Cell Systems

An efficient, not-only-linear correlation coefficient based on clustering

Highlights

- CCC is a not-only-linear, computationally efficient correlation coefficient
- CCC is a clustering-based statistic
- CCC reveals meaningful linear and nonlinear patterns missed by standard coefficients
- It can readily be applied to genome-scale data across different data types

Authors

Milton Pividori, Marylyn D. Ritchie, Diego H. Milone, Casey S. Greene

Correspondence

milton.pividori@cuanschutz.edu (M.P.), casey.s.greene@cuanschutz.edu (C.S.G.)

In brief

Correlation coefficients are widely used to identify patterns. We introduce CCC, an efficient, easy-to-use coefficient based on clustering that reveals biologically meaningful linear and nonlinear patterns. When applied to human gene expression data, CCC identifies nonlinear patterns explained by sex differences that are not captured by standard and linear-only coefficients.

Pividori et al., 2024, Cell Systems *15*, 1–15
September 18, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2024.08.005

CellPress

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

Cell Systems

CellPress
OPEN ACCESS

Methods

An efficient, not-only-linear correlation coefficient based on clustering

Milton Pividori,[1,2,5,*] Marylyn D. Ritchie,[2] Diego H. Milone,[3] and Casey S. Greene[1,4,*]
[1]Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO 80045, USA
[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]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
[4]Center for Health AI, University of Colorado School of Medicine, Aurora, CO 80045, USA
[5]Lead contact
*Correspondence: milton.pividori@cuanschutz.edu (M.P.), casey.s.greene@cuanschutz.edu (C.S.G.)
https://doi.org/10.1016/j.cels.2024.08.005

SUMMARY

Identifying meaningful patterns in data is crucial for understanding complex biological processes, particularly in transcriptomics, where genes with correlated expression often share functions or contribute to disease mechanisms. Traditional correlation coefficients, which primarily capture linear relationships, may overlook important nonlinear patterns. We introduce the clustermatch correlation coefficient (CCC), a not-only-linear coefficient that utilizes clustering to efficiently detect both linear and nonlinear associations. CCC outperforms standard methods by revealing biologically meaningful patterns that linear-only coefficients miss and is faster than state-of-the-art coefficients such as the maximal information coefficient. When applied to human gene expression data from genotype-tissue expression (GTEx), CCC identified robust linear relationships and nonlinear patterns, such as sex-specific differences, that are undetectable by standard methods. Highly ranked gene pairs were enriched for interactions in integrated networks built from protein-protein interactions, transcription factor regulation, and chemical and genetic perturbations, suggesting that CCC can detect functional relationships missed by linear-only approaches. CCC is a highly efficient, next-generation, not-only-linear correlation coefficient for genome-scale data. A record of this paper's transparent peer review process is included in the supplemental information.

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] These coefficients 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-sequencing (RNA-seq) datasets[10,11] can also reveal complex transcriptional mechanisms underlying human diseases.[2,12–15] 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,[15,18–20] even in specific fields such as polygenic risk scores.[21] In this context, recent approaches combine disease-associated genes from genome-wide association studies (GWASs) with gene co-expression networks to prioritize "core" genes directly affecting diseases.[19,20,22] 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

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

proposed as metrics that capture nonlinear patterns, such as the maximal information coefficient (MIC)[23] and the distance correlation (DC).[24] MIC, in particular, is one of the most commonly used statistics to capture more complex relationships, with successful applications across several domains.[4,25,26] However, the computational complexity makes them impractical for even moderately sized datasets.[25,27] Recent implementations of MIC, for example, take several seconds to compute on a single variable pair across a few thousand objects or conditions.[25] We previously developed a clustering method for highly diverse datasets that outperformed approaches based on 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. 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 us 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 (GTEx) v8 project across different tissues.[10] 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 integrated analysis of gene networks in tissues (GIANT).[7] 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

Overview of CCC: A not-only-linear correlation coefficient

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 clusterings 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). For categorical values, CCC uses the categories themselves to separate data points into different clusters (e.g., if feature color has three values, red, green, and blue, then data will be clustered into three clusters defined by those colors). Once all clusterings are generated according to each variable, we define the CCC as the maximum adjusted Rand index (ARI)[29] between them, ranging between 0 and 1. Details of the CCC algorithm can be found in Box 1.

We examined how the Pearson (p), Spearman (s), and CCC (c) correlation coefficients behaved on different simulated data patterns. Figure 1 shows different types of relationships between two variables of different data types, where x and y are numerical and w and z are categorical. For each variable pair, we show the coefficient values and their statistical significance, where asterisks indicate different p values (p). The red lines show how CCC clustered numerical data points using x (vertical lines) and y (horizontal lines).

