An efficient not-only-linear correlation coefficient based on machine learning

A DOI-citable version of this manuscript is available at https://doi.org/10.1101/2022.06.15.496326.

Authors

- **Milton Pividori**
  0000-0002-3035-4403 · miltondp · miltondp
  Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA · Funded by The Gordon and Betty Moore Foundation GBMF 4552; The National Human Genome Research Institute (R01 HG010067); The National Human Genome Research Institute (K99HG011898)

- **Marylyn D. Ritchie**
  0000-0002-1208-1720 · MarylynRitchie
  Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

- **Diego H. Milone**
  0000-0003-2182-4351 · dmilone · d1001
  Research Institute for Signals, Systems and Computational Intelligence (sinc(i)), Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Santa Fe CP3000, Argentina

- **Casey S. Greene** ✉
  0000-0001-8713-9213 · cgreene · GreeneScientist
  Center for Health AI, University of Colorado School of Medicine, Aurora, CO 80045, USA; Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA · Funded by The Gordon and Betty Moore Foundation (GBMF 4552); The National Human Genome Research Institute (R01 HG010067); The National Cancer Institute (R01 CA237170)

✉ — Correspondence possible via GitHub Issues or email to Casey S. Greene <casey.s.greene@cuanschutz.edu>.

Abstract

Correlation coefficients are widely used to identify patterns in data that may be of particular interest. In transcriptomics, genes with correlated expression often share functions or are part of disease-relevant biological processes. Here we introduce the Clustermatch Correlation Coefficient (CCC), an efficient, easy-to-use and not-only-linear coefficient based on machine learning models. CCC reveals biologically meaningful linear and nonlinear patterns missed by standard, linear-only correlation coefficients. CCC captures general patterns in data by comparing clustering solutions while being much faster than state-of-the-art coefficients such as the Maximal Information Coefficient. When applied to human gene expression data, CCC identifies robust linear relationships while detecting nonlinear patterns associated, for example, with sex differences that are not captured by linear-only coefficients. Gene pairs highly ranked by CCC were enriched for interactions in integrated networks built from protein-protein interaction, transcription factor regulation, and chemical and genetic perturbations, suggesting that CCC could detect functional relationships that linear-only methods missed. CCC is a highly-efficient, next-generation not-only-linear correlation coefficient that can readily be applied to genome-scale data and other domains across different data types.

Introduction

New technologies have vastly improved data collection, generating a deluge of information across different disciplines. This large amount of data provides new opportunities to address unanswered scientific questions, provided we have efficient tools capable of identifying multiple types of underlying patterns. Correlation analysis is an essential statistical technique for discovering relationships between variables [1]. Correlation coefficients are often used in exploratory data mining techniques, such as clustering or community detection algorithms, to compute a similarity value between a pair of objects of interest such as genes [2] or disease-relevant lifestyle factors [3]. Correlation methods are also used in supervised tasks, for example, for feature selection to improve prediction accuracy [4,5]. The Pearson correlation coefficient is ubiquitously deployed across application domains and diverse scientific areas. Thus, even minor and significant improvements in these techniques could have enormous consequences in industry and research.

In transcriptomics, many analyses start with estimating the correlation between genes. More sophisticated approaches built on correlation analysis can suggest gene function [6], aid in discovering common and cell lineage-specific regulatory networks [7], and capture important interactions in a living organism that can uncover molecular mechanisms in other species [8,9]. The analysis of large RNA-seq datasets [10,11] can also reveal complex transcriptional mechanisms underlying human diseases [2,12,13,14,15]. Since the introduction of the omnigenic model of complex traits [16,17], gene-gene relationships are playing an increasingly important role in genetic studies of human diseases [18,19,20,21], even in specific fields such as polygenic risk scores [22]. In this context, recent approaches combine disease-associated genes from genome-wide association studies (GWAS) with gene co-expression networks to prioritize “core” genes directly affecting diseases [19,20,23]. These core genes are not captured by standard statistical methods but are believed to be part of highly-interconnected, disease-relevant regulatory networks. Therefore, advanced correlation coefficients could immediately find wide applications across many areas of biology, including the prioritization of candidate drug targets in the precision medicine field.

The Pearson and Spearman correlation coefficients are widely used because they reveal intuitive relationships and can be computed quickly. However, they are designed to capture linear or monotonic patterns (referred to as linear-only) and may miss complex yet critical relationships. Novel coefficients have been proposed as metrics that capture nonlinear patterns such as the Maximal Information Coefficient (MIC) [24] and the Distance Correlation (DC) [25]. MIC, in particular, is one of

the most commonly used statistics to capture more complex relationships, with successful applications across several domains [4,26,27]. However, the computational complexity makes them impractical for even moderately sized datasets [26,28]. Recent implementations of MIC, for example, take several seconds to compute on a single variable pair across a few thousand objects or conditions [26]. We previously developed a clustering method for highly diverse datasets that significantly outperformed approaches based on Pearson, Spearman, DC and MIC in detecting clusters of simulated linear and nonlinear relationships with varying noise levels [29]. Here we introduce the Clustermatch Correlation Coefficient (CCC), an efficient not-only-linear coefficient that works across quantitative and qualitative variables. CCC has a single parameter that limits the maximum complexity of relationships found (from linear to more general patterns) and computation time. CCC provides a high level of flexibility to detect specific types of patterns that are more important for the user, while providing safe defaults to capture general relationships. We also provide an efficient CCC implementation that is highly parallelizable, allowing to speed up computation across variable pairs with millions of objects or conditions. To assess its performance, we applied our method to gene expression data from the Genotype-Tissue Expression v8 (GTEx) project across different tissues [30]. CCC captured both strong linear relationships and novel nonlinear patterns, which were entirely missed by standard coefficients. For example, some of these nonlinear patterns were associated with sex differences in gene expression, suggesting that CCC can capture strong relationships present only in a subset of samples. We also found that the CCC behaves similarly to MIC in several cases, although it is much faster to compute. Gene pairs detected in expression data by CCC had higher interaction probabilities in tissue-specific gene networks from the Genome-wide Analysis of gene Networks in Tissues (GIANT) [31]. Furthermore, its ability to efficiently handle diverse data types (including numerical and categorical features) reduces preprocessing steps and makes it appealing to analyze large and heterogeneous repositories.

Results

A robust and efficient not-only-linear dependence coefficient

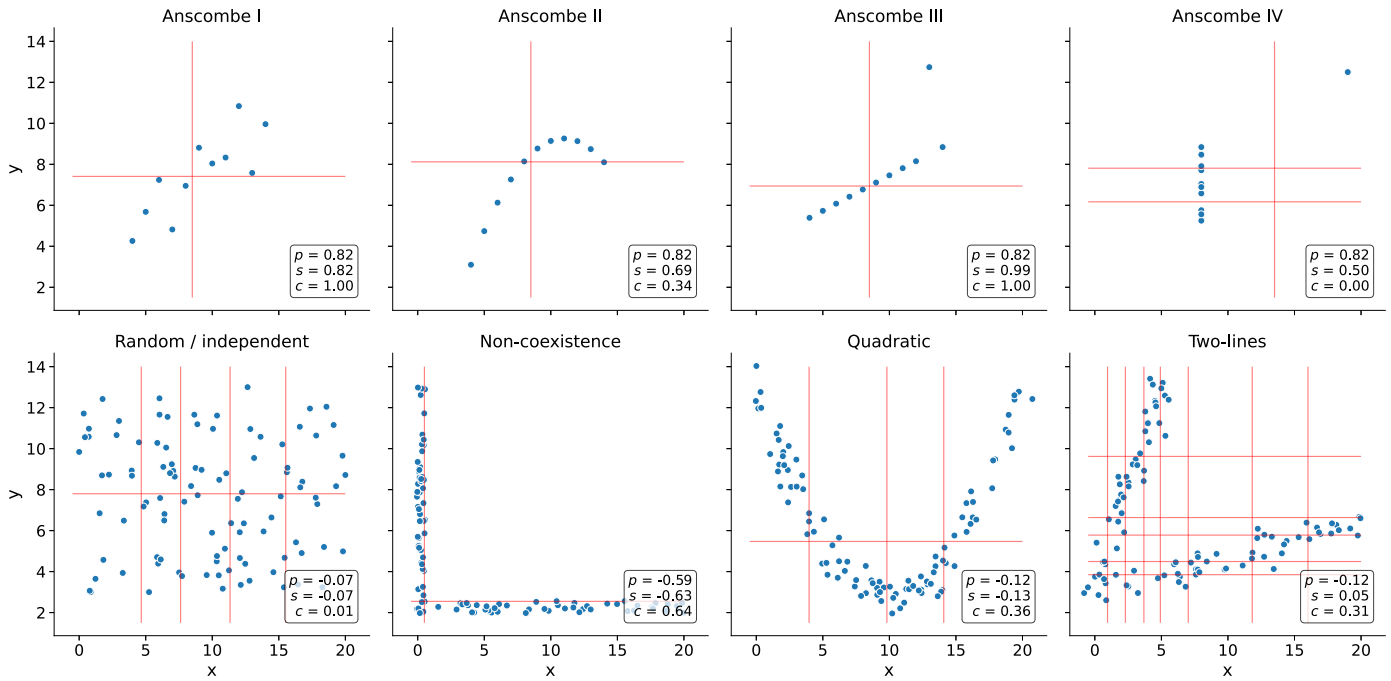

Figure 1: Different types of relationships in data. Each panel contains a set of simulated data points described by two generic variables: x and y. The first row shows Anscombe's quartet with four different datasets (from Anscombe I to IV) and 11 data points each. The second row contains a set of general patterns with 100 data points each. Each panel shows the correlation value using Pearson (p), Spearman (s) and CCC (c). Vertical and horizontal red lines show how CCC clustered data points using x and y.

The CCC provides a similarity measure between any pair of variables, either with numerical or categorical values. The method assumes that if there is a relationship between two variables/features describing n data points/objects, then the **cluster**ings of those objects using each variable should **match**. In the case of numerical values, CCC uses quantiles to efficiently separate data points into different clusters (e.g., the median separates numerical data into two clusters). Once all clusterings are generated according to each variable, we define the CCC as the maximum adjusted Rand index (ARI) [32] between them, ranging between 0 and 1. Details of the CCC algorithm can be found in Methods.

We examined how the Pearson (p), Spearman (s) and CCC (c) correlation coefficients behaved on different simulated data patterns. In the first row of Figure 1, we examine the classic Anscombe's quartet [33], which comprises four synthetic datasets with different patterns but the same data statistics (mean, standard deviation and Pearson's correlation). This kind of simulated data, recently revisited with the "Datasaurus" [34,35,36], is used as a reminder of the importance of going beyond simple statistics, where either undesirable patterns (such as outliers) or desirable ones (such as biologically meaningful nonlinear relationships) can be masked by summary statistics alone.

