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

This is a testing version edited by an AI bot.

This manuscript (permalink) was automatically generated from miltondp/ccc-manuscript-chatgpt@1eb2937 on December 24, 2022.

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)

- **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

Abstract

We introduce the Clustermatch Correlation Coefficient (CCC), a next-generation not-only-linear correlation coefficient that can efficiently detect linear and nonlinear patterns in data. CCC is based on machine learning models and is much faster than state-of-the-art coefficients such as the Maximal Information Coefficient. When applied to human gene expression data, CCC reveals biologically meaningful linear and nonlinear patterns missed by standard, linear-only correlation coefficients. These patterns include nonlinear associations associated with sex differences that are not detected by linear-only methods. Moreover, gene pairs highly ranked by CCC were enriched for interactions in integrated networks, suggesting that CCC can detect functional relationships that linear-only methods miss. CCC is an easy-to-use and highly-efficient tool that can be applied to genome-scale data and other domains across different data types.

Introduction

The availability of large datasets has enabled researchers to explore complex scientific questions. To identify patterns in these datasets, correlation analysis is an essential statistical technique. Correlation coefficients are widely used to measure the similarity between two objects, such as genes [1] or lifestyle factors [2], and to select features that improve prediction accuracy [3,4]. The Pearson correlation coefficient is widely applied across various application domains and scientific areas. Therefore, even minor improvements to this technique could have a major impact on research and industry.

Analysis of gene expression in transcriptomics often starts with estimating correlations between genes. Sophisticated correlation analysis can suggest gene function [5], identify common and cell-type specific regulatory networks [6], and uncover important interactions in living organisms [7,8]. RNA-seq datasets [9,10] are also used to detect complex transcriptional mechanisms underlying human diseases [1,11,12,13,14]. The introduction of the omnigenic model of complex traits [15,16] has led to increased focus on gene-gene relationships in the study of human diseases [17,18,19,20], including in the field of polygenic risk scores [21]. Combining disease-associated genes from genome-wide association studies (GWAS) with gene co-expression networks to prioritize “core” genes directly affecting diseases [18,19,22] is a recent approach that can capture core genes not identified by standard statistical methods. These genes are believed to be part of highly interconnected, disease-relevant regulatory networks. Advanced correlation coefficients could therefore have wide applications in biology, including in the precision medicine field for the prioritization of candidate drug targets.

Pearson and Spearman correlation coefficients are widely used to identify intuitive linear or monotonic relationships, but they may fail to capture complex yet critical nonlinear patterns. To address this issue, several novel coefficients have been proposed, such as the Maximal Information Coefficient (MIC) [23] and the Distance Correlation (DC) [24], which have been successfully applied in various domains [3,25,26]. However, their computational complexity makes them impractical for even moderately sized datasets [25,27]. We previously developed a clustering method that outperformed Pearson, Spearman, DC and MIC in detecting clusters of simulated linear and nonlinear relationships with varying noise levels [28]. 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. It also provides a high level of flexibility to detect specific types of patterns, while providing safe defaults to capture general relationships. Moreover, its efficient implementation is highly parallelizable, allowing to speed up computation across variable pairs with millions of objects or conditions. To assess its performance, we applied CCC to gene expression data

from the Genotype-Tissue Expression v8 (GTEx) project across different tissues [29]. Our results showed that CCC captured both strong linear relationships and novel nonlinear patterns, which were entirely missed by standard coefficients. Additionally, CCC behaved similarly to MIC in several cases, although it was much faster to compute. Moreover, 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) [30]. Finally, 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

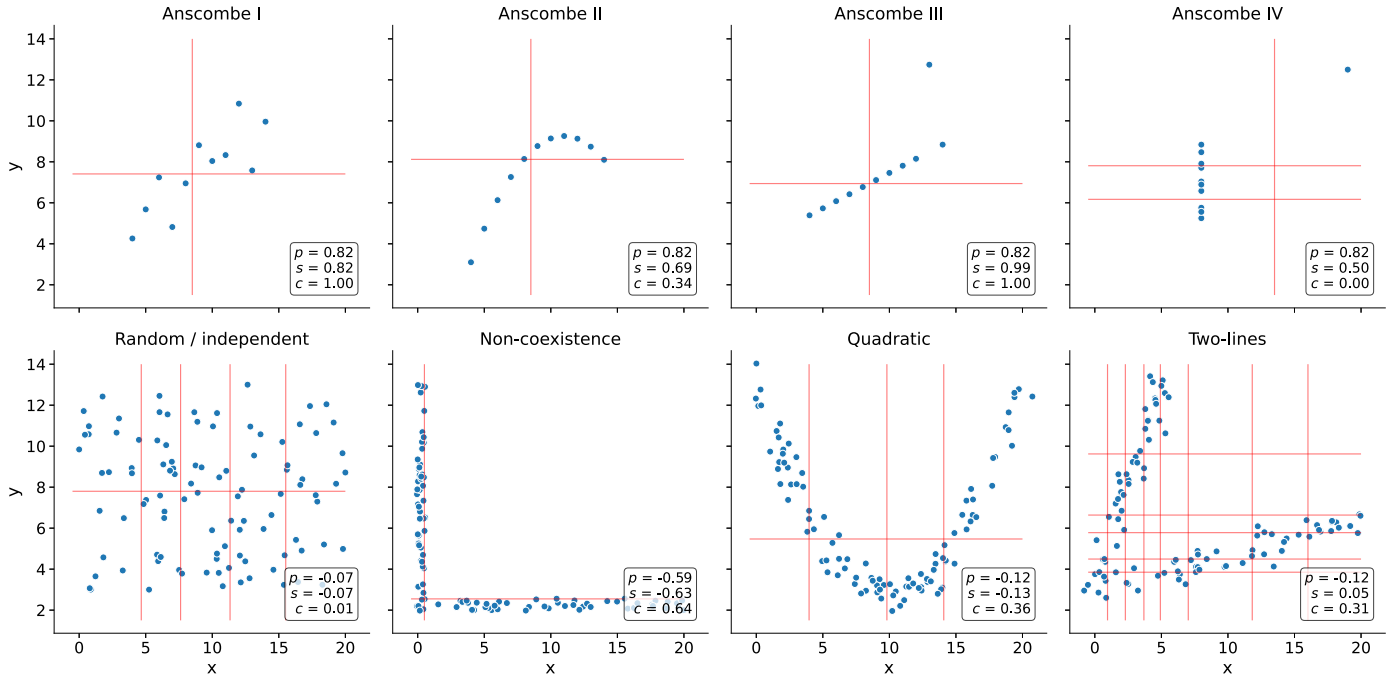

Figure 1: Different types of relationships in data. Each panel contains a set of simulated data points described by two generic variables: x and y. The first row shows Anscombe's quartet with four different datasets (from Anscombe I to IV) and 11 data points each. The second row contains a set of general 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.