In Figure 1A, we examine the classic Anscombe's quartet,[30] 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,"[31,32] 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 a statistically significant value of 1.0 (the maximum for CCC) and thus indicating a strong relationship. Anscombe II seems to follow a partially quadratic relationship, interpreted as linear by Pearson and Spearman. By contrast, for this potentially undersampled quadratic pattern, CCC yields a lower and not statistically significant 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 (which correctly identifies no relationship) as it does for Pearson or Spearman, although only Pearson yields a statistically significant result. 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 of Anscombe's examples ($p = 0.82$), whereas Spearman's is 0.50 or greater.

We also simulated additional types of numerical relationships (Figure 1B), including some previously described from gene expression data.[23,33,34] For the random/independent pair of variables, all coefficients correctly agree with a value close to zero and $p > 0.05$. 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), only CCC is able to yield a high and statistically significant correlation value, whereas Pearson and Spearman fail to capture these nonlinear patterns. These relationships 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

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

CellPress
OPEN ACCESS

Box 1. The CCC algorithm

Algorithm 1: CCC algorithm

```
1  Function get_partitions(v, k_max):
       Input:
           v: feature values on n objects
           k_max: maximum number of clusters
       Output:
           Π: a set of partitions over n objects
2      if v ∈ ℝ^n then
3          for k ← 2 to min{k_max, n − 1} do
4              ρ ← (ρ_ℓ | Pr(v_i < ρ_ℓ) ≤ (ℓ − 1)/k), ∀ℓ ∈ [1, k + 1]
5              π_ℓ ← {i | ρ_ℓ < v_i ≤ ρ_{ℓ+1}}, ∀ℓ ∈ [1, k]
6              Π_k ← π
7      else
8          C ← {c_1, c_2, ..., c_m} (set of m unique categorical values in v)
9          π_ℓ ← {i | v_i = C_ℓ}, ∀ℓ ∈ [1, m]
10         Π_m ← π
11     Π ← {Π_k | |Π_k| > 1}, ∀k
12     return Π
13
14 Function ccc(x, y, k_max):
       Input:
           x: feature values on n objects
           y: feature values on n objects
           k_max: maximum number of clusters
       Output:
           c: correlation value for x and y (c ∈ [0, 1])
15     Π^x = get_partitions(x, k_max)
16     Π^y = get_partitions(y, k_max)
17     c ← max{ARI(π_j, π_l)}, ∀π_j ∈ Π^x, π_l ∈ Π^y
18     return max(c, 0)
```

The main function of the algorithm, CCC, generates a set of partitions Π^x for variable x (line 16) and another set of partitions Π^y for variable y (line 17). Then, it computes the ARI between each partition $\pi_j \in \Pi^x$ and $\pi_l \in \Pi^y$ and gets the maximum (line 18), returning either this value or zero if this is negative (line 19).

Since CCC only needs a set of partitions to compute a correlation value, any type of variable that can be used to perform clustering is supported. If variable v is numerical (lines 2–6 in the get_partitions function), each partition π is generated using a set of quantiles ρ. For example, if $k = 2$, then $\rho = (\rho_1, \rho_2, \rho_3)$, where ρ_1 is the minimum value of v, ρ_2 is the median, and ρ_3 is the maximum value of v. Then, the first cluster π_1 contains all values of v that are less than or equal to ρ_2, and π_2 contains all values of v that are greater than ρ_2. If variable v is categorical (lines 8–11), we compute a single partition π with $m = |C|$ clusters, where $C = \{c_1, c_2, \ldots, c_m\}$ is a set of unique categorical values in v. Therefore, all variable types are internally represented as partitions, and it is not necessary to access the original data values to compute the ARI. Consequently, numerical and categorical variables can be naturally integrated.

Our algorithm implementation uses $k_{max} = 10$ as the default. This means that for a variable pair, 18 partitions are generated (9 for each variable, 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 relationship is desired. In addition, $k_{max} = 2$ implies that only two partitions are generated and only one ARI comparison is performed.

As a final remark, generating partitions (lines 15 and 16) and computing their similarity (line 17) can be easily parallelized. We used three CPU cores in our analyses to speed up the computation of CCC, and this could be potentially extended to a large number of processors using a graphics processing unit (GPU).