Anscombe I contains a noisy but clear linear pattern, similar to Anscombe III where the linearity is perfect besides one outlier. In these two examples, CCC separates data points using two clusters (one red line for each variable x and y), yielding 1.0 and thus indicating a strong relationship. Anscombe II seems to follow a partially quadratic relationship interpreted as linear by Pearson and Spearman. In contrast, for this potentially undersampled quadratic pattern, CCC yields a lower yet non-zero value of 0.34, reflecting a more complex relationship than a linear pattern. Anscombe IV shows a vertical line of data points where x values are almost constant except for one outlier. This outlier does not influence CCC as it does for Pearson or Spearman. Thus $c = 0.00$ (the minimum value) correctly indicates no association for this variable pair because, besides the outlier, for a single value of x there are ten different values for y. This pair of variables does not fit the CCC assumption: the two clusters formed with x (approximately separated by $x = 13$) do not match the three clusters formed with y. The Pearson's correlation coefficient is the same across all these Anscombe's examples ($p = 0.82$), whereas Spearman is 0.50 or greater. These simulated datasets show that both Pearson and Spearman are powerful in detecting linear patterns. However, any deviation in this assumption (like nonlinear relationships or outliers) affects their robustness.

We simulated additional types of relationships (Figure 1, second row), including some previously described from gene expression data [37,38,39]. For the random/independent pair of variables, all coefficients correctly agree with a value close to zero. The non-coexistence pattern, captured by all coefficients, represents a case where one gene (x) might be expressed while the other one (y) is inhibited, highlighting a potentially strong biological relationship (such as a microRNA negatively regulating another gene). For the other two examples (quadratic and two-lines), Pearson and Spearman do not capture the nonlinear pattern between variables x and y. These patterns also show how CCC uses different degrees of complexity to capture the relationships. For the quadratic pattern, for example, CCC separates x into more clusters (four in this case) to reach the maximum ARI. The two-lines example shows two embedded linear relationships with different slopes, which neither Pearson nor Spearman detect ($p = -0.12$ and $s = 0.05$, respectively). Here, CCC increases the complexity of the model by using eight clusters for x and six for y, resulting in $c = 0.31$.

The CCC reveals linear and nonlinear patterns in human transcriptomic data

We next examined the characteristics of these correlation coefficients in gene expression data from GTEx v8 across different tissues. We selected the top 5,000 genes with the largest variance for our

initial analyses on whole blood and then computed the correlation matrix between genes using Pearson, Spearman and CCC (see Methods).

We examined the distribution of each coefficient's absolute values in GTEx (Figure 2). CCC (mean=0.14, median=0.08, sd=0.15) has a much more skewed distribution than Pearson (mean=0.31, median=0.24, sd=0.24) and Spearman (mean=0.39, median=0.37, sd=0.26). The coefficients reach a cumulative set containing 70% of gene pairs at different values (Figure 2 b), $c = 0.18$, $p = 0.44$ and $s = 0.56$, suggesting that for this type of data, the coefficients are not directly comparable by magnitude, so we used ranks for further comparisons. In GTEx v8, CCC values were closer to Spearman and vice versa than either was to Pearson (Figure 2 c). We also compared the Maximal Information Coefficient (MIC) in this data (see Supplementary Note 1). We found that CCC behaved very similarly to MIC, although CCC was up to two orders of magnitude faster to run (see Supplementary Note 2). MIC, an advanced correlation coefficient able to capture general patterns beyond linear relationships, represented a significant step forward in correlation analysis research and has been successfully used in various application domains [4,26,27]. These results suggest that our findings for CCC generalize to MIC, therefore, in the subsequent analyses we focus on CCC and linear-only coefficients.

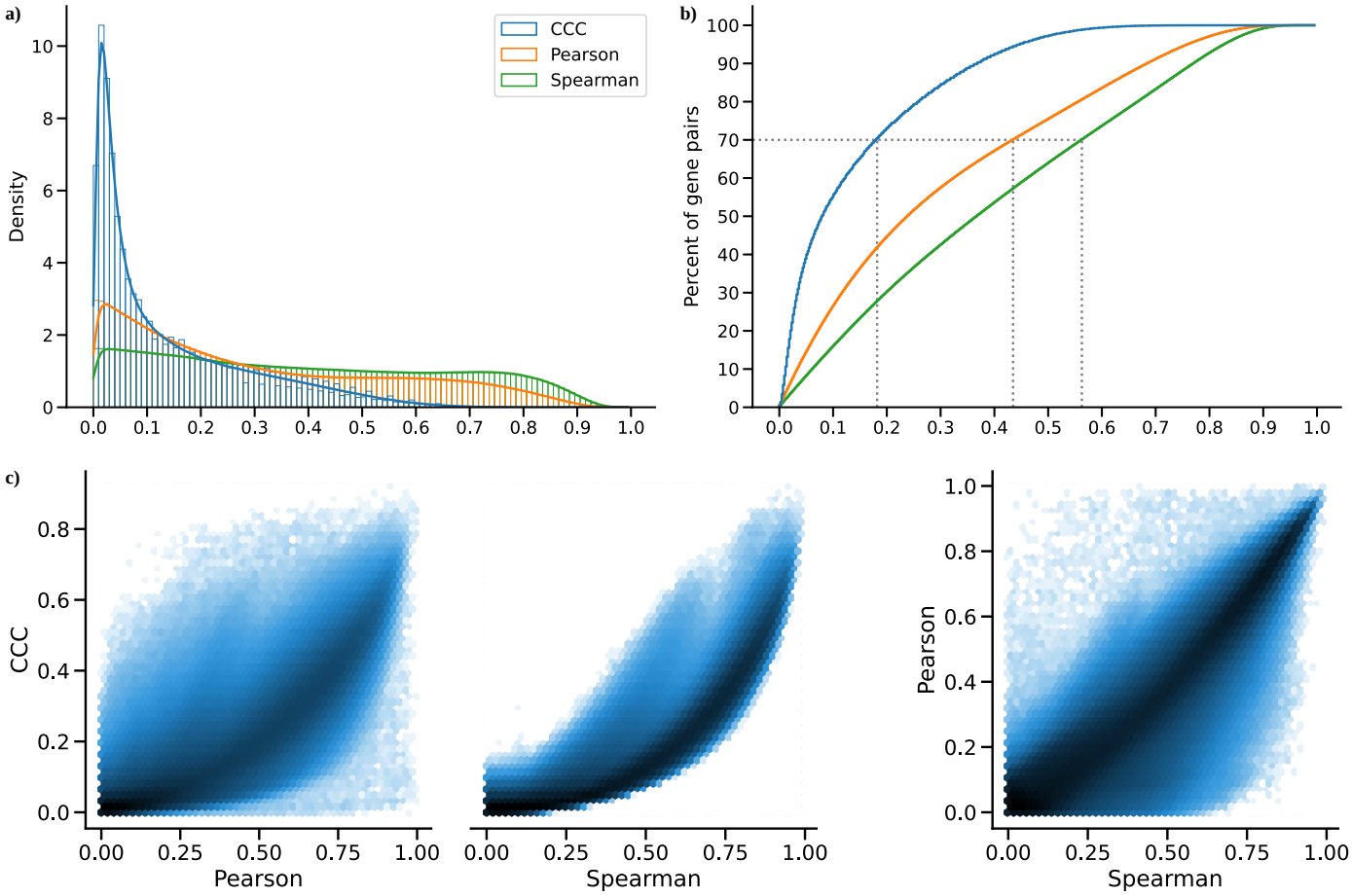

Figure 2: Distribution of coefficient values on gene expression (GTEx v8, whole blood). a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all coefficients, where a logarithmic scale was used to color each hexagon.

A closer inspection of gene pairs that were either prioritized or disregarded by these coefficients revealed that they captured different patterns. We analyzed the agreements and disagreements by obtaining, for each coefficient, the top 30% of gene pairs with the largest correlation values ("high" set) and the bottom 30% ("low" set), resulting in six potentially overlapping categories. For most cases (76.4%), an UpSet analysis [40] (Figure 3 a) showed that the three coefficients agreed on whether there is a strong correlation (42.1%) or there is no relationship (34.3%). Since Pearson and Spearman are linear-only, and CCC can also capture these patterns, we expect that these concordant gene pairs

represent clear linear patterns. CCC and Spearman agree more on either highly or poorly correlated pairs (4.0% in “high”, and 7.0% in “low”) than any of these with Pearson (all between 0.3%-3.5% for “high”, and 2.8%-5.5% for “low”). In summary, CCC agrees with either Pearson or Spearman in 90.5% of gene pairs by assigning a high or a low correlation value.