Figure ??? shows an example of the CCC calculation.

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, then the clustering of data points using each variable should match. For numerical values, CCC uses quantiles to separate data points into different clusters (e.g., the median separates numerical data into two clusters). The CCC is then defined as the maximum adjusted Rand index (ARI) [31] between the clusterings, ranging from 0 to 1. Further details of the CCC algorithm can be found in the Methods section. An example of the CCC calculation is shown in Figure ???.

We examined the behavior of Pearson (p), Spearman (s) and CCC (c) correlation coefficients on different simulated data patterns. Figure 1 shows the classic Anscombe's quartet [32], 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" [33,34,35], is to remind us of the importance of going beyond simple statistics.

Unwanted patterns (such as outliers) and desirable ones (such as biologically meaningful nonlinear relationships) can be masked by summary statistics alone.

In contrast, CCC is more robust to nonlinear relationships and outliers, as illustrated in Figure $???$.

The Anscombe datasets show that CCC is more robust than Pearson and Spearman when detecting nonlinear relationships and outliers. In Anscombe I, CCC separates data points using two clusters, yielding a strong relationship of 1.0. In Anscombe II, which follows a partially quadratic relationship, CCC yields a lower yet non-zero value of 0.34, reflecting a more complex relationship than a linear pattern. In Anscombe IV, where x values are almost constant except for one outlier, CCC correctly indicates no association for this variable pair with a value of $c = 0.00$. This is in contrast to Pearson's and Spearman's correlation coefficients, which are the same across all these Anscombe's examples ($p = 0.82$ and $s = 0.50$ or greater, respectively). This is illustrated in Figure $???$.

We simulated different types of relationships (see Figure 1, second row), including some previously described from gene expression data [36,37,38]. For the random/independent pair of variables, all coefficients correctly agreed with a value close to zero. The non-coexistence pattern, captured by all coefficients, indicated a case where one gene (x) was expressed while the other one (y) was inhibited, highlighting a potentially strong biological relationship (such as a microRNA negatively regulating another gene). However, for the other two examples (quadratic and two-lines), Pearson and Spearman did not capture the nonlinear pattern between variables x and y. CCC, on the other hand, used different degrees of complexity to capture the relationships. For instance, in the quadratic pattern, CCC separated x into four clusters to reach the maximum ARI. The two-lines example showed two embedded linear relationships with different slopes, which neither Pearson nor Spearman detected ($p = -0.12$ and $s = 0.05$, respectively). Here, CCC increased 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

As shown in Fig. 1, CCC outperformed Pearson and Spearman in capturing nonlinear relationships between genes in whole blood.

We then examined the correlation coefficients in gene expression data from GTEx v8 across different tissues. We chose the top 5,000 genes with the highest variance for our initial analyses in whole blood. We computed the correlation matrix between genes using Pearson, Spearman and CCC (see Methods). As Fig. 1 shows, CCC was better than Pearson and Spearman at capturing nonlinear relationships between genes in whole blood.

We examined the distribution of each coefficient's absolute values in GTEx (Figure 2). CCC (mean=0.14, median=0.08, sd=0.15) had 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 reached a cumulative set containing 70% of gene pairs at different values (Figure 2 b), with CCC at 0.18, Pearson at 0.44, and Spearman at 0.56. This suggests 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 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, has been successfully used in various application domains [3,25,26]. These results suggest that our findings for CCC generalize to MIC. Therefore, in the subsequent analyses, we focused on CCC and linear-only coefficients.