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

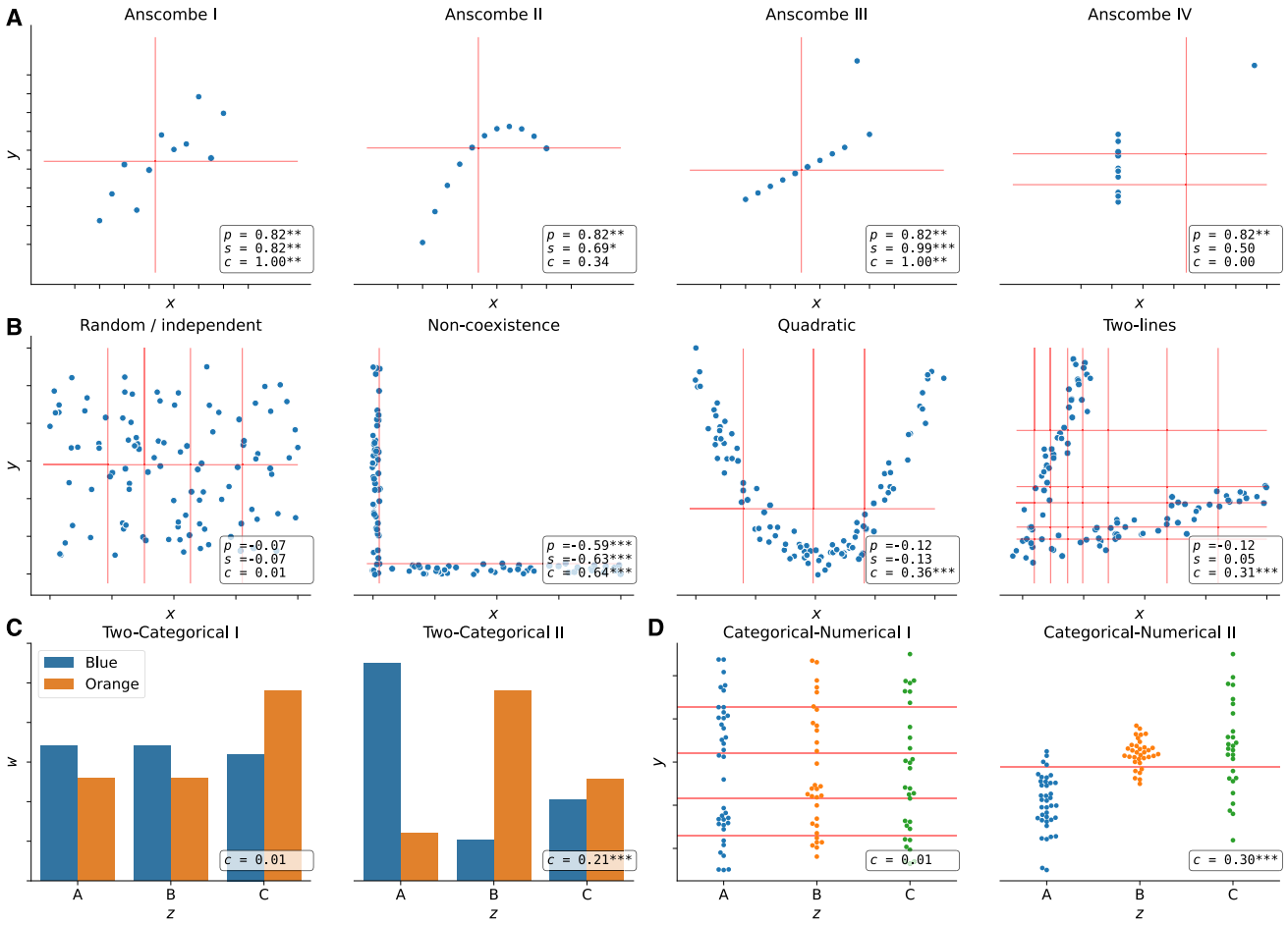


Figure 1. Different patterns across data types

Each panel contains a set of simulated datasets described by two generic variables.

(A) The Anscombe's quartet with four different datasets (from Anscombe I to IV) with numerical variables *x* and *y* and 11 data points.

(B) Four datasets with 100 data points.

(C) Two datasets with categorical variables w (with values orange and blue) and *z* (with values A, B, and C) and 100 data points.

(D) Two datasets with categorical and numerical variables and 100 data points. Each panel shows the correlation value using Pearson (*p*) and Spearman (*s*) for numerical variables only and CCC (*c*) for both numerical and categorical, and their statistical significance is indicated with asterisks: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***). For CCC, we computed the *p* value using 10,000 permutations. Vertical and horizontal red lines show how CCC clustered data points using *x* and *y*, respectively. For categorical variables, CCC uses the categories to cluster data points.

shows two embedded linear relationships with different slopes, which neither Pearson nor Spearman detects ($p = -0.12$ and $s = 0.05$, respectively, both with $p > 0.05$). Here, CCC increases the complexity of the model by using eight clusters for *x* and six for *y*, resulting in $c = 0.31$ ($p < 0.001$).

Furthermore, we also simulated categorical variables, which only CCC can handle. Figure 1C shows two patterns between variables *w* (with categories orange and blue) and *z* (with categories A, B, and C). The first case (two-categorical I) represents a random/independent pattern where categorical values in one variable are approximately uniformly distributed across the categorical values of the other variable. Here, as expected, CCC yields a very low and non-significant value. In the second case (two-categorical II), the category blue of *w* is overrepresented in data points with *z* equal to A and, less strongly, the category orange of *w* is overrepresented in data points with *z* equal to B. In this case, since CCC clusters data points using the categorical values, it detects that clusters of data points with *w* = blue match clusters with *z*= A, yielding a statistically significant $c = 0.21$. Figure 1D mixes a categorical variable (*z*) with a numerical one (*y*). The first case (categorical-numerical I) represents a random/independent pattern where numerical values in *y* are approximately uniformly distributed across the categorical values in *z*. Similarly, as in the other random/independent cases, CCC yields a very low and non-significant value since the clusters formed by *y* do not match the clusters (given by the categorical values) formed by *z*. Conversely, in the second case (categorical-numerical II), clusters of data points with similar values in *y* tend to have also similar categorical values in *z*. In this example, for data points with *z* = A, we assigned $y \sim N(0, 0.5^2)$, whereas for *z*= B and C, we assigned $y \sim N(1, 0.25^2)$ and $y \sim N(1, 0.75^2)$, respectively. Here, CCC uses *y* values to group data points into two clusters, and these clusters match the clusters obtained from *z*, yielding a statistically significant $c = 0.30$.