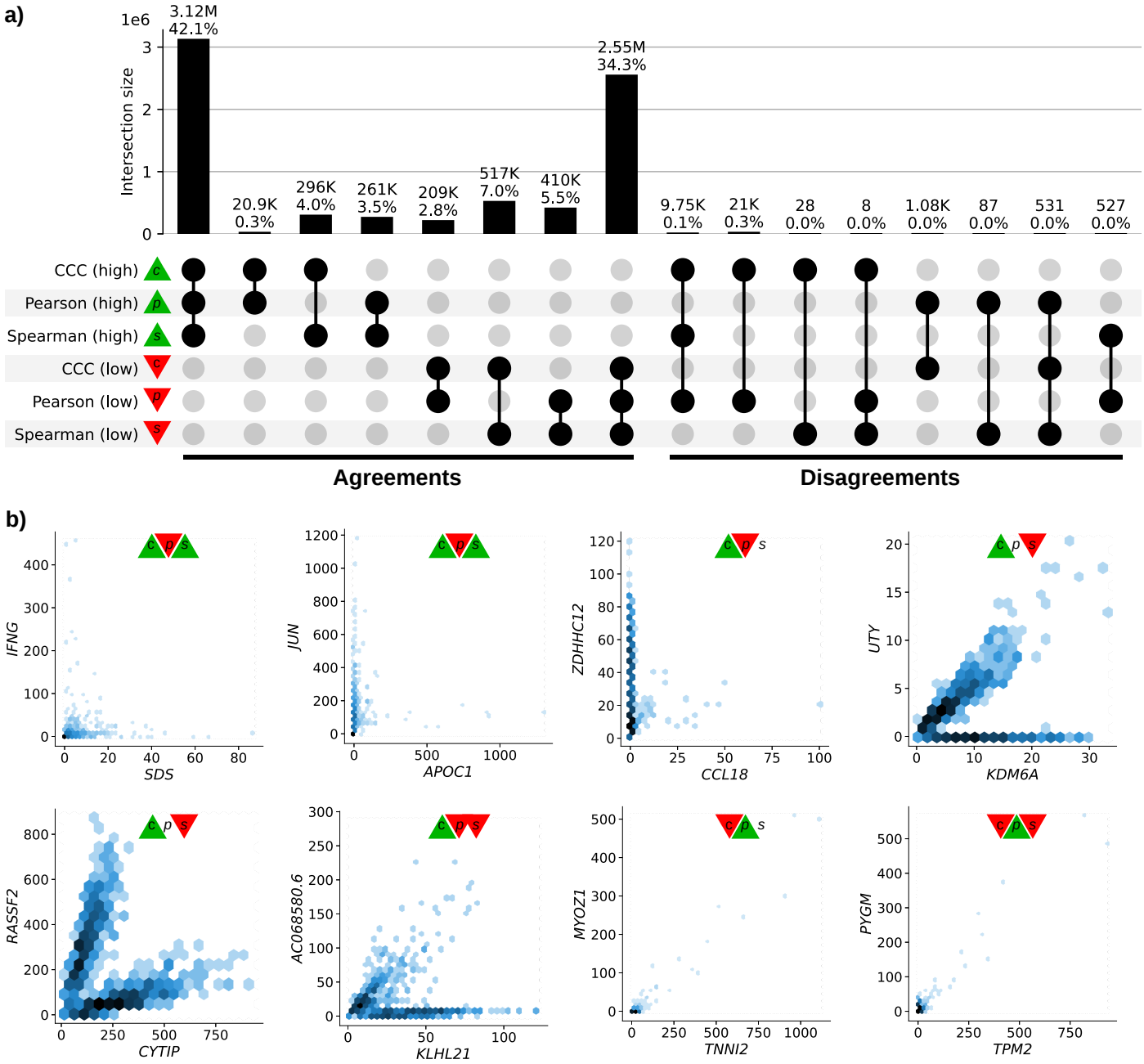

Figure 3: Intersection of gene pairs with high and low correlation coefficient values (GTEx v8, whole blood). a) UpSet plot with six categories (rows) grouping the 30% of the highest (green triangle) and lowest (red triangle) values for each coefficient. Columns show different intersections of categories grouped by agreements and disagreements. **b)** Hexagonal binning plots with examples of gene pairs where CCC (c) disagrees with Pearson (p) and Spearman (s). For each method, colors in the triangles indicate if the gene pair is among the top (green) or bottom (red) 30% of coefficient values. No triangle means that the correlation value for the gene pair is between the 30th and 70th percentiles (neither low nor high). A logarithmic scale was used to color each hexagon.

While there was broad agreement, more than 20,000 gene pairs with a high CCC value were not highly ranked by the other coefficients (right part of Figure 3 a). There were also gene pairs with a high Pearson value and either low CCC (1,075), low Spearman (87) or both low CCC and low Spearman values (531). However, our examination suggests that many of these cases appear to be driven by potential outliers (Figure 3 b, and analyzed later). We analyzed gene pairs among the top five of each intersection in the “Disagreements” group (Figure 3 a, right) where CCC disagrees with Pearson, Spearman or both.

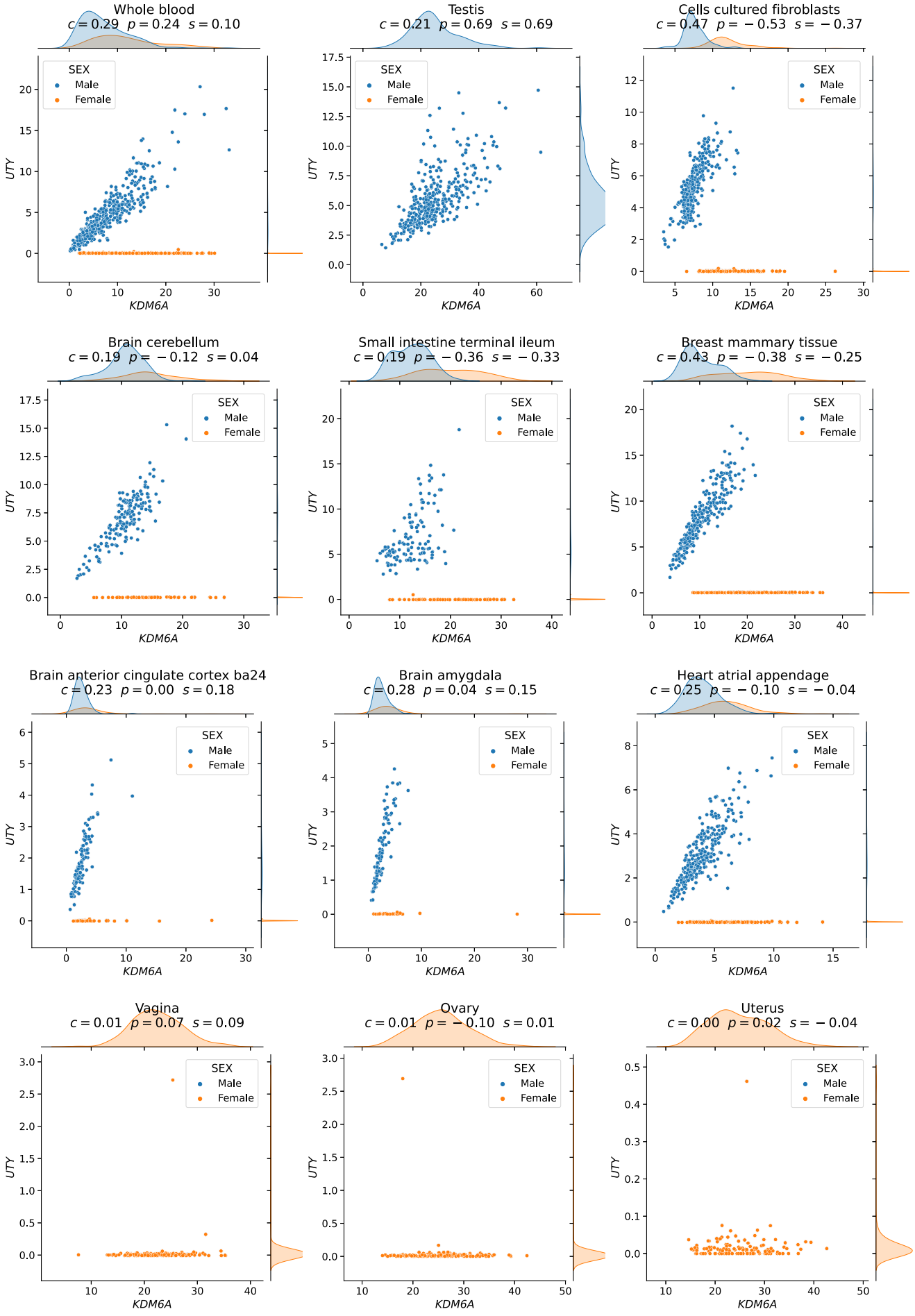

Figure 4: The expression levels of *KDM6A* and *UTY* display sex-specific associations across GTEx tissues. CCC captures this nonlinear relationship in all GTEx tissues (nine examples are shown in the first three rows), except in female-specific organs (last row).

The first three gene pairs at the top (*IFNG* - *SDS*, *JUN* - *APOC1*, and *ZDHHC12* - *CCL18*), with high CCC and low Pearson values, appear to follow a non-coexistence relationship: in samples where one of the genes is highly (slightly) expressed, the other is slightly (highly) activated, suggesting a potentially inhibiting effect. The following three gene pairs (*UTY* - *KDM6A*, *RASSF2* - *CYTIP*, and *AC068580.6* - *KLHL21*) follow patterns combining either two linear or one linear and one independent relationships. In particular, genes *UTY* and *KDM6A* (paralogs) show a nonlinear relationship where a subset of samples follows a robust linear pattern and another subset has a constant (independent) expression of one gene. This relationship is explained by the fact that *UTY* is in chromosome Y (Yq11) whereas *KDM6A* is in chromosome X (Xp11), and samples with a linear pattern are males, whereas those with no expression for *UTY* are females. This combination of linear and independent patterns is captured by CCC ($c = 0.29$, above the 80th percentile) but not by Pearson ($p = 0.24$, below the 55th percentile) or Spearman ($s = 0.10$, below the 15th percentile). Furthermore, the same gene pair pattern is highly ranked by CCC in all other tissues in GTEx, except for female-specific organs (Figure 4).

Replication of gene associations using tissue-specific gene networks from GIANT

We sought to systematically analyze discrepant scores to assess whether associations were replicated in other datasets besides GTEx. This is challenging and prone to bias because linear-only correlation coefficients are usually used in gene co-expression analyses. We used 144 tissue-specific gene networks from the Genome-wide Analysis of gene Networks in Tissues (GIANT) [41,42], where nodes represent genes and each edge a functional relationship weighted with a probability of interaction between two genes (see Methods). Importantly, the version of GIANT used in this study did not include GTEx samples [43], making it an ideal case for replication. These networks were built from expression and different interaction measurements, including protein-interaction, transcription factor regulation, chemical/genetic perturbations and microRNA target profiles from the Molecular Signatures Database (MSigDB [44]). We reasoned that highly-ranked gene pairs using three different coefficients in a single tissue (whole blood in GTEx, Figure 3) that represented real patterns should often replicate in a corresponding tissue or related cell lineage using the multi-cell type functional interaction networks in GIANT. In addition to predicting a network with interactions for a pair of genes, the GIANT web application can also automatically detect a relevant tissue or cell type where genes are predicted to be specifically expressed (the approach uses a machine learning method introduced in [45] and described in Methods). For example, we obtained the networks in blood and the automatically-predicted cell type for gene pairs *RASSF2* - *CYTIP* (CCC high, Figure 5 a) and *MYOZ1* - *TNNI2* (Pearson high, Figure 5 b). In addition to the gene pair, the networks include other genes connected according to their probability of interaction (up to 15 additional genes are shown), which allows estimating whether genes are part of the same tissue-specific biological process. Two large black nodes in each network's top-left and bottom-right corners represent our gene pairs. A green edge means a close-to-zero probability of interaction, whereas a red edge represents a strong predicted relationship between the two genes. In this example, genes *RASSF2* and *CYTIP* (Figure 5 a), with a high CCC value ($c = 0.20$, above the 73th percentile) and low Pearson and Spearman ($p = 0.16$ and $s = 0.11$, below the 38th and 17th percentiles, respectively), were both strongly connected to the blood network, with interaction scores of at least 0.63 and an average of 0.75 and 0.84, respectively (Supplementary Table 1). The autodetected cell type for this pair was leukocytes, and interaction scores were similar to the blood network (Supplementary Table 1). However, genes *MYOZ1* and *TNNI2*, with a very high Pearson value ($p = 0.97$), moderate Spearman ($s = 0.28$) and very low CCC ($c = 0.03$), were predicted to belong to much less cohesive networks (Figure 5 b), with average interaction scores of 0.17 and 0.22 with the rest of the genes, respectively. Additionally, the autodetected cell type (skeletal muscle) is not related to blood or one of its cell lineages. These preliminary results suggested that CCC might be capturing blood-specific patterns missed by the other coefficients.