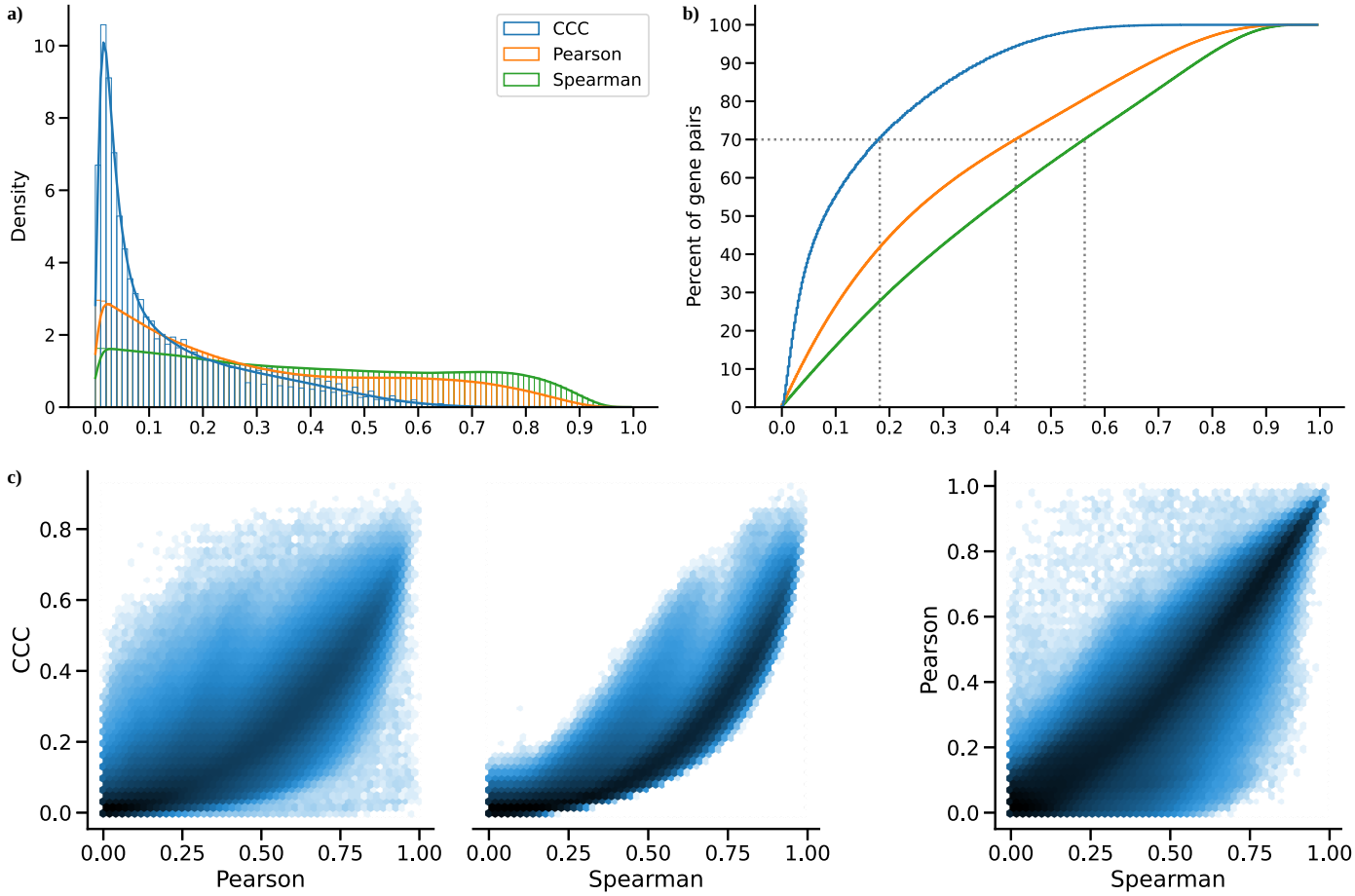

Figure 2: Distribution of coefficient values on gene expression (GTEx v8, whole blood). a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all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

We found that the correlation coefficients identified different patterns in gene expression data. To analyze their agreement and disagreement, we obtained 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. Most of the time (76.4%), the three coefficients agreed on whether there is a strong correlation (42.1%) or no relationship (34.3%). This suggests that these concordant gene pairs most likely represent linear patterns, since Pearson and Spearman are linear-only and CCC can also capture these patterns. Moreover, CCC and Spearman tended to agree more on either highly or poorly correlated pairs (4.0% in the "high" set, and 7.0% in the "low" set) than either of them with Pearson (all between 0.3%-3.5% for the "high" set, and 2.8%-5.5% for the "low" set). In summary, CCC agreed with either Pearson or Spearman in 90.5% of gene pairs by assigning a high or a low correlation value (Figure 3 a).

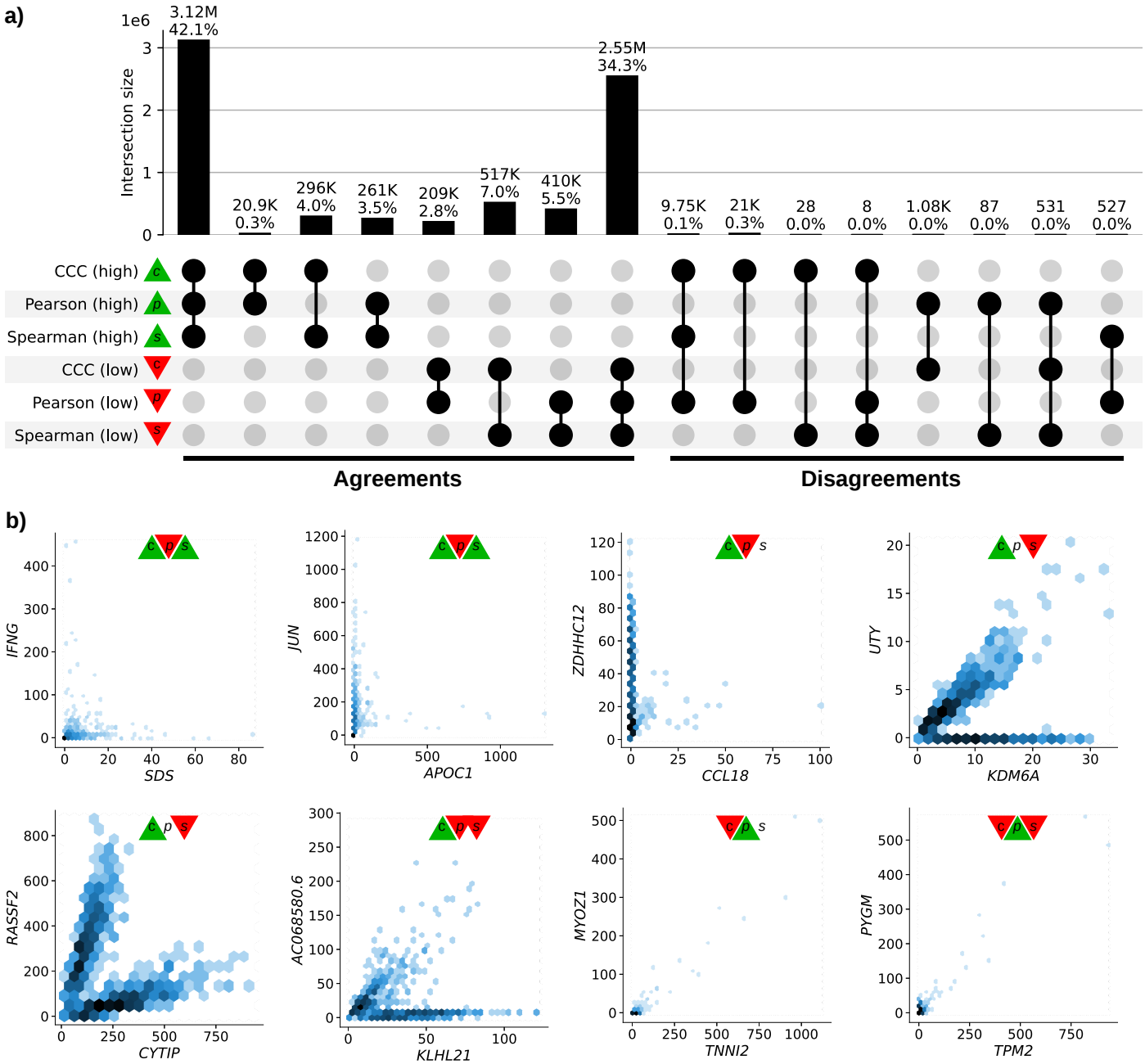

Figure 3: Intersection of gene pairs with high and low correlation coefficient values (GTEx v8, whole blood). a) UpSet plot with six categories (rows) grouping the 30% of the highest (green triangle) and lowest (red triangle) values for each coefficient. Columns show different intersections of categories 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.

We found that more than 20,000 gene pairs had a high CCC value but were not highly ranked by other correlation coefficients (Figure 3 a, right). Additionally, 1,075 gene pairs had a high Pearson value but low CCC, 87 gene pairs had a high Spearman value but low CCC, and 531 gene pairs had both low CCC and low Spearman values. However, our analysis suggests that many of these cases may be caused by potential outliers (Figure 3 b, and discussed in detail later). We further examined the gene pairs that were in the top five of each intersection in the “Disagreements” group (Figure 3 a, right), where CCC disagreed with Pearson, Spearman, or both.