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

CellPress
OPEN ACCESS

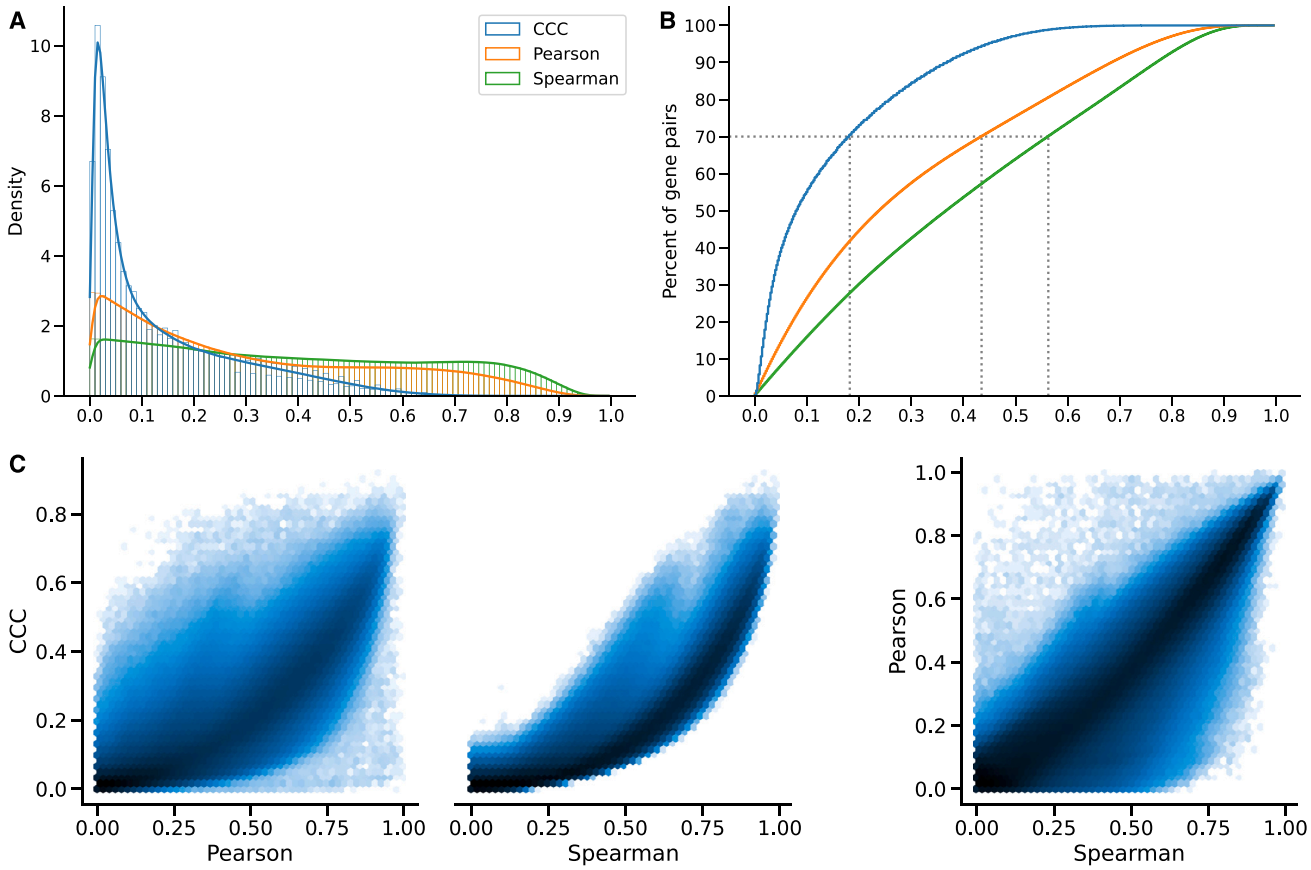


Figure 2. Distribution of coefficient values on gene expression (GTEx v8, whole blood)

(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.

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. For our initial analyses, we selected the top 5,000 genes with the largest variance on whole blood and then computed the correlation matrix between genes using Pearson, Spearman, and CCC (see STAR Methods). Although we always considered the statistical significance of the coefficients, we focused on the strength of the association (i.e., the coefficient value) for our analyses.