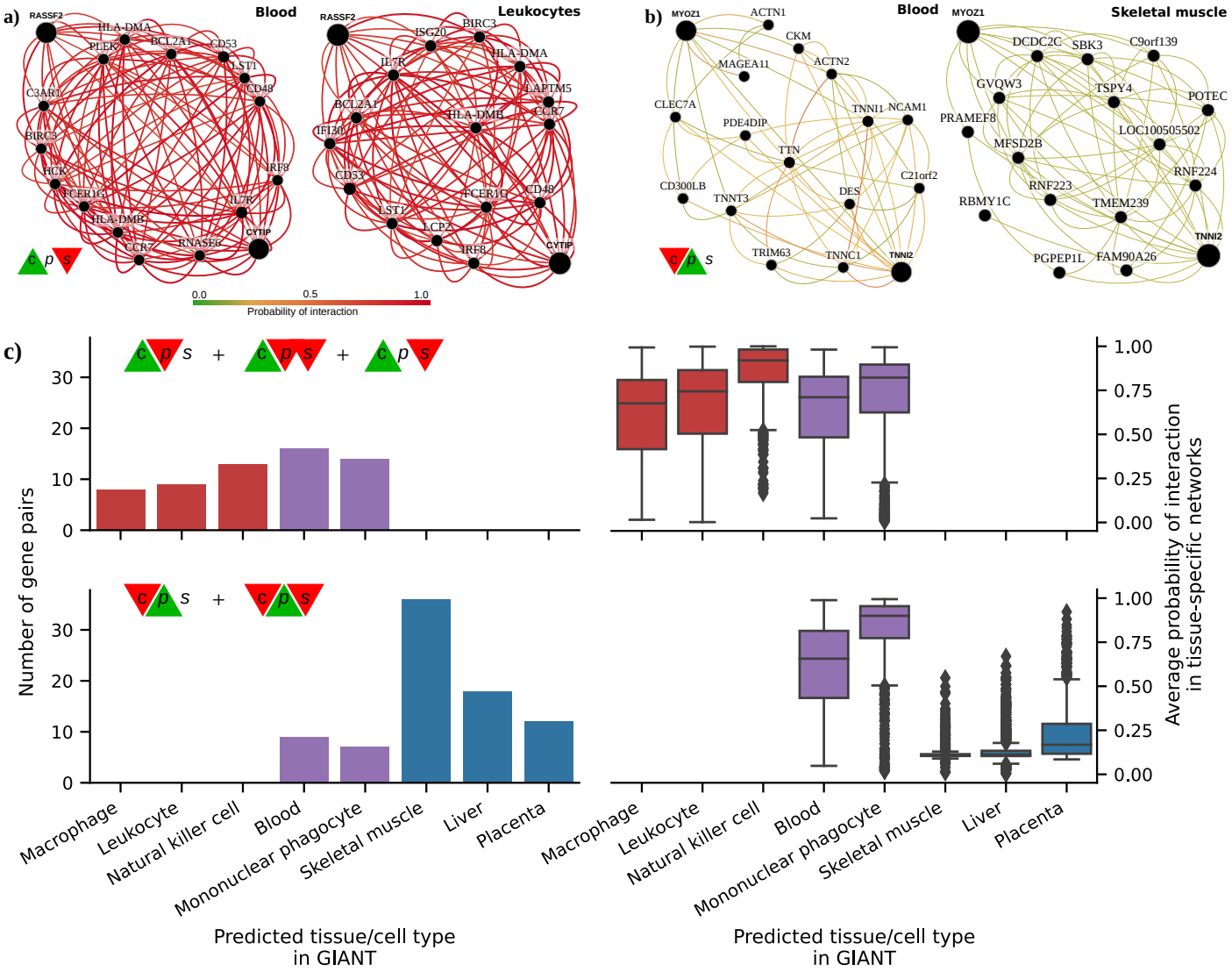

Figure 5: Analysis of GIANT tissue-specific predicted networks for gene pairs prioritized by correlation coefficients. a-b) Two gene pairs prioritized by correlation coefficients (from Figure 3 b) with their predicted networks in blood (left) and an automatically selected tissue/cell type (right) using the method described in [45]. A node represents a gene and an edge the probability that two genes are part of the same biological process in a specific cell type. A maximum of 15 genes are shown for each network. The GIANT web application automatically determined a minimum interaction confidence (edges' weights) to be shown. These networks can be analyzed online using the following links: *RASSF2 - CYTIP* [46], *MYOZ1 - TNNI2* [47]. **c)** Summary of predicted tissue/cell type networks for gene pairs exclusively prioritized by CCC and Pearson. The first row combines all gene pairs where CCC is high and Pearson or Spearman are low. The second row combines all gene pairs where Pearson is high and CCC or Spearman are low. Bar plots (left) show the number of gene pairs for each predicted tissue/cell type. Box plots (right) show the average probability of interaction between genes in these predicted tissue-specific networks. Red indicates CCC-only tissues/cell types, blue are Pearson-only, and purple are shared.

We next performed a systematic evaluation using the top 100 discrepant gene pairs between CCC and the other two coefficients. For each gene pair prioritized in GTEx (whole blood), we autodetected a relevant cell type using GIANT to assess whether genes were predicted to be specifically expressed in a blood-relevant cell lineage. For this, we used the top five most commonly autodetected cell types for each coefficient and assessed connectivity in the resulting networks (see Methods). The top 5 predicted cell types for gene pairs highly ranked by CCC and not by the rest were all blood-specific (Figure 5 c, top left), including macrophage, leukocyte, natural killer cell, blood and mononuclear phagocyte. The average probability of interaction between genes in these CCC-ranked networks was significantly higher than the other coefficients (Figure 5 c, top right), with all medians larger than 67% and first quartiles above 41% across predicted cell types. In contrast, most Pearson's gene pairs were predicted to be specific to tissues unrelated to blood (Figure 5 c, bottom left), with skeletal muscle being the most commonly predicted tissue. The interaction probabilities in these Pearson-ranked

networks were also generally lower than in CCC, except for blood-specific gene pairs (Figure 5 c, bottom right). The associations exclusively detected by CCC in whole blood from GTEx were more strongly replicated in these independent networks that incorporated multiple data modalities. CCC-ranked gene pairs not only had high probabilities of belonging to the same biological process but were also predicted to be specifically expressed in blood cell lineages. Conversely, most Pearson-ranked gene pairs were not predicted to be blood-specific, and their interaction probabilities were relatively low. This lack of replication in GIANT suggests that top Pearson-ranked gene pairs in GTEx might be driven mainly by outliers, which is consistent with our earlier observations of outlier-driven associations (Figure 3 b).

Discussion

We introduce the Clustermatch Correlation Coefficient (CCC), an efficient not-only-linear machine learning-based statistic. Applying CCC to GTEx v8 revealed that it was robust to outliers and detected linear relationships as well as complex and biologically meaningful patterns that standard coefficients missed. In particular, CCC alone detected gene pairs with complex nonlinear patterns from the sex chromosomes, highlighting the way that not-only-linear coefficients can play in capturing sex-specific differences. The ability to capture these nonlinear patterns, however, extends beyond sex differences: it provides a powerful approach to detect complex relationships where a subset of samples or conditions are explained by other factors (such as differences between health and disease). We found that top CCC-ranked gene pairs in whole blood from GTEx were replicated in independent tissue-specific networks trained from multiple data types and attributed to cell lineages from blood, even though CCC did not have access to any cell lineage-specific information. This suggests that CCC can disentangle intricate cell lineage-specific transcriptional patterns missed by linear-only coefficients. In addition to capturing nonlinear patterns, the CCC was more similar to Spearman than Pearson, highlighting their shared robustness to outliers. The CCC results were concordant with MIC, but much faster to compute and thus practical for large datasets. Another advantage over MIC is that CCC can also process categorical variables together with numerical values. CCC is conceptually easy to interpret and has a single parameter that controls the maximum complexity of the detected relationships while also balancing compute time.

Datasets such as Anscombe or “Datasaurus” highlight the value of visualization instead of relying on simple data summaries. While visual analysis is helpful, for many datasets examining each possible relationship is infeasible, and this is where more sophisticated and robust correlation coefficients are necessary. Advanced yet interpretable coefficients like CCC can focus human interpretation on patterns that are more likely to reflect real biology. The complexity of these patterns might reflect heterogeneity in samples that mask clear relationships between variables. For example, genes *UTY*-*KDM6A* (from sex chromosomes), detected by CCC, have a strong linear relationship but only in a subset of samples (males), which was not captured by linear-only coefficients. This example, in particular, highlights the importance of considering sex as a biological variable (SABV) [48] to avoid overlooking important differences between men and women, for instance, in disease manifestations [49,50]. More generally, a not-only-linear correlation coefficient like CCC could identify significant differences between variables (such as genes) that are explained by a third factor (beyond sex differences), that would be entirely missed by linear-only coefficients.

It is well-known that biomedical research is biased towards a small fraction of human genes [51,52]. Some genes highlighted in CCC-ranked pairs (Figure 3 b), such as *SDS* (12q24) and *ZDHHC12* (9q34), were previously found to be the focus of fewer than expected publications [53]. It is possible that the widespread use of linear coefficients may bias researchers away from genes with complex coexpression patterns. A beyond-linear gene co-expression analysis on large compendia might shed light on the function of understudied genes. For example, gene *KLHL21* (1p36) and *AC068580.6* (*ENSG00000235027*, in 11p15) have a high CCC value and are missed by the other coefficients. *KLHL21*

was suggested as a potential therapeutic target for hepatocellular carcinoma [54] and other cancers [55,56]. Its nonlinear correlation with *AC068580.6* might unveil other important players in cancer initiation or progression, potentially in subsets of samples with specific characteristics (as suggested in Figure 3 b).

Not-only-linear correlation coefficients might also be helpful in the field of genetic studies. In this context, genome-wide association studies (GWAS) have been successful in understanding the molecular basis of common diseases by estimating the association between genotype and phenotype [57]. However, the estimated effect sizes of genes identified with GWAS are generally modest, and they explain only a fraction of the phenotype variance, hampering the clinical translation of these findings [58]. Recent theories, like the omnigenic model for complex traits [16,17], argue that these observations are explained by highly-interconnected gene regulatory networks, with some core genes having a more direct effect on the phenotype than others. Using this omnigenic perspective, we and others [19,20,23] have shown that integrating gene co-expression networks in genetic studies could potentially identify core genes that are missed by linear-only models alone like GWAS. Our results suggest that building these networks with more advanced and efficient correlation coefficients could better estimate gene co-expression profiles and thus more accurately identify these core genes. Approaches like CCC could play a significant role in the precision medicine field by providing the computational tools to focus on more promising genes representing potentially better candidate drug targets.