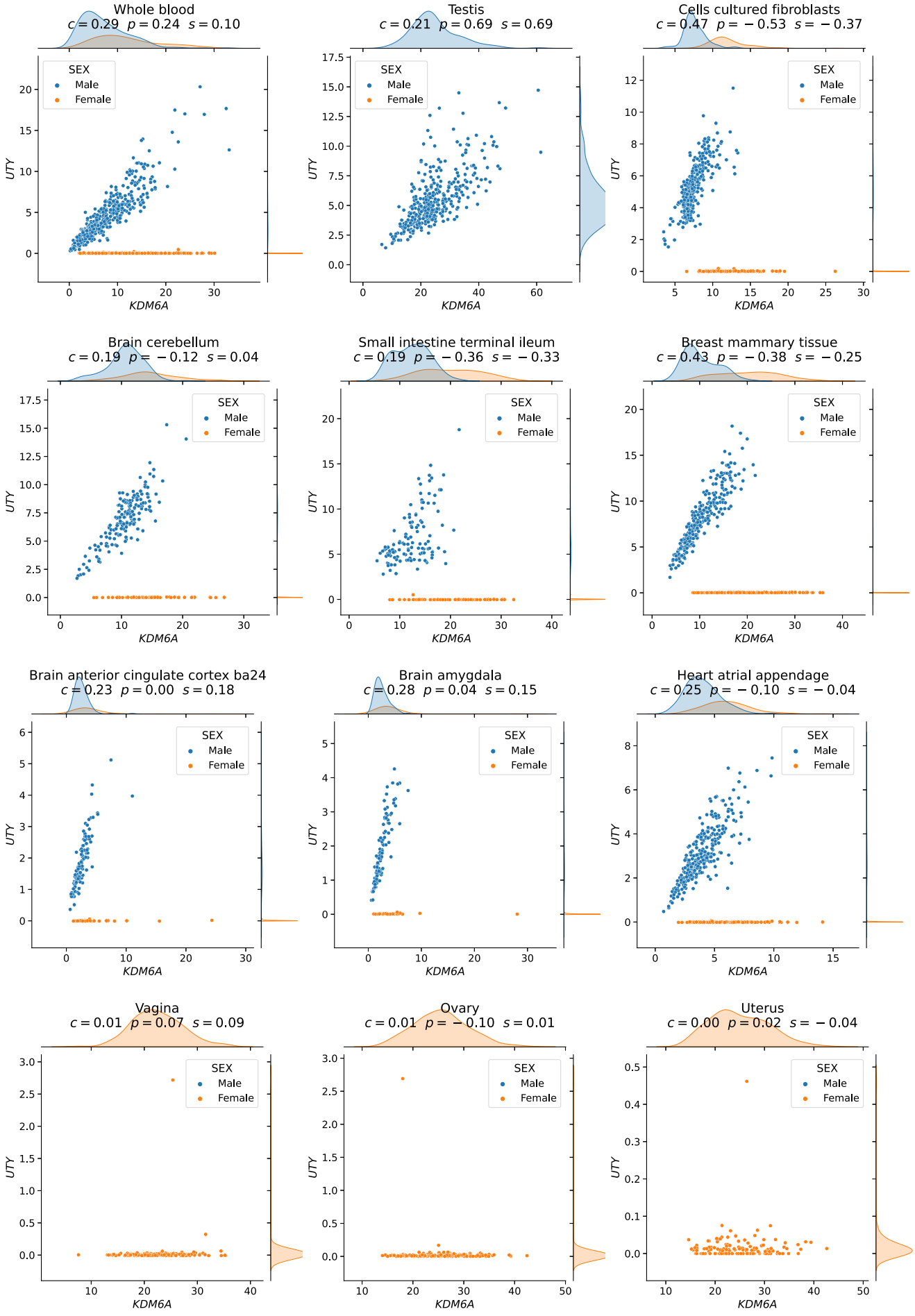

Figure 4: The expression levels of *KDM6A* and *UTY* display sex-specific associations across GTEx tissues. CCC captures this nonlinear relationship in all GTEx tissues (nine examples are shown in the first three rows), except in female-specific organs (last row).

The first three gene pairs (*IFNG* - *SDS*, *JUN* - *APOC1*, and *ZDHHC12* - *CCL18*) show a non-coexistence relationship. Samples with high expression of one gene tend to have low expression of the other, pointing to a potentially inhibiting effect. The following three gene pairs (*UTY* - *KDM6A*, *RASSF2* - *CYTIP*, and *AC068580.6* - *KLHL21*) show patterns combining either two linear or one linear and one independent relationships. For example, *UTY* and *KDM6A* (paralogs) have a nonlinear relationship, where a subset of samples follows a linear pattern and another subset has a constant expression of one gene. 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) [39,40], 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 [41], 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 [42]). 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 [43] 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 ??). The autodetected cell type for this pair was leukocytes, and interaction scores were similar to the blood network (Supplementary Table ??). 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.