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 2B), $c = 0.18$, $p = 0.44$, 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 than either was to Pearson (Figure 2C). As shown later, we also compared with the MIC, another advanced, not-only-linear correlation coefficient that has been successfully used in various application domains.[4,25,26] We found that CCC behaved very similarly to MIC, although CCC was up to two orders of magnitude faster to run. These results suggest that our findings for CCC generalize to MIC; therefore, in the subsequent analyses, we focus on CCC and linear-only coefficients.

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 (see "All pairwise gene correlations in GTEx whole blood" and "Coefficients percentiles in GTEx whole blood" in the key resources table). For most cases (76.4%), an UpSet analysis[35] (Figure 3A) 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

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

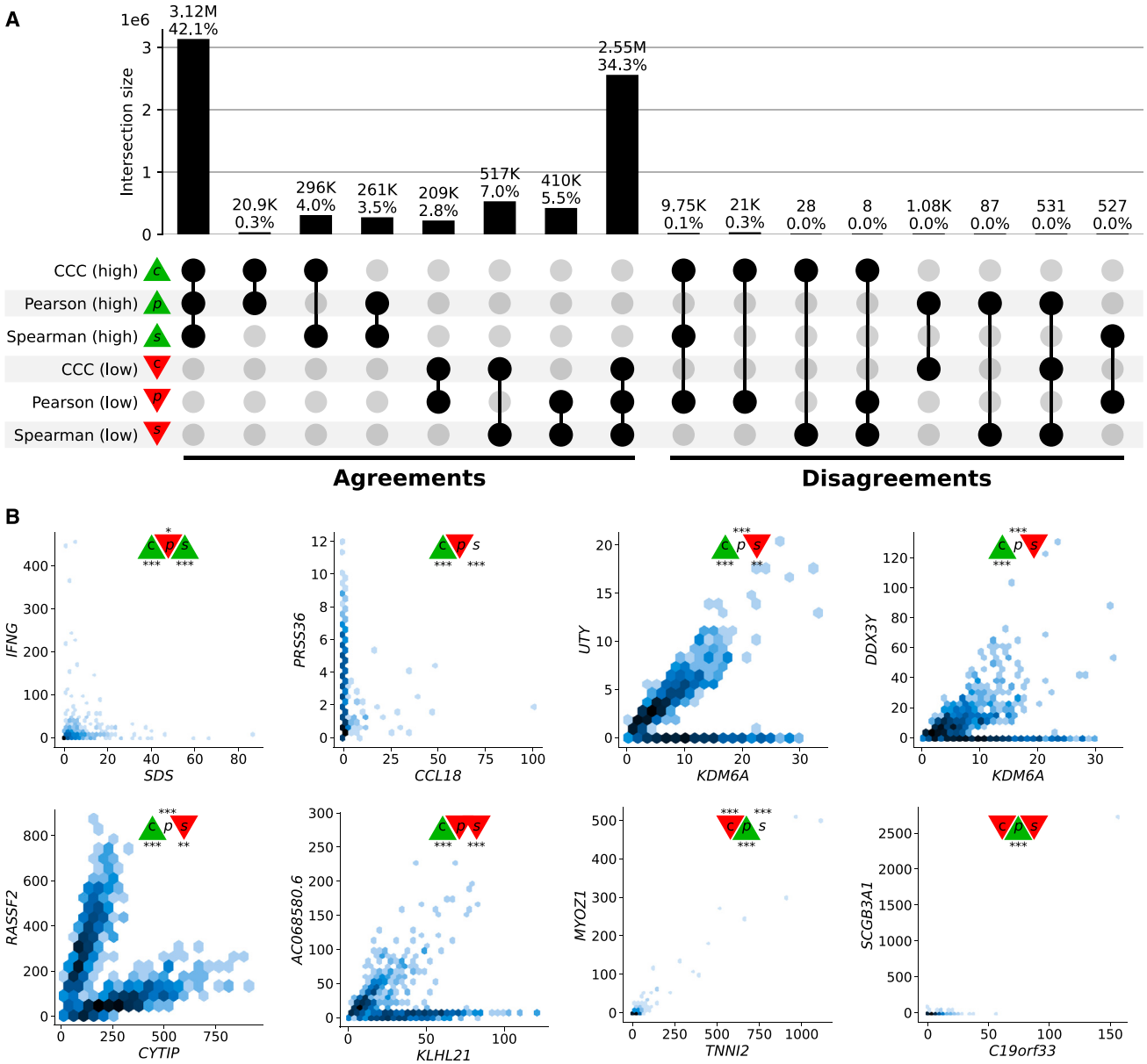


Figure 3. Intersection of gene pairs with high and low correlation coefficient values (GTEx v8, whole blood)

(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). The statistical significance is indicated with asterisks using the false discovery rate (FDR) adjusted *p* values, calculated using the Benjamini and Hochberg method[36]: $FDR < 0.05$ (*), $FDR < 0.01$ (**), and $FDR < 0.001$ (***). A logarithmic scale was used to color each hexagon.

summary, CCC agrees with either Pearson or Spearman in 90.5% of gene pairs by assigning a high or a low correlation value.