Our analyses have some limitations. We worked on a sample with the top variable genes to keep computation time feasible. Although CCC is much faster than MIC, Pearson and Spearman are still the most computationally efficient since they only rely on simple data statistics. Our results, however, reveal the advantages of using more advanced coefficients like CCC for detecting and studying more intricate molecular mechanisms that replicated in independent datasets. The application of CCC on larger compendia, such as recount3 [11] with thousands of heterogeneous samples across different conditions, can reveal other potentially meaningful gene interactions. The single parameter of CCC, $k_{\max}$, controls the maximum complexity of patterns found and also impacts the compute time. Our analysis suggested that $k_{\max} = 10$ was sufficient to identify both linear and more complex patterns in gene expression. A more comprehensive analysis of optimal values for this parameter could provide insights to adjust it for different applications or data types.

While linear and rank-based correlation coefficients are exceptionally fast to calculate, not all relevant patterns in biological datasets are linear. For example, patterns associated with sex as a biological variable are not apparent to the linear-only coefficients that we evaluated but are revealed by not-only-linear methods. Beyond sex differences, being able to use a method that inherently identifies patterns driven by other factors is likely to be desirable. Not-only-linear coefficients can also disentangle intricate yet relevant patterns from expression data alone that were replicated in models integrating different data modalities. CCC, in particular, is highly parallelizable, and we anticipate efficient GPU-based implementations that could make it even faster. The CCC is an efficient, next-generation correlation coefficient that is highly effective in transcriptome analyses and potentially useful in a broad range of other domains.

Methods

The code needed to reproduce all of our analyses and generate the figures is available in https://github.com/greenelab/ccc. We provide scripts to download the required data and run all the steps. A Docker image is provided to use the same runtime environment.

The CCC algorithm

The Clustermatch Correlation Coefficient (CCC) computes a similarity value $c \in [0, 1]$ between any pair of numerical or categorical features/variables $\mathbf{x}$ and $\mathbf{y}$ measured on n objects. CCC assumes that if two features $\mathbf{x}$ and $\mathbf{y}$ are similar, then the partitioning by clustering of the n objects using each feature separately should match. For example, given $\mathbf{x} = (11, 27, 32, 40)$ and $\mathbf{y} = 10x = (110, 270, 320, 400)$, where $n = 4$, partitioning each variable into two clusters ($k = 2$) using their medians (29.5 for $\mathbf{x}$ and 295 for $\mathbf{y}$) would result in partition $\Omega^{\mathbf{x}}_{k=2} = (1, 1, 2, 2)$ for $\mathbf{x}$, and partition $\Omega^{\mathbf{y}}_{k=2} = (1, 1, 2, 2)$ for $\mathbf{y}$. Then, the agreement between $\Omega^{\mathbf{x}}_{k=2}$ and $\Omega^{\mathbf{y}}_{k=2}$ can be computed using any measure of similarity between partitions, like the adjusted Rand index (ARI) [32]. In that case, it will return the maximum value (1.0 in the case of ARI). Note that the same value of k might not be the right one to find a relationship between any two features. For instance, in the quadratic example in Figure 1, CCC returns a value of 0.36 (grouping objects in four clusters using one feature and two using the other). If we used only two clusters instead, CCC would return a similarity value of 0.02. Therefore, the CCC algorithm (shown below) searches for this optimal number of clusters given a maximum k, which is its single parameter $k_{\max}$.

Algorithm 1: CCC algorithm

```
1  Function get_partitions(v, k_max):
       Output:
           Ω_r: clustering with r clusters over n objects
2      if v ∈ ℝ^n then
3          for r ← 2 to min{k_max, |v| − 1} do
4              ρ ← (ρ_ℓ | Pr(v_i < ρ_ℓ) ≤ (ℓ − 1)/r), ∀ℓ ∈ [1, r + 1]
5              Ω_rℓ ← {i | ρ_ℓ < v_i ≤ ρ_{ℓ+1}}, ∀ℓ ∈ [1, r]
6      else
7          C ← ∪_j {v_i}
8          r ← |C|
9          Ω_rc ← {i | v_i = C_c}, ∀c ∈ [1, r]
10     Ω ← {Ω_r | |Ω_r| > 1}, ∀r
11     return Ω
12
13 Function ccc(x, y, k_max):
       Input:
           x: feature values on n objects
           y: feature values on n objects
           k_max: maximum number of internal clusters
       Output:
           c: similarity value for x and y (c ∈ [0, 1])
14     Ω^x = get_partitions(x, k_max)
15     Ω^y = get_partitions(y, k_max)
16     c ← max{A(Ω^x_p, Ω^y_q)}, ∀p, q
17     return max(c, 0)
```

The main function of the algorithm, `ccc`, generates a list of partitionings $\Omega^{\mathbf{x}}$ and $\Omega^{\mathbf{y}}$ (lines 14 and 15), for each feature $\mathbf{x}$ and $\mathbf{y}$. Then, it computes the ARI between each partition in $\Omega^{\mathbf{x}}$ and $\Omega^{\mathbf{y}}$ (line 16), and then it keeps the pair that generates the maximum ARI. Finally, since ARI does not have a lower bound (it could return negative values, which in our case are not meaningful), CCC returns only values between 0 and 1 (line 17).

Interestingly, since CCC only needs a pair of partitions to compute a similarity value, any type of feature that can be used to perform clustering/grouping is supported. If the feature is numerical (lines 2 to 5 in the `get_partitions` function), then quantiles are used for clustering (for example, the median generates $k = 2$ clusters of objects), from $k = 2$ to $k = k_{\max}$. If the feature is categorical (lines 7 to 9), the categories are used to group objects together. Consequently, since features are internally categorized into clusters, numerical and categorical variables can be naturally integrated since clusters do not need an order.

For all our analyses we used $k_{\max} = 10$. This means that for each gene pair, 18 partitions are generated (9 for each gene, from $k = 2$ to $k = 10$), and 81 ARI comparisons are performed. Smaller values of $k_{\max}$ can reduce computation time, although at the expense of missing more complex/general relationships. Our examples in Figure 1 suggest that using $k_{\max} = 2$ would force CCC to find linear-only patterns, which could be a valid use case scenario where only this kind of relationships are desired. In addition, $k_{\max} = 2$ implies that only two partitions are generated, and only one ARI comparison is performed. In this regard, our Python implementation of CCC provides flexibility in specifying $k_{\max}$. For instance, instead of the maximum k (an integer), the parameter could be a custom list of integers: for example, `[2, 5, 10]` will partition the data into two, five and ten clusters.

For a single pair of features (genes in our study), generating partitions or computing their similarity can be parallelized. We used three CPU cores in our analyses to speed up the computation of CCC. A future improved implementation of CCC could potentially use graphical processing units (GPU) to parallelize its computation further.

A Python implementation of CCC (optimized with `numba` [59]) can be found in our Github repository [60], as well as a package published in the Python Package Index (PyPI) that can be easily installed.

Gene expression data and preprocessing

We downloaded GTEx v8 data for all tissues, normalized using TPM (transcripts per million), and focused our primary analysis on whole blood, which has a good sample size (755). We selected the top 5,000 genes from whole blood with the largest variance after standardizing with $log(x + 1)$ to avoid a bias towards highly-expressed genes. We then computed Pearson, Spearman, MIC and CCC on these 5,000 genes across all 755 samples on the TPM-normalized data, generating a pairwise similarity matrix of size 5,000 x 5,000.

Tissue-specific network analyses using GIANT

We accessed tissue-specific gene networks of GIANT using both the web interface and web services provided by HumanBase [42]. The GIANT version used in this study included 987 genome-scale datasets with approximately 38,000 conditions from around 14,000 publications. Details on how these networks were built are described in [31]. Briefly, tissue-specific gene networks were built using gene expression data (without GTEx samples [43]) from the NCBI's Gene Expression Omnibus (GEO) [61], protein-protein interaction (BioGRID [62], IntAct [63], MINT [64] and MIPS [65]), transcription factor regulation using binding motifs from JASPAR [66], and chemical and genetic perturbations from MSigDB [67]. Gene expression data were log-transformed, and the Pearson correlation was computed for each gene pair, normalized using the Fisher's z transform, and z-scores discretized into different bins. Gold standards for tissue-specific functional relationships were built using expert curation and experimentally derived gene annotations from the Gene Ontology. Then, one naive Bayesian classifier (using C++ implementations from the Sleipnir library [68]) for each of the 144 tissues was trained using these gold standards. Finally, these classifiers were used to estimate the probability of tissue-specific interactions for each gene pair.

For each pair of genes prioritized in our study using GTEx, we used GIANT through HumanBase to obtain 1) a predicted gene network for blood (manually selected to match whole blood in GTEx) and 2) a gene network with an automatically predicted tissue using the method described in [45] and provided by HumanBase web interfaces/services. Briefly, the tissue prediction approach trains a machine learning model using comprehensive transcriptional data with human-curated markers of different cell lineages (e.g., macrophages) as gold standards. Then, these models are used to predict other cell lineage-specific genes. In addition to reporting this predicted tissue or cell lineage, we computed the average probability of interaction between all genes in the network retrieved from GIANT. Following the default procedure used in GIANT, we included the top 15 genes with the highest probability of interaction with the queried gene pair for each network.

Maximal Information Coefficient (MIC)

We used the Python package `minepy` [69,70] (version 1.2.5) to estimate the MIC coefficient. In GTEx v8 (whole blood), we used MIC_e (an improved implementation of the original MIC introduced in [71]) with the default parameters `alpha=0.6`, `c=15` and `estimator='mic_e'`. We used the `pairwise_distances` function from `scikit-learn` [72] to parallelize the computation of MIC on GTEx. For our computational complexity analyses (see Supplementary Material), we ran the original MIC (using parameter `estimator='mic_approx'`) and MIC_e (`estimator='mic_e'`).