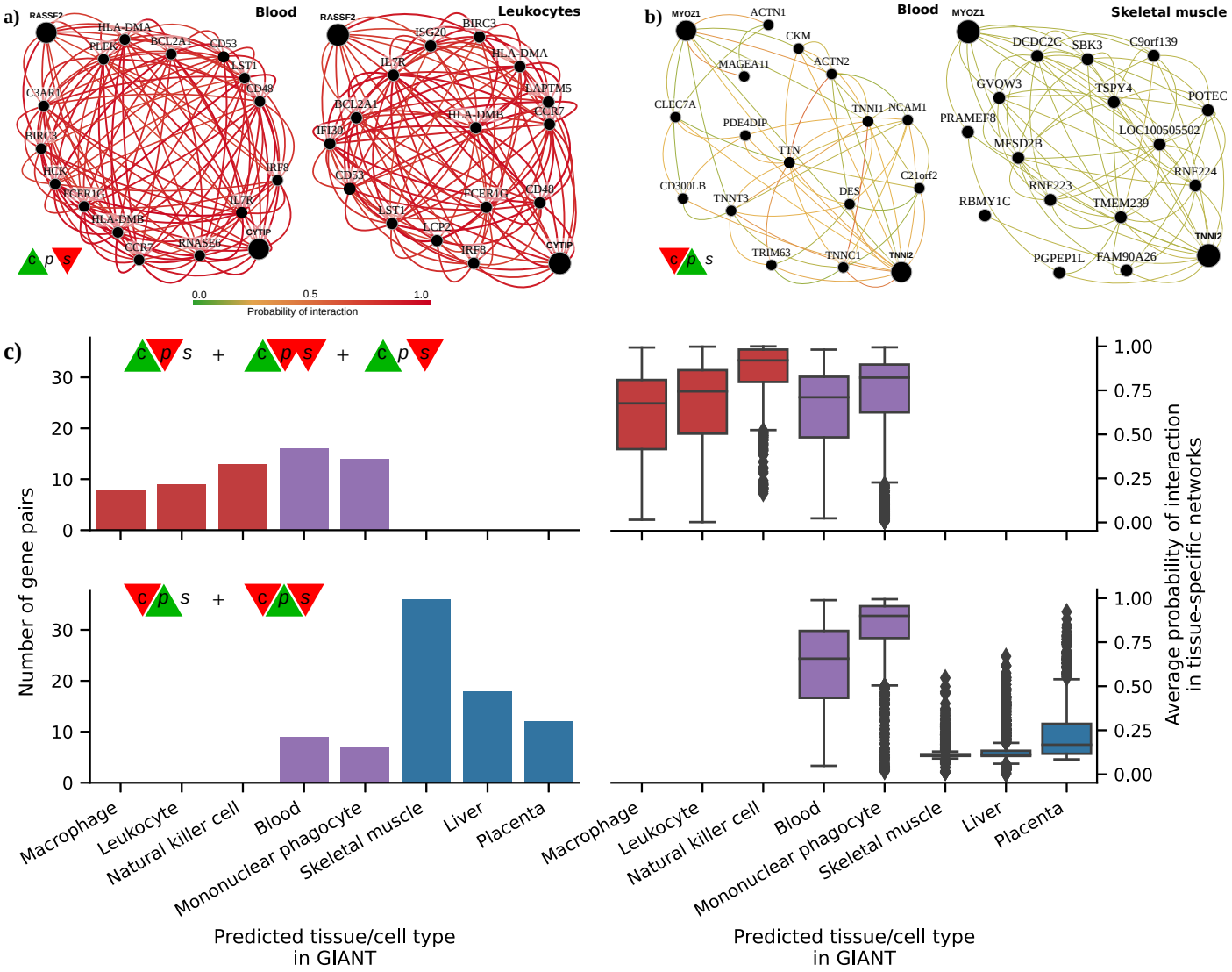

Figure 5: Analysis of GIANT tissue-specific predicted networks for gene pairs prioritized by correlation coefficients. a-b) Two gene pairs prioritized by correlation coefficients (from Figure 3 b) with their predicted networks in blood (left) and an automatically selected tissue/cell type (right) using the method described in [43]. 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* [44], *MYOZ1 - TNNI2* [45]. **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 evaluated the top 100 discrepant gene pairs between CCC and the other two correlation coefficients in GTEx (whole blood). We used GIANT to automatically detect relevant cell types for each gene pair. The top five most commonly predicted cell types for CCC-ranked gene pairs 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-ranked gene pairs were predicted to be specific to tissues unrelated to blood (Figure 5 c, bottom left). 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). These results suggest that 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. In contrast, most Pearson-ranked gene

pairs were not predicted to be blood-specific, and their interaction probabilities were relatively low. This suggests that the associations exclusively detected by CCC in whole blood from GTEx were more strongly replicated in these independent networks. Our findings are consistent with our earlier observations of outlier-driven associations (Figure 3 b).

Discussion

We introduce the Clustermatch Correlation Coefficient (CCC), a machine learning-based statistic that is efficient for detecting not-only-linear relationships. Applying CCC to GTEx v8 data revealed it was robust to outliers and could detect both linear and complex, biologically meaningful patterns that standard coefficients missed. It was able to capture nonlinear patterns from the sex chromosomes, providing insight into sex-specific differences. CCC also detected complex relationships in which a subset of samples or conditions were explained by other factors, such as differences between health and disease. Furthermore, 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 [46]. This suggests that CCC can disentangle intricate cell lineage-specific transcriptional patterns missed by linear-only coefficients. Compared to Spearman and Pearson, CCC was more similar to Spearman in terms of robustness to outliers. Additionally, CCC was more concordant with MIC than Pearson and Spearman, but much faster to compute and thus practical for large datasets. Moreover, CCC can 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.

Visual analysis is helpful in datasets such as Anscombe and “Datasaurus”, but it is infeasible to examine each possible relationship in many datasets. This is where more sophisticated and robust correlation coefficients like CCC become necessary. CCC can detect patterns that may reflect real biology, such as the strong linear relationship between genes *UTY* and *KDM6A* (from sex chromosomes) that was only found in a subset of samples (males). This example highlights the importance of considering sex as a biological variable to avoid overlooking differences between men and women, such as in disease manifestations. A not-only-linear correlation coefficient like CCC can also identify significant differences between variables (such as genes) that are explained by a third factor, which would be entirely missed by linear-only coefficients.

Previous studies have shown that a small portion of human genes receive a disproportionate amount of attention in biomedical research [47,48]. Some of these genes, such as *SDS* (12q24) and *ZDHHC12* (9q34), were found to be the focus of fewer publications than expected [49]. This might be due to the common use of linear correlation coefficients, which could lead researchers to overlook genes with complex coexpression patterns. Therefore, the application of our beyond-linear correlation coefficient on large datasets could provide insights into the function of understudied genes. For instance, *KLHL21* (1p36) and *AC068580.6* (*ENSG00000235027*, in 11p15) have a high CCC value and are not detected by the other coefficients. *KLHL21* has been suggested as a potential therapeutic target for hepatocellular carcinoma [50] and other cancers [51,52]. Its nonlinear correlation with *AC068580.6* could uncover other important players in cancer initiation or progression, particularly in subsets of samples with specific characteristics (as shown in Figure 3 b).

Not-only-linear correlation coefficients can be useful in genetic studies, such as genome-wide association studies (GWAS). GWAS has been successful in understanding the molecular basis of common diseases by estimating the association between genotype and phenotype [53], but the estimated effect sizes of genes identified with GWAS are generally modest and explain only a fraction of the phenotype variance [54]. This can be explained by the omnigenic model for complex traits [15,16], which states that highly-interconnected gene regulatory networks exist, with some core genes

having greater direct effects on the phenotype than others. We and others [18,19,22] have found that integrating gene co-expression networks into genetic studies can help identify these core genes, which are missed by linear-only models like GWAS. Our results suggest that using more advanced and efficient correlation coefficients, such as CCC, to build gene co-expression networks could better estimate gene co-expression profiles and thus more accurately identify core genes. This could lead to more promising candidate drug targets, which could be beneficial for precision medicine.