While there was broad agreement, more than 20,000 gene pairs with a high CCC value were not highly ranked by the other coefficients (“disagreements” group on the right of Figure 3A). 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). No gene pairs were found to have a high Spearman value and a low CCC. Considering the correlation values and their statistical significance, we analyzed gene pairs among the top ten of each intersection in the disagreements group (Figure 3A, right) where CCC disagrees with Pearson, Spearman, or both.

The first two gene pairs at the top of Figure 3B (*IFNG-SDS*, with high CCC and Spearman, and low Pearson; *PRSS36-CCL18*, with high CCC and low Pearson) appear to follow a non-coexistence relationship: in samples where one of the genes is highly expressed, the other is slightly activated, suggesting a potentially

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

inhibiting effect. The following four gene pairs (*UTY-KDM6A*, *DDX3Y-KDM6A*, *RASSF2-CYTIP*, and *AC068580.6-KLHL21*) follow patterns combining either two linear or one linear and one independent relationship. In particular, genes *UTY-KDM6A* (paralogs) and *DDX3Y-KDM6A* 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. The relationships in these two gene pairs are explained by sex differences in expression: *UTY* and *DDX3Y* are in chromosome Y (Yq11), whereas *KDM6A* is in chromosome X (Xp11), and therefore samples with a linear pattern are males, whereas those with no expression for *UTY* or *DDX3Y* are females. Furthermore, for this sex-specific gene pair pattern, CCC yields a statistically significant coefficient value in 45 out of 47 tissues in GTEx, except for female-specific organs (Figures 4 and S1; see "Correlations for *UTY - KDM6A* and *DDX3Y - KDM6A* in all GTEx tissues" in the key resources table). The gene pair *RASSF2-CYTIP* was replicated in an independent dataset as we explain later. Even though we have not found a biological explanation for gene pair *AC068580.6-KLHL21* (there is limited information about *AC068580.6*, ENSG00000235027, a long non-coding RNA), its strong nonlinear connection with *KLHL21* (linked with some cancers[37]) is robustly captured by CCC only. Notably, these four gene pairs contain strong linear relationships, and CCC is the only coefficient able to consistently capture these nonlinear patterns across a variety of tissues with a statistically significant and high correlation value. Pearson and Spearman show a statistically significant correlation value for some of these gene pairs, although these values are low and would very likely not be prioritized for further research. In addition, these two linear-only coefficients are unable to robustly capture the same pattern in other tissues (Figures 4 and S1; see "Correlations for *UTY - KDM6A* and *DDX3Y - KDM6A* in all GTEx tissues" in the key resources table). For instance, although the three coefficients are statistically significant in whole blood for the gene pair *UTY-KDM6A*, Pearson and Spearman fail to capture the same pattern in the brain cerebellum, and in many cases, such as small intestine, the sign of the coefficient is negative despite the strong positive linear correlation among male samples (Figure 4).

Finally, the last two gene pairs in Figure 3B are highly ranked by Pearson but not by CCC or Spearman. Although all coefficients are significant for the gene pair *MYOZ1-TNNI2*, the low CCC ($c = 0.03$) and moderate Spearman ($s = 0.28$) contrast with Pearson's ($p = 0.97$), suggesting a statistically significant but very weak linear relationship. The high and statistically significant Pearson value for *SCGB3A1-C19orf33* seems to be driven by outliers.

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. Therefore, we used 144 tissue-specific gene networks from the GIANT project,[7,38] where nodes represent genes and each edge a functional relationship weighted with a probability of interaction between two genes (see STAR Methods). The version of GIANT used in this study did not include GTEx samples,[39] 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[40]). We reasoned that statistically significant and 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 Ju et al.[41] and described in STAR Methods).

As an example of our evaluation procedure, we obtained the networks in blood and the automatically predicted cell type for gene pairs *RASSF2-CYTIP* (strong nonlinear pattern and CCC high, Figure 5A) and *MYOZ1-TNNI2* (weak linear pattern and Pearson high, Figure 5B). 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 5A), 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 0.69 and an average of 0.77 and 0.85, respectively (Table S1). The autodetected cell type for this pair was leukocytes, and interaction scores were similar to the blood network (Table S1). 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 5B), 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.