References

1. **Making data maximally available.**
Brooks Hanson, Andrew Sugden, Bruce Alberts
Science (New York, N.Y.) (2011-02-11) https://www.ncbi.nlm.nih.gov/pubmed/21310971
DOI: 10.1126/science.1203354 · PMID: 21310971

2. **Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder.**
Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra L Theesfeld, Aaron K Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, Olga G Troyanskaya
Nature neuroscience (2016-08-01) https://www.ncbi.nlm.nih.gov/pubmed/27479844
DOI: 10.1038/nn.4353 · PMID: 27479844 · PMCID: PMC5803797

3. **Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality**
Jing Kong, Barbara EK Klein, Ronald Klein, Kristine E Lee, Grace Wahba
Proceedings of the National Academy of Sciences (2012-11-21) https://doi.org/f4htm9
DOI: 10.1073/pnas.1217269109 · PMID: 23175793 · PMCID: PMC3528609

4. **McTwo: a two-step feature selection algorithm based on maximal information coefficient.**
Ruiquan Ge, Manli Zhou, Youxi Luo, Qinghan Meng, Guoqin Mai, Dongli Ma, Guoqing Wang, Fengfeng Zhou
BMC bioinformatics (2016-03-23) https://www.ncbi.nlm.nih.gov/pubmed/27006077
DOI: 10.1186/s12859-016-0990-0 · PMID: 27006077 · PMCID: PMC4804474

5. **A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data.**
Xian-Fang Song, Yong Zhang, Dun-Wei Gong, Xiao-Zhi Gao
IEEE transactions on cybernetics (2022-08-18) https://www.ncbi.nlm.nih.gov/pubmed/33729976
DOI: 10.1109/tcyb.2021.3061152 · PMID: 33729976

6. **Densely interconnected transcriptional circuits control cell states in human hematopoiesis.**
Noa Novershtern, Aravind Subramanian, Lee N Lawton, Raymond H Mak, WNicholas Haining, Marie E McConkey, Naomi Habib, Nir Yosef, Cindy Y Chang, Tal Shay, … Benjamin L Ebert
Cell (2011-01-21) https://www.ncbi.nlm.nih.gov/pubmed/21241896
DOI: 10.1016/j.cell.2011.01.004 · PMID: 21241896 · PMCID: PMC3049864

7. **Understanding multicellular function and disease with human tissue-specific networks.**
Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, … Olga G Troyanskaya
Nature genetics (2015-04-27) https://www.ncbi.nlm.nih.gov/pubmed/25915600
DOI: 10.1038/ng.3259 · PMID: 25915600 · PMCID: PMC4828725

8. **Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice.**
Stephen P Ficklin, FAlex Feltus
Plant physiology (2011-05-23) https://www.ncbi.nlm.nih.gov/pubmed/21606319
DOI: 10.1104/pp.111.173047 · PMID: 21606319 · PMCID: PMC3135956

9. **Global similarity and local divergence in human and mouse gene co-expression networks.**
Panayiotis Tsaparas, Leonardo Mariño-Ramírez, Olivier Bodenreider, Eugene V Koonin, IKing Jordan
BMC evolutionary biology (2006-09-12) https://www.ncbi.nlm.nih.gov/pubmed/16968540
DOI: 10.1186/1471-2148-6-70 · PMID: 16968540 · PMCID: PMC1601971

10. **The GTEx Consortium atlas of genetic regulatory effects across human tissues.** *Science (New York, N.Y.)* (2020-09-11) https://www.ncbi.nlm.nih.gov/pubmed/32913098
DOI: 10.1126/science.aaz1776 · PMID: 32913098 · PMCID: PMC7737656

11. **recount3: summaries and queries for large-scale RNA-seq expression and splicing.**
Christopher Wilks, Shijie C Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P Ling, Eddie Luidy Imada, David Zhang, Lance Joseph, Jeffrey T Leek, ... Ben Langmead
Genome biology (2021-11-29) https://www.ncbi.nlm.nih.gov/pubmed/34844637
DOI: 10.1186/s13059-021-02533-6 · PMID: 34844637 · PMCID: PMC8628444

12. **MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease.**
Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, Casey S Greene
Cell systems (2019-05-22) https://www.ncbi.nlm.nih.gov/pubmed/31121115
DOI: 10.1016/j.cels.2019.04.003 · PMID: 31121115 · PMCID: PMC6538307

13. **Integrating predicted transcriptome from multiple tissues improves association detection.**
Alvaro N Barbeira, Milton Pividori, Jiamao Zheng, Heather E Wheeler, Dan L Nicolae, Hae Kyung Im
PLoS genetics (2019-01-22) https://www.ncbi.nlm.nih.gov/pubmed/30668570
DOI: 10.1371/journal.pgen.1007889 · PMID: 30668570 · PMCID: PMC6358100

14. **Quantifying genetic effects on disease mediated by assayed gene expression levels.**
Douglas W Yao, Luke J O'Connor, Alkes L Price, Alexander Gusev
Nature genetics (2020-05-18) https://www.ncbi.nlm.nih.gov/pubmed/32424349
DOI: 10.1038/s41588-020-0625-2 · PMID: 32424349 · PMCID: PMC7276299

15. **Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression.**
Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, ... Lude Franke
Nature genetics (2021-09-02) https://www.ncbi.nlm.nih.gov/pubmed/34475573
DOI: 10.1038/s41588-021-00913-z · PMID: 34475573 · PMCID: PMC8432599

16. **An Expanded View of Complex Traits: From Polygenic to Omnigenic.**
Evan A Boyle, Yang I Li, Jonathan K Pritchard
Cell (2017-06-15) https://www.ncbi.nlm.nih.gov/pubmed/28622505
DOI: 10.1016/j.cell.2017.05.038 · PMID: 28622505 · PMCID: PMC5536862

17. **Trans Effects on Gene Expression Can Drive Omnigenic Inheritance.**
Xuanyao Liu, Yang I Li, Jonathan K Pritchard
Cell (2019-05-02) https://www.ncbi.nlm.nih.gov/pubmed/31051098
DOI: 10.1016/j.cell.2019.04.014 · PMID: 31051098 · PMCID: PMC6553491

18. **Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics.**

Karthik A Jagadeesh, Kushal K Dey, Daniel T Montoro, Rahul Mohan, Steven Gazal, Jesse M Engreitz, Ramnik J Xavier, Alkes L Price, Aviv Regev
bioRxiv : the preprint server for biology (2021-11-23)
https://www.ncbi.nlm.nih.gov/pubmed/34845454
DOI: 10.1101/2021.03.19.436212 · PMID: 34845454 · PMCID: PMC8629197

19. **Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms**
Milton Pividori, Sumei Lu, Binglan Li, Chun Su, Matthew E Johnson, Wei-Qi Wei, Qiping Feng, Bahram Namjou, Krzysztof Kiryluk, Iftikhar Kullo, … Casey S Greene
Cold Spring Harbor Laboratory (2021-07-06) https://doi.org/gk9g25
DOI: 10.1101/2021.07.05.450786

20. **Linking common and rare disease genetics through gene regulatory networks**
Olivier B Bakker, Annique Claringbould, Harm-Jan Westra, Henry Wiersma, Floranne Boulogne, Urmo Võsa, Sophie Mulcahy Symmons, Iris H Jonkers, Lude Franke, Patrick Deelen
Cold Spring Harbor Laboratory (2021-10-26) https://doi.org/gpdftn
DOI: 10.1101/2021.10.21.21265342

21. **Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression**
Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, …
Nature Genetics (2021-09) https://doi.org/gmpj66
DOI: 10.1038/s41588-021-00913-z · PMID: 34475573 · PMCID: PMC8432599

22. **The omnigenic model and polygenic prediction of complex traits**
Iain Mathieson
The American Journal of Human Genetics (2021-09) https://doi.org/gmv9s5
DOI: 10.1016/j.ajhg.2021.07.003 · PMID: 34331855 · PMCID: PMC8456163

23. **Identification of therapeutic targets from genetic association studies using hierarchical component analysis**
Hao-Chih Lee, Osamu Ichikawa, Benjamin S Glicksberg, Aparna A Divaraniya, Christine E Becker, Pankaj Agarwal, Joel T Dudley
BioData Mining (2020-06-17) https://doi.org/gjp5pf
DOI: 10.1186/s13040-020-00216-9 · PMID: 32565911 · PMCID: PMC7301559

24. **Detecting novel associations in large data sets.**
David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, Pardis C Sabeti
Science (New York, N.Y.) (2011-12-16) https://www.ncbi.nlm.nih.gov/pubmed/22174245
DOI: 10.1126/science.1205438 · PMID: 22174245 · PMCID: PMC3325791

25. **Measuring and testing dependence by correlation of distances**
Gábor J Székely, Maria L Rizzo, Nail K Bakirov
The Annals of Statistics (2007-12-01) https://doi.org/dkgjb4
DOI: 10.1214/009053607000000505

26. **An improved algorithm for the maximal information coefficient and its application.**
Dan Cao, Yuan Chen, Jin Chen, Hongyan Zhang, Zheming Yuan
Royal Society open science (2021-02-10) https://www.ncbi.nlm.nih.gov/pubmed/33972855
DOI: 10.1098/rsos.201424 · PMID: 33972855 · PMCID: PMC8074658

27. **Time-Frequency Maximal Information Coefficient Method and its Application to Functional Corticomuscular Coupling.**
Tie Liang, Qingyu Zhang, Xiaoguang Liu, Cunguang Lou, Xiuling Liu, Hongrui Wang
IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society (2020-11-06)
https://www.ncbi.nlm.nih.gov/pubmed/33001806
DOI: 10.1109/tnsre.2020.3028199 · PMID: 33001806

28. **A New Algorithm to Optimize Maximal Information Coefficient.**
Yuan Chen, Ying Zeng, Feng Luo, Zheming Yuan
PloS one (2016-06-22) https://www.ncbi.nlm.nih.gov/pubmed/27333001
DOI: 10.1371/journal.pone.0157567 · PMID: 27333001 · PMCID: PMC4917098

29. **Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization**
Milton Pividori, Andres Cernadas, Luis A de Haro, Fernando Carrari, Georgina Stegmayer, Diego H Milone
Bioinformatics (2018-10-24) https://doi.org/gfg4bt
DOI: 10.1093/bioinformatics/bty899 · PMID: 30357313

30. **The GTEx Consortium atlas of genetic regulatory effects across human tissues**
, François Aguet, Shankara Anand, Kristin G Ardlie, Stacey Gabriel, Gad A Getz, Aaron Graubert, Kane Hadley, Robert E Handsaker, Katherine H Huang, … Simona Volpi
Science (2020-09-11) https://doi.org/ghbnhr
DOI: 10.1126/science.aaz1776 · PMID: 32913098 · PMCID: PMC7737656

31. **Understanding multicellular function and disease with human tissue-specific networks**
Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, … Olga G Troyanskaya
Nature Genetics (2015-04-27) https://doi.org/f7dvkv
DOI: 10.1038/ng.3259 · PMID: 25915600 · PMCID: PMC4828725

32. **Comparing partitions**
Lawrence Hubert, Phipps Arabie
Journal of Classification (1985-12) https://doi.org/bphmzh
DOI: 10.1007/bf01908075

33. **Graphs in Statistical Analysis**
FJ Anscombe
The American Statistician (1973-02) https://doi.org/gfpn48
DOI: 10.1080/00031305.1973.10478966

34. **Download the Datasaurus: Never trust summary statistics alone; always visualize your data**
Alberto Cairo
http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

35. **Same Stats, Different Graphs**
Justin Matejka, George Fitzmaurice
Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (2017-05-02)
https://doi.org/gdtg2w
DOI: 10.1145/3025453.3025912