Our analysis has some limitations. We used a sample with the top variable genes to keep computation time feasible. Although the proposed correlation coefficient (CCC) is faster than existing methods such as Maximum Information Coefficient (MIC), Pearson and Spearman coefficients, which rely on simple data statistics, are still the most computationally efficient. However, our results reveal the advantages of using more advanced coefficients like CCC to detect and study more intricate molecular mechanisms that replicate in independent datasets. Applying CCC on larger compendia, such as recount3 [10] 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. Conducting a more comprehensive analysis of optimal values for this parameter could provide insights to adjust it for different applications or data types.

[55]

Linear and rank-based correlation coefficients are very efficient to calculate, but they cannot capture nonlinear relationships. For example, sex-based patterns are not detectable by linear-only coefficients, but they can be identified by not-only-linear methods. Furthermore, not-only-linear coefficients can detect intricate patterns from expression data that are replicated in models that integrate different data modalities. The CCC (correlation coefficient based on machine learning) is particularly efficient and can be further accelerated with GPU-based implementations. This next-generation correlation coefficient is highly effective in transcriptome analyses and may be useful in many other areas. [55]

Methods

We used the CCC coefficient to measure the correlation between gene expression and nonlinear relationships. The CCC coefficient is defined as $CCC(X,Y) = \frac{2 \cdot \mathrm{cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}$, where μ_X and μ_Y are the means of X and Y, respectively, and σ_X^2 and σ_Y^2 are the variances of X and Y, respectively [56].

The code needed to reproduce all of our analyses and generate the figures is available at https://github.com/greenelab/ccc. We provide scripts to download the required data and run all the steps. A Docker image is also available to use the same runtime environment. The CCC coefficient was used to measure the correlation between gene expression and nonlinear relationships. The CCC coefficient is defined as $CCC(X, Y) = \frac{2 (X, Y)}{_X^2 + _Y^2 + (_X -

The CCC algorithm

The Clustermatch Correlation Coefficient (CCC) computes a similarity value $c \in [0, 1]$ between two numerical or categorical features $\mathbf{x}$ and $\mathbf{y}$ measured on n objects. CCC assumes that if two features are similar, then the partitioning of the objects using each feature separately should match. For example, given two features $\mathbf{x}$ and $\mathbf{y}$, partitioning each variable into two clusters using their medians would result in the same partition for both features. The agreement between these partitions can be computed using any measure of similarity between partitions, such as the adjusted Rand index (ARI) [31], which returns the maximum value (1.0). Note that the same number of clusters 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 when 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 searches for the optimal number of clusters given a maximum k, which is its single parameter $k_{\max}$.

Algorithm 1: CCC algorithm

1 **Function** `get_partitions`($\mathbf{v}$, $k_{\max}$)**:**
  Output:
    Ω_r: clustering with r clusters over n objects
2 **if** $\mathbf{v} \in \mathbb{R}^n$ **then**
3 **for** $r \leftarrow 2$ **to** $\min\{k_{\max}, |\mathbf{v}| - 1\}$ **do**
4 $\boldsymbol{\rho} \leftarrow (\rho_\ell \mid \Pr(v_i < \rho_\ell) \leq (\ell - 1)/r), \forall \ell \in [1, r + 1]$
5 $\Omega_{r\ell} \leftarrow \{i \mid \rho_\ell < v_i \leq \rho_{\ell+1}\}, \forall \ell \in [1, r]$
6 **else**
7 $\mathcal{C} \leftarrow \cup_j \{v_i\}$
8 $r \leftarrow |\mathcal{C}|$
9 $\Omega_{rc} \leftarrow \{i \mid v_i = \mathcal{C}_c\}, \forall c \in [1, r]$
10 $\Omega \leftarrow \{\Omega_r \mid |\Omega_r| > 1\}, \forall r$
11 **return** Ω
12
13 **Function** `ccc`($\mathbf{x}$, $\mathbf{y}$, $k_{\max}$)**:**
  Input:
    $\mathbf{x}$: feature values on n objects
    $\mathbf{y}$: feature values on n objects
    $k_{\max}$: maximum number of internal clusters
  Output:
    c: similarity value for $\mathbf{x}$ and $\mathbf{y}$ ($c \in [0, 1]$)
14 $\Omega^{\mathbf{x}} = \text{get\_partitions}(\mathbf{x}, k_{\max})$
15 $\Omega^{\mathbf{y}} = \text{get\_partitions}(\mathbf{y}, k_{\max})$
16 $c \leftarrow \max\{\mathcal{A}(\Omega^{\mathbf{x}}_p, \Omega^{\mathbf{y}}_q)\}, \forall p, q$
17 **return** $\max(c, 0)$

The algorithm `ccc` generates partitionings for each feature $\mathbf{x}$ and $\mathbf{y}$ (lines 14 and 15). It then computes the Adjusted Rand Index (ARI) between each partition in $\Omega^{\mathbf{x}}$ and $\Omega^{\mathbf{y}}$ (line 16), and keeps the pair that produces the highest ARI. Since ARI can return negative values, which are not meaningful in our case, CCC returns values between 0 and 1 (line 17).

This is an important advantage compared to other correlation coefficients that are based on linear relationships [57].

CCC only needs a pair of partitions to compute a similarity value, so any type of feature that can be used for clustering/grouping is supported. For numerical features (lines 2 to 5 in the `get_partitions` function), quantiles are used for clustering from $k = 2$ to $k = k_{\max}$, for example the median generates $k = 2$ clusters of objects. For categorical features (lines 7 to 9), categories are used to group objects. This means numerical and categorical variables can be integrated since clusters do not need an order, which is an advantage compared to correlation coefficients that rely on linear relationships [57].

For our analyses, we set the maximum number of clusters, $k_{\max}$, to 10. This means that for each gene pair, 18 partitions were generated (9 for each gene, ranging from 2 to 10 clusters) and 81 Adjusted Rand Index (ARI) comparisons were performed. To reduce computation time, smaller values of $k_{\max}$ can be used, although this may lead to missing more complex/general relationships. Our examples in Figure 1 suggest that using $k_{\max} = 2$ would force CCC to find only linear relationships, which could be a valid use case scenario if only this type of relationship is desired. Additionally, $k_{\max} = 2$ implies that only two partitions are generated, and only one ARI comparison is performed. Our Python implementation of CCC provides flexibility in specifying $k_{\max}$. For instance, instead of the maximum k (an integer), the parameter can be a custom list of integers, such as `[2, 5, 10]`, which will partition the data into two, five and ten clusters.