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 STAR Methods). The top 5 predicted cell types for gene pairs highly ranked by CCC and not by the rest were all blood-specific (Figure 5C, 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 5C,

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

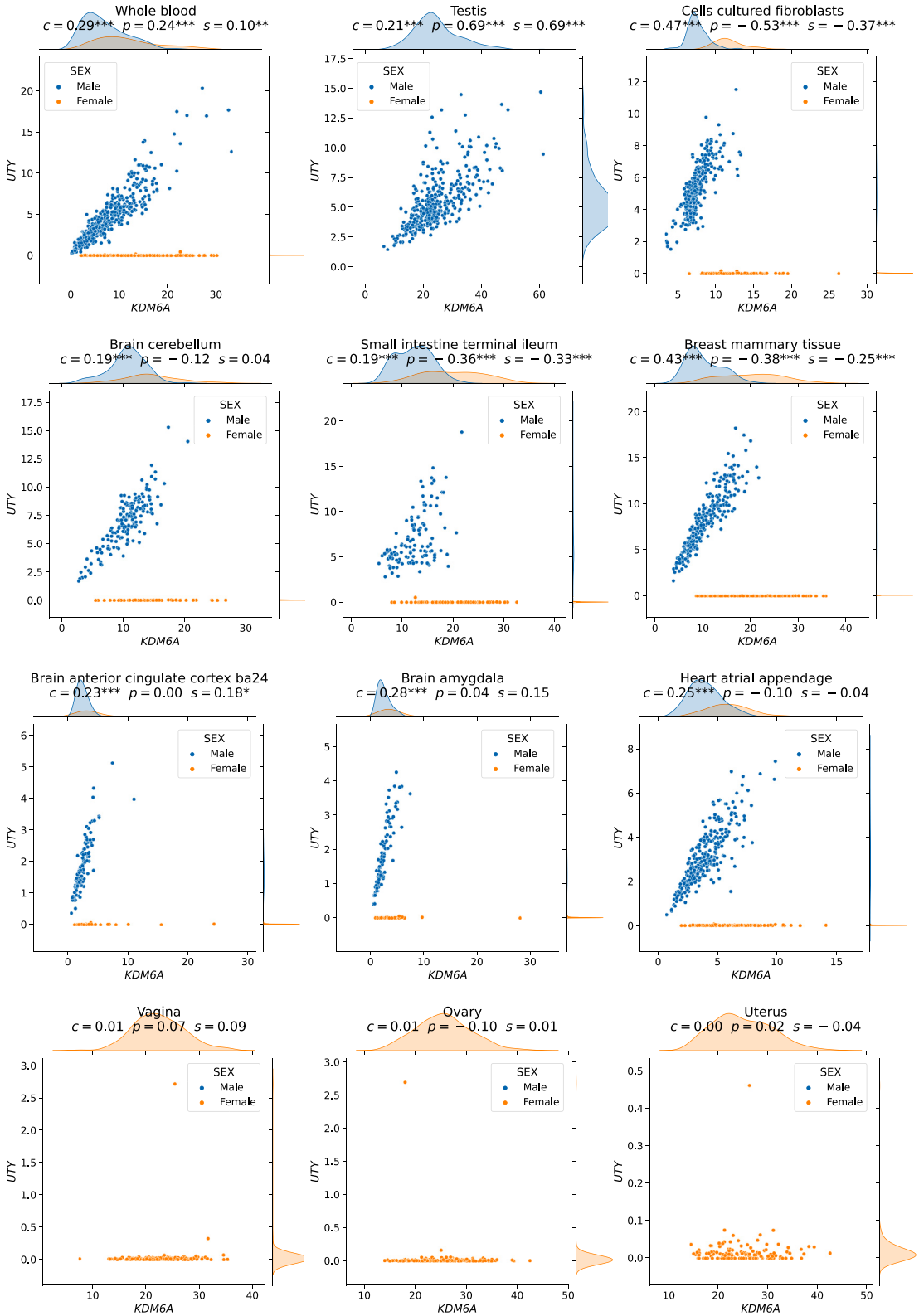


Figure 4. The expression levels of KDM6A and UTY display sex-specific associations across GTEx tissues

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).