36. **Generating data sets for teaching the importance of regression analysis**

Lori L Murray, John G Wilson
Decision Sciences Journal of Innovative Education (2021-03-31) https://doi.org/gjmgqt
DOI: 10.1111/dsji.12233

37. **Detecting Novel Associations in Large Data Sets**
David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, Pardis C Sabeti
Science (2011-12-16) https://doi.org/bzn5c3
DOI: 10.1126/science.1205438 · PMID: 22174245 · PMCID: PMC3325791

38. **A Novel Method to Efficiently Highlight Nonlinearly Expressed Genes**
Qifei Wang, Haojian Zhang, Yuqing Liang, Heling Jiang, Siqiao Tan, Feng Luo, Zheming Yuan, Yuan Chen
Frontiers in Genetics (2020-01-31) https://doi.org/gnr5k7
DOI: 10.3389/fgene.2019.01410 · PMID: 32082366 · PMCID: PMC7006292

39. **Comprehensive Identification of Cell Cycle–regulated Genes of the YeastSaccharomyces cerevisiaeby Microarray Hybridization**
Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, Bruce Futcher
Molecular Biology of the Cell (1998-12) https://doi.org/gnr5k5
DOI: 10.1091/mbc.9.12.3273 · PMID: 9843569 · PMCID: PMC25624

40. **UpSet: Visualization of Intersecting Sets**
Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister
IEEE Transactions on Visualization and Computer Graphics (2014-12-31) https://doi.org/f3ssr5
DOI: 10.1109/tvcg.2014.2346248 · PMID: 26356912 · PMCID: PMC4720993

41. **Understanding multicellular function and disease with human tissue-specific networks**
Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, … Olga G Troyanskaya
Nature genetics (2015-06) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4828725/
DOI: 10.1038/ng.3259 · PMID: 25915600 · PMCID: PMC4828725

42. **HumanBase: data-driven predictions of gene function and interactions**
https://hb.flatironinstitute.org/

43. **Data sources** https://hb.flatironinstitute.org/data

44. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.**
Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, Jill P Mesirov
Proceedings of the National Academy of Sciences of the United States of America (2005-09-30) https://www.ncbi.nlm.nih.gov/pubmed/16199517
DOI: 10.1073/pnas.0506580102 · PMID: 16199517 · PMCID: PMC1239896

45. **Defining cell-type specificity at the transcriptional level in human disease**
Wenjun Ju, Casey S Greene, Felix Eichinger, Viji Nair, Jeffrey B Hodgin, Markus Bitzer, Young-suk Lee, Qian Zhu, Masami Kehata, Min Li, … Matthias Kretzler
Genome Research (2013-08-15) https://doi.org/f5g4hm
DOI: 10.1101/gr.155697.113 · PMID: 23950145 · PMCID: PMC3814886

46. **RASSF2, CYTIP - HumanBase** https://hb.flatironinstitute.org/gene/9770+9595

47. **MYOZ1, TNNI2 - HumanBase** https://hb.flatironinstitute.org/gene/58529+7136

48. **Policy: NIH to balance sex in cell and animal studies**
Janine A Clayton, Francis S Collins
Nature (2014-05) https://doi.org/gfzc82
DOI: 10.1038/509282a · PMID: 24834516 · PMCID: PMC5101948

49. **Considering Sex as a Biological Variable in Basic and Clinical Studies: An Endocrine Society Scientific Statement**
Aditi Bhargava, Arthur P Arnold, Debra A Bangasser, Kate M Denton, Arpana Gupta, Lucinda M Hilliard Krause, Emeran A Mayer, Margaret McCarthy, Walter L Miller, Armin Raznahan, Ragini Verma
Endocrine Reviews (2021-03-11) https://doi.org/gm642r
DOI: 10.1210/endrev/bnaa034 · PMID: 33704446 · PMCID: PMC8348944

50. **Considering sex as a biological variable will require a global shift in science culture**
Rebecca M Shansky, Anne Z Murphy
Nature Neuroscience (2021-03-01) https://doi.org/gjhkx8
DOI: 10.1038/s41593-021-00806-8 · PMID: 33649507

51. **Temporal patterns of genes in scientific publications.**
Thomas Pfeiffer, Robert Hoffmann
Proceedings of the National Academy of Sciences of the United States of America (2007-07-09)
https://www.ncbi.nlm.nih.gov/pubmed/17620606
DOI: 10.1073/pnas.0701315104 · PMID: 17620606 · PMCID: PMC1924584

52. **Power-law-like distributions in biomedical publications and research funding.**
Andrew I Su, John B Hogenesch
Genome biology (2007) https://www.ncbi.nlm.nih.gov/pubmed/17472739
DOI: 10.1186/gb-2007-8-4-404 · PMID: 17472739 · PMCID: PMC1895997

53. **Large-scale investigation of the reasons why potentially important genes are ignored.**
Thomas Stoeger, Martin Gerlach, Richard I Morimoto, Luís A Nunes Amaral
PLoS biology (2018-09-18) https://www.ncbi.nlm.nih.gov/pubmed/30226837
DOI: 10.1371/journal.pbio.2006643 · PMID: 30226837 · PMCID: PMC6143198

54. **KLHL21, a novel gene that contributes to the progression of hepatocellular carcinoma.**
Lei Shi, Wenfa Zhang, Fagui Zou, Lihua Mei, Gang Wu, Yong Teng
BMC cancer (2016-10-21) https://www.ncbi.nlm.nih.gov/pubmed/27769251
DOI: 10.1186/s12885-016-2851-7 · PMID: 27769251 · PMCID: PMC5073891

55. **Inhibition of KLHL21 prevents cholangiocarcinoma progression through regulating cell proliferation and motility, arresting cell cycle and reducing Erk activation.**
Jian Chen, Wenfeng Song, Yehui Du, Zequn Li, Zefeng Xuan, Long Zhao, Jun Chen, Yongchao Zhao, Biguang Tuo, Shusen Zheng, Penghong Song
Biochemical and biophysical research communications (2018-03-31)
https://www.ncbi.nlm.nih.gov/pubmed/29574153
DOI: 10.1016/j.bbrc.2018.03.152 · PMID: 29574153

56. **Tumor-promoting mechanisms of macrophage-derived extracellular vesicles-enclosed microRNA-660 in breast cancer progression.**
Changchun Li, Ruiqing Li, Xingchi Hu, Guangjun Zhou, Guoqing Jiang
Breast cancer research and treatment (2022-01-27)
https://www.ncbi.nlm.nih.gov/pubmed/35084622
DOI: 10.1007/s10549-021-06433-y · PMID: 35084622

57. **10 Years of GWAS Discovery: Biology, Function, and Translation**
Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, Jian Yang
The American Journal of Human Genetics (2017-07) https://doi.org/gcsmnm
DOI: 10.1016/j.ajhg.2017.06.005 · PMID: 28686856 · PMCID: PMC5501872

58. **Benefits and limitations of genome-wide association studies**
Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, David Meyre
Nature Reviews Genetics (2019-05-08) https://doi.org/ggcxxb
DOI: 10.1038/s41576-019-0127-1 · PMID: 31068683

59. **Numba**
Siu Kwan Lam, Antoine Pitrou, Stanley Seibert
Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15 (2015) https://doi.org/gf3nks
DOI: 10.1145/2833157.2833162

60. **Clustermatch Correlation Coefficient (CCC)**
Greene Laboratory
(2022-12-08) https://github.com/greenelab/ccc

61. **NCBI GEO: archive for functional genomics data sets—update**
Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, … Alexandra Soboleva
Nucleic Acids Research (2012-11-26) https://doi.org/f3mn62
DOI: 10.1093/nar/gks1193 · PMID: 23193258 · PMCID: PMC3531084

62. **The BioGRID interaction database: 2013 update**
Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, … Mike Tyers
Nucleic acids research (2013-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531226/
DOI: 10.1093/nar/gks1158 · PMID: 23203989 · PMCID: PMC3531226

63. **The IntAct molecular interaction database in 2012**
S Kerrien, B Aranda, L Breuza, A Bridge, F Broackes-Carter, C Chen, M Duesbury, M Dumousseau, M Feuermann, U Hinz, … H Hermjakob
Nucleic Acids Research (2011-11-24) https://doi.org/bpmrdk
DOI: 10.1093/nar/gkr1088 · PMID: 22121220 · PMCID: PMC3245075

64. **MINT, the molecular interaction database: 2012 update**
Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, … Gianni Cesareni
Nucleic Acids Research (2011-11-16) https://doi.org/cqvx3b
DOI: 10.1093/nar/gkr930 · PMID: 22096227 · PMCID: PMC3244991

65. **MIPS: a database for genomes and protein sequences**
HW Mewes, K Heumann, A Kaps, K Mayer, F Pfeiffer, S Stocker, D Frishman
Nucleic acids research (1999-01-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC148093/
DOI: 10.1093/nar/27.1.44 · PMID: 9847138 · PMCID: PMC148093

66. **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles**

Elodie Portales-Casamar, Supat Thongjuea, Andrew T Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W Wasserman, Albin Sandelin
Nucleic Acids Research (2009-11-10) https://doi.org/ddwfqp
DOI: 10.1093/nar/gkp950 · PMID: 19906716 · PMCID: PMC2808906

67. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**
Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, Jill P Mesirov
Proceedings of the National Academy of Sciences (2005-09-30) https://doi.org/d4qbh8
DOI: 10.1073/pnas.0506580102 · PMID: 16199517 · PMCID: PMC1239896

68. **The Sleipnir library for computational functional genomics.**
Curtis Huttenhower, Mark Schroeder, Maria D Chikina, Olga G Troyanskaya
Bioinformatics (Oxford, England) (2008-05-21) https://www.ncbi.nlm.nih.gov/pubmed/18499696
DOI: 10.1093/bioinformatics/btn237 · PMID: 18499696 · PMCID: PMC2718674

69. **minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers**
Davide Albanese, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman, Cesare Furlanello
Bioinformatics (2012-12-14) https://doi.org/f4nxg6
DOI: 10.1093/bioinformatics/bts707 · PMID: 23242262

70. **minepy - Maximal Information-based Nonparametric Exploration**
minepy - Maximal Information-based Nonparametric Exploration (MINE) in C and Python (2022-11-26) https://github.com/minepy/minepy

71. **Measuring Dependence Powerfully and Equitably**
Yakir Reshef, David Reshef, Hilary Finucane, Pardis Sabeti, Michael Mitzenmacher
Journal of Machine Learning Research (2010) https://jmlr.org/papers/v17/15-308.html

72. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, … Edouard Duchesnay
Journal of Machine Learning Research (2011)
https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

73. **An improved algorithm for the maximal information coefficient and its application**
Dan Cao, Yuan Chen, Jin Chen, Hongyan Zhang, Zheming Yuan
Royal Society Open Science (2021-02) https://doi.org/gpcwkd
DOI: 10.1098/rsos.201424 · PMID: 33972855 · PMCID: PMC8074658