We used three CPU cores to speed up the computation of our correlation coefficient CCC. This allowed us to parallelize the process of generating partitions and computing the similarity of a single pair of features (genes in our study). In the future, we could potentially use graphical processing units (GPU) to further speed up the computation of CCC [**doi:10.1186/s12859-016-1044-6?**].

We have created a Python implementation of the CCC, which is optimized using the `numba` package [58]. This implementation is available in our Github repository [59]. Additionally, we have published a package in the Python Package Index (PyPI) that can be easily installed.

Gene expression data and preprocessing

We obtained GTEx v8 data for all tissues and normalized it using transcripts per million (TPM). We focused our primary analysis on whole blood, which had a good sample size (755). To avoid a bias towards highly-expressed genes, we standardized the data with $log(x + 1)$ and then selected the top 5,000 genes based on their variance. We then computed Pearson, Spearman, MIC, and CCC correlations on these 5,000 genes across all 755 samples on the TPM-normalized data. This generated a pairwise similarity matrix of size 5,000 x 5,000.

Tissue-specific network analyses using GIANT

The final correlation coefficient (ρ) was computed as:

We used gene expression data from the NCBI's Gene Expression Omnibus (GEO) [60], protein-protein interaction (BioGRID [61], IntAct [62], MINT [63] and MIPS [64]), transcription factor regulation using binding motifs from JASPAR [65], and chemical and genetic perturbations from MSigDB [66] to build tissue-specific gene networks of GIANT [40]. This GIANT version included 987 genome-scale datasets with approximately 38,000 conditions from around 14,000 publications. Details on the building of the networks are described in [30]. We used a naive Bayesian classifier (implemented in the Sleipnir library [67]) to estimate the probability of tissue-specific interactions for each gene pair. This classifier was trained with gold standards built from expert curation and experimentally derived gene annotations from the Gene Ontology. The final correlation coefficient (ρ) was computed as:

For each gene pair prioritized in our study using GTEx, we used GIANT and HumanBase to obtain two gene networks. The first network was manually selected to match the whole blood tissue in GTEx. The second network was automatically predicted using a machine learning model described in [43], which was provided by HumanBase web interfaces/services. This model was trained using comprehensive transcriptional data, with human-curated markers of different cell lineages (e.g., macrophages) as gold standards. Then, these models were used to predict other cell lineage-specific genes. Besides the predicted tissue or cell lineage, we computed the average probability of interaction between all genes in the networks retrieved from GIANT. We included the top 15 genes with the highest probability of interaction with the queried gene pair for each network, following the default procedure used in GIANT ($P_{ij} \geq 0.5$).

Maximal Information Coefficient (MIC)

We used the Python package `minepy` [68,69] (version 1.2.5) to estimate the MIC coefficient. For the GTEx v8 (whole blood) dataset, we used MIC_e (an improved implementation of the original MIC introduced in [70]) with the default parameters `alpha=0.6`, `c=15` and `estimator='mic_e'`. To parallelize the computation of MIC, we used the `pairwise_distances` function from `scikit-learn` [71]. For our computational complexity analyses (see Supplementary Material), we ran both the original MIC (with parameter `estimator='mic_approx'`) and MIC_e (`estimator='mic_e'`).

References

1. **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

2. **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

3. **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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **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

23. **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

24. **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

25. **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

26. **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

27. **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

28. **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

29. **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

30. **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

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

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

33. **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

34. **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

35. **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

36. **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

37. **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

38. **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

39. **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

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

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

42. **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

43. **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

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

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

46. **Priority setting in head and neck oncology in low-resource environments.**
Luiz P Kowalski, Alvaro Sanabria
Current opinion in otolaryngology & head and neck surgery (2019-06)
https://www.ncbi.nlm.nih.gov/pubmed/30870186
DOI: 10.1097/moo.0000000000000530 · PMID: 30870186

47. **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

48. **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

49. **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

50. **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

51. **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

52. **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

53. **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

54. **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

55. **[The laboratory in programs for enteric infection control].**
OB Grados
Boletin de la Oficina Sanitaria Panamericana. Pan American Sanitary Bureau (1975-04)
https://www.ncbi.nlm.nih.gov/pubmed/123456
PMID: 123456

56. **Denpasar Declaration on Population and Development.** *Integration (Tokyo, Japan)* (1994-06)
https://www.ncbi.nlm.nih.gov/pubmed/12345678
DOI: 10.1234/2013/999990 · PMID: 12345678

57. **Superconductivity with twofold symmetry in Bi**
Mingyang Chen, Xiaoyu Chen, Huan Yang, Zengyi Du, Hai-Hu Wen
Science advances (2018-06-08) https://www.ncbi.nlm.nih.gov/pubmed/29888330

DOI: 10.1126/sciadv.aat1084 · PMID: 29888330 · PMCID: PMC5993472

58. **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

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

60. **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

61. **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

62. **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

63. **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

64. **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

65. **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

66. **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

67. **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

68. **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

69. **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

70. **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

71. **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

72. **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

73. **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

74. **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

75. **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 the correlation coefficients in this study to MIC [23], a popular nonlinear method that can find complex relationships in data. MIC is computationally intensive [72], so we ran the modified version (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 to 9 hours for CCC. We then analyzed the distribution of coefficients (the same as in the main text), shown in Figure 6. We found that CCC and MIC behaved similarly in this dataset, with essentially the same distribution but only shifted. Figure 6 c shows that these two coefficients almost linearly relate to each other, and both compare very similarly to Pearson and Spearman correlation coefficients.

