinted in the placenta. *Human molecular genetics*, page ddq272, 2010.

30. Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3):243–248, 2008.

31. Igor Ponomarev, Shi Wang, Lingling Zhang, R Adron Harris, and R Dayne Mayfield. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. *The Journal of Neuroscience*, 32(5):1884–1897, 2012.

32. Volkhard Rippe, Lea Dittberner, Verena N Lorenz, Norbert Drieschner, Rolf Nimzyk, Wolfgang Sendt, Klaus Junker, Gazanfer Belge, and Jörn Bullerdiek. The two stem cell microrna gene clusters c19mc and mir-371-3 are activated by specific chromosomal rearrangements in a subgroup of thyroid adenomas. *PloS one*, 5(3):e9485, 2010.

33. Daniel Sánchez-Taltavull, Parameswaran Ramachandran, Nelson Lau, and Theodore J Perkins. Bayesian correlation analysis for sequence count data. *PloS one*, 11(10):e0163595, 2016.

34. Juliane Schäfer, Korbinian Strimmer, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

35. Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

36. ML Si, S Zhu, H Wu, Z Lu, F Wu, and YY Mo. miR-21-mediated tumor growth. *Oncogene*, 26(19):2799–2803, 2007.

37. Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.

38. Hiromu Suzuki, Shintaro Takatsuka, Hirofumi Akashi, Eiichiro Yamamoto, Masanori Nojima, Reo Maruyama, Masahiro Kai, Hiro-o Yamano, Yasushi Sasaki, Takashi Tokino, et al. Genome-wide profiling of chromatin signatures reveals epigenetic regulation of microrna genes in colorectal cancer. *Cancer research*, 71(17):5646–5658, 2011.

39. Kuo-Wang Tsai, Hsiao-Wei Kao, Hua-Chien Chen, Su-Jen Chen, and Wen-chang Lin. Epigenetic control of the expression of a primate-specific microrna cluster in human cancer cells. *Epigenetics*, 4(8):587–592, 2009.

40. Haleh Vosgha, Ali Salajegheh, Robert Anthony Smith, and Alfred King-Yin Lam. The important roles of mir-205 in normal physiology, cancers and as a potential therapeutic target. *Current cancer drug targets*, 14(7):621–637, 2014.

41. John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

42. Dongxiao Zhu, Youjuan Li, and Hua Li. Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data. *Bioinformatics*, 23(17):2298–2305, 2007.