74. **A New Algorithm to Optimize Maximal Information Coefficient**
Yuan Chen, Ying Zeng, Feng Luo, Zheming Yuan
PLOS ONE (2016-06-22) https://doi.org/gbpjt7
DOI: 10.1371/journal.pone.0157567 · PMID: 27333001 · PMCID: PMC4917098

75. **RapidMic: Rapid Computation of the Maximal Information Coefficient**
Dongming Tang, Mingwen Wang, Weifan Zheng, Hongjun Wang
Evolutionary Bioinformatics (2014-01) https://doi.org/gpt7c8
DOI: 10.4137/ebo.s13121 · PMID: 24526831 · PMCID: PMC3921152

76. **A Novel Algorithm for the Precise Calculation of the Maximal Information Coefficient**

Yi Zhang, Shili Jia, Haiyun Huang, Jiqing Qiu, Changjie Zhou
Scientific Reports (2014-10-17) https://doi.org/gpt7c7
DOI: 10.1038/srep06662 · PMID: 25322794 · PMCID: PMC4200418

Supplementary material

Supplementary Note 1: Comparison with the Maximal Information Coefficient (MIC) on gene expression data

We compared all the coefficients in this study with MIC [24], a popular nonlinear method that can find complex relationships in data, although very computationally intensive [73]. We ran MIC_e (see Methods) on all possible pairwise comparisons of our 5,000 highly variable genes from whole blood in GTEx v8. This took 4 days and 19 hours to finish (compared with 9 hours for CCC). Then we performed the analysis on the distribution of coefficients (the same as in the main text), shown in Figure 6. We verified that CCC and MIC behave similarly in this dataset, with essentially the same distribution but only shifted. Figure 6 c shows that these two coefficients relate almost linearly, and both compare very similarly with Pearson and Spearman.

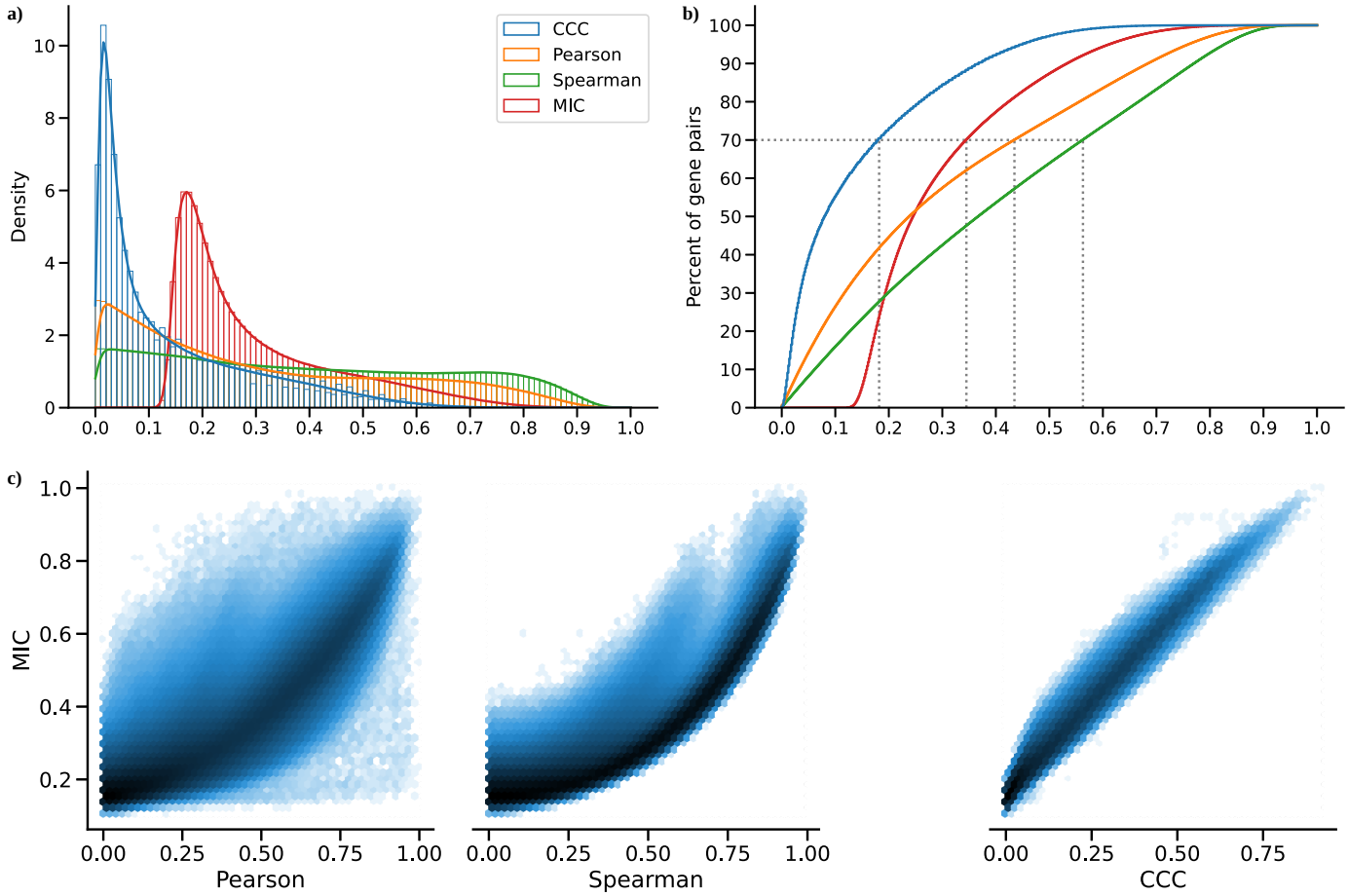

Figure 6: Distribution of MIC values on gene expression (GTEx v8, whole blood) and comparison with other methods. a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all coefficients, where a logarithmic scale was used to color each hexagon.

Supplementary Note 2: Computational complexity of coefficients

We also compared CCC with the other coefficients in terms of computational complexity. Although CCC and MIC might identify similar gene pairs in gene expression data (see here), the use of MIC in large datasets remains limited due to its very long computation time, despite some methodological/implementation improvements [69,73,74,75,76]. The original MIC implementation uses ApproxMaxMI, a computationally demanding heuristic estimator [37]. Recently, a more efficient implementation called MIC_e was proposed [71]. These two MIC estimators are provided by the `minepy` package [69], a C implementation available for Python. We compared all these coefficients in terms of computation time on randomly generated variables of different sizes, which simulates a scenario of gene expression data with different numbers of conditions. Differently from the rest, CCC allows us to easily parallelize the computation of a single gene pair (see Methods), so we also tested the cases using 1 and 3 CPU cores. Figure 7 shows the time in seconds in log scale.

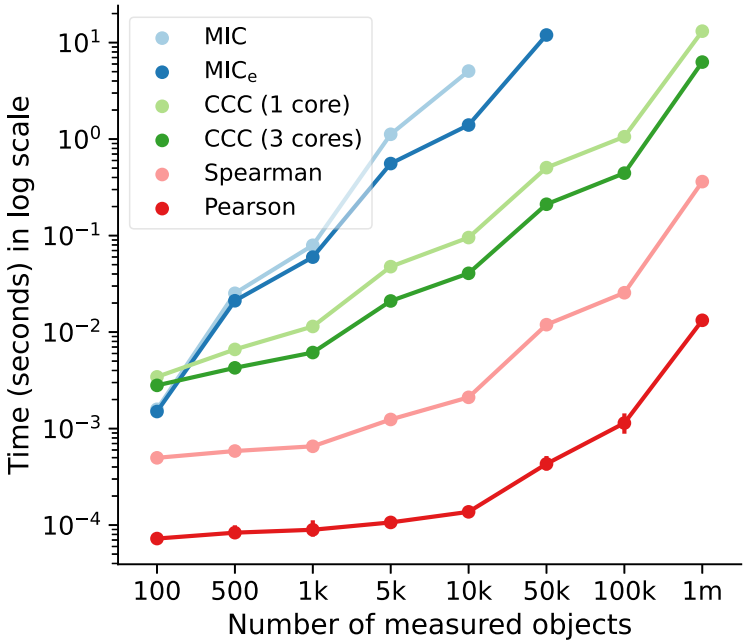

Figure 7: Computational complexity of all correlation coefficients on simulated data. We simulated variables/features with varying data sizes (from 100 to a million, x-axis). The plot shows the average time in seconds (log-scale) taken for each coefficient on ten repetitions (1000 repetitions were performed for data size 100). CCC was run using 1 and 3 CPU cores. MIC and MIC_e did not finish running in a reasonable amount of time for data sizes of 10,000 and 100,000, respectively.

As we already expected, Pearson and Spearman were the fastest, given that they only need to compute basic summary statistics from the data. For example, Pearson is three orders of magnitude faster than CCC. Among the nonlinear coefficients, CCC was faster than the two MIC variations (up to two orders of magnitude), with the only exception in very small data sizes. The difference is important because both MIC variants were implemented in C [69], a high-performance programming language, whereas CCC was implemented in Python (optimized with `numba`). For a data size of a million, the multi-core CCC was twice as fast as the single-core CCC. This suggests that new implementations using more advanced processing units (such as GPUs) are feasible and could make CCC reach speeds closer to Pearson.

Tissue-specific gene networks with GIANT

Table 1: Network statistics of six gene pairs shown in Figure 3 b for blood and predicted cell types. Only gene pairs present in GIANT models are listed. For each gene in the pair (first column), the minimum, average and maximum interaction coefficients with the other genes in the network are shown.

	Interaction confidence						
	Blood			**Predicted cell type**			
Gene	**Min.**	**Avg.**	**Max.**	**Cell type**	**Min.**	**Avg.**	**Max.**
IFNG	0.19	0.42	0.54	Natural killer cell	0.74	0.90	0.99
SDS	0.18	0.29	0.41		0.65	0.81	0.94
JUN	0.26	0.68	0.97	Mononuclear phagocyte	0.36	0.73	0.94
APOC1	0.22	0.47	0.77		0.29	0.50	0.80
ZDHHC12	0.05	0.07	0.10	Macrophage	0.03	0.12	0.33
CCL18	0.74	0.79	0.86		0.36	0.70	0.90
RASSF2	0.69	0.77	0.90	Leukocyte	0.66	0.74	0.88
CYTIP	0.74	0.85	0.91		0.76	0.84	0.91
MYOZ1	0.09	0.17	0.37	Skeletal muscle	0.11	0.11	0.12
TNNI2	0.10	0.22	0.44		0.10	0.11	0.12
PYGM	0.02	0.04	0.14	Skeletal muscle	0.01	0.02	0.04
TPM2	0.05	0.56	0.80		0.01	0.28	0.47