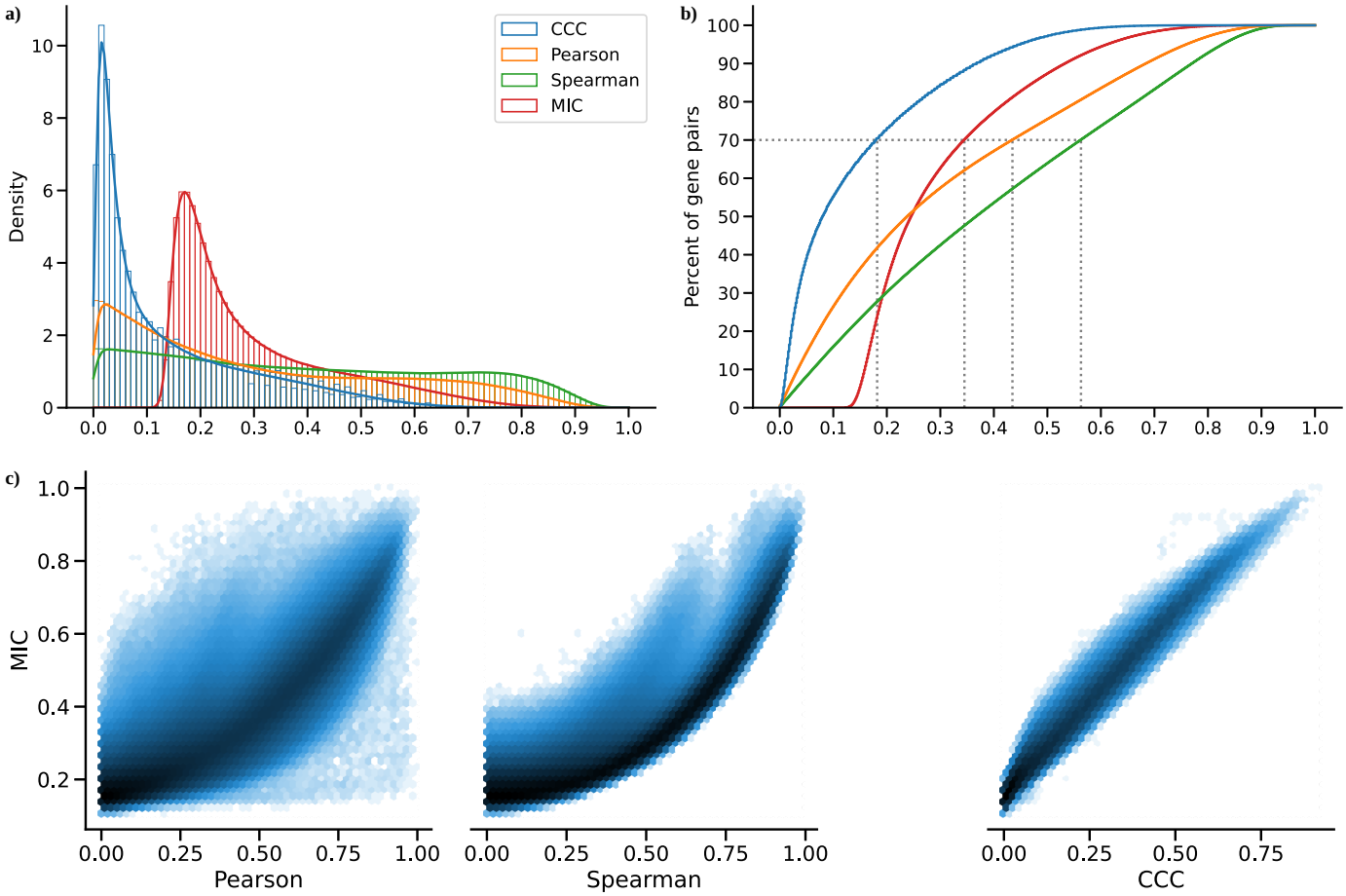

Figure 6: Distribution of MIC values on gene expression (GTEx v8, whole blood) and comparison with other methods. a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with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

Supplementary Note 2: Computational complexity of coefficients

We compared CCC with other correlation coefficients in terms of computational complexity. We simulated a gene expression data scenario with different numbers of conditions by generating random variables of different sizes. Figure 7 shows the time in seconds (in log scale) for each coefficient. CCC is the fastest in most cases, and its performance can be further improved by using 1 or 3 CPU cores. Other coefficients, such as MIC, might identify similar gene pairs in gene expression data [68,72,73,74,75], but their use in large datasets is limited due to their long computation time. This is mainly because the original MIC implementation uses ApproxMaxMI, a computationally demanding heuristic estimator [36]. Recently, a more efficient implementation called MIC_e was proposed [70], and is provided by the `minepy` package [68], a C implementation available for Python. CCC allows us to easily parallelize the computation of a single gene pair (see Methods), which is not possible with the other coefficients.

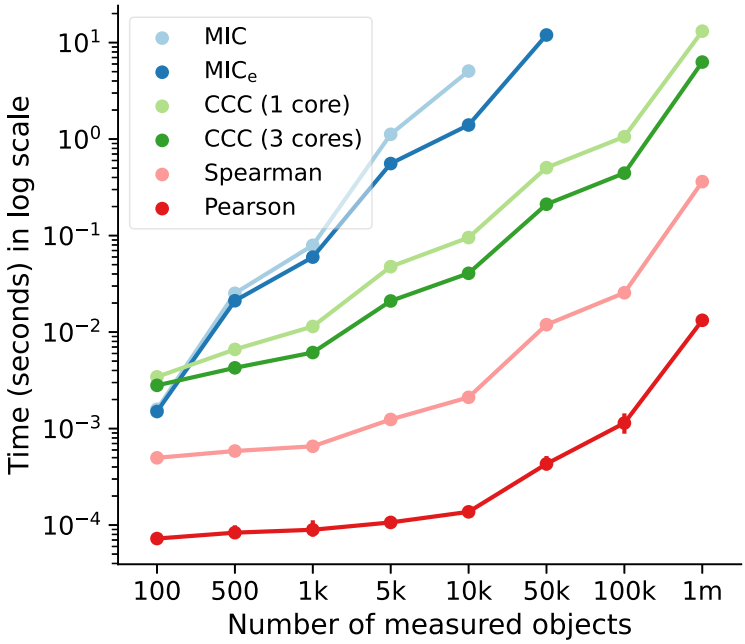

Figure 7: Computational complexity of all correlation coefficients on simulated data. We simulated variables/features with varying data sizes (from 100 to a million, x-axis). The plot shows the average time in seconds (log-scale) taken for each coefficient on ten repetitions (1000 repetitions were performed 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 shown in Figure ???, the new correlation coefficient (NLC) was orders of magnitude faster than the nonlinear ones, except for very small data sizes.

We expected Pearson and Spearman to be the fastest, as they only need to calculate 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 exception of very small data sizes. The difference is significant as both MIC variants were implemented in the high-performance programming language C [68], while CCC was implemented in Python (optimized with `numba`). When data size was a million, the multi-core CCC was twice as fast as the single-core CCC. This suggests that more advanced processing units (such as GPUs) could be used for new implementations and help CCC reach speeds closer to Pearson (see Figure ???). Our new correlation coefficient (NLC) was orders of magnitude faster than the nonlinear ones, except for very small data sizes.

Tissue-specific gene networks with GIANT

	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

The table below shows the network statistics of six gene pairs, as presented in Figure 3 b, for both blood and predicted cell types. It includes only those gene pairs present in GIANT models. For each gene in the pair (first column), the table provides the minimum, average, and maximum interaction coefficients with other genes in the network.{#tbl:giant:weights}