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

CellPress
OPEN ACCESS

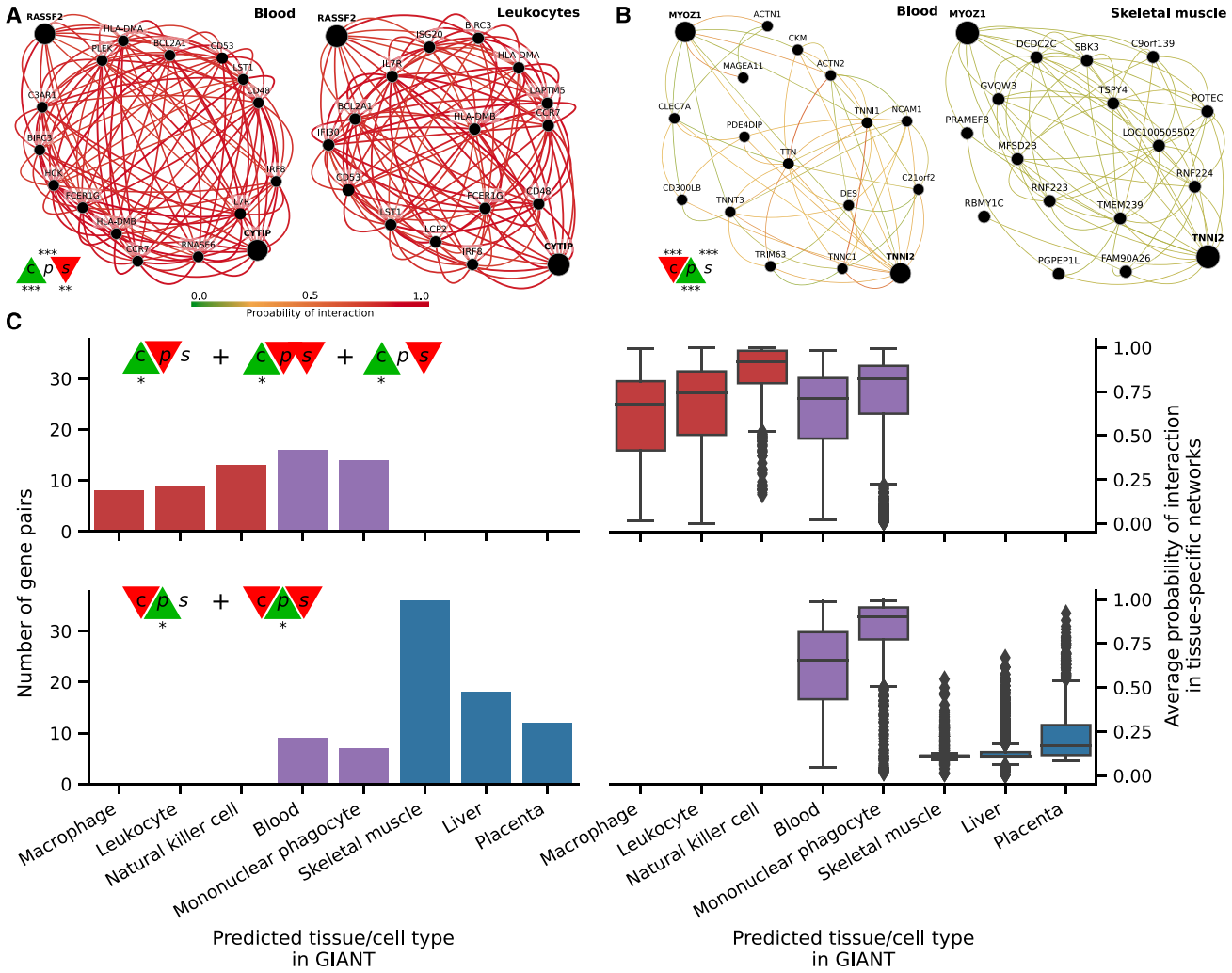


Figure 5. Analysis of GIANT tissue-specific predicted networks for gene pairs prioritized by correlation coefficients

(A and B) Two gene pairs prioritized by correlation coefficients (from Figure 3B) with their predicted networks in blood (left) and an automatically selected tissue/cell type (right) using the method described in Zhang et al.[41] A node represents a gene, and an edge represents 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*,[42] *MYOZ1-TNNI2*.[43]

(C) Summary of predicted tissue/cell type networks for gene pairs exclusively prioritized by CCC and Pearson (FDR < 0.05). The first row combines all gene pairs where CCC is high, and Pearson and Spearman are not. The second row combines all gene pairs where Pearson is high, and CCC and Spearman are not. Bar plots (left) show the number of gene pairs for each predicted tissue/cell type. Boxplots (right) show the average probability of interaction between genes in these predicted tissue-specific networks. Red indicates CCC-only tissues/cell types, blue is Pearson-only, and purple is shared.

top right), with all medians larger than 67% and first quartiles above 41% across predicted cell types. By contrast, most Pearson's gene pairs were predicted to be specific to tissues unrelated to blood (Figure 5C, 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 5C, 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 higher 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 much lower. This lack of replication in GIANT suggests that top Pearson-exclusive-ranked gene pairs in GTEx might be driven mainly by outliers, which is consistent with our earlier observations of outlier-driven associations (Figure 3B).

CCC and the MIC

Conceptual and statistical differences

The CCC and the MIC[23] are measures designed to capture nonlinear relationships between variables. While they share

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

certain similarities, there are also notable differences between them.

Conceptually, CCC is grounded in clustering input data using each variable separately. This process effectively transforms each variable into a set of partitions, each containing a different number of clusters. The CCC then quantifies the correlation between variables by assessing the similarity of these partitions. This allows to process of various types of variables, including both numerical and categorical variables, even when the categories are nominal (i.e., they lack intrinsic order), as explained in STAR Methods. MIC, however, is specifically designed for numerical variables. Additionally, in theory, CCC should also support correlating variables with different dimensions. For one-dimensional variables (such as genes), CCC obtains partitions using a quantile-based approach. For multidimensional variables, CCC could potentially use a standard clustering algorithm (such as *k*-means) to obtain partitions.

Now, consider two variables with n data points on a scatterplot. We can overlay a grid on this scatterplot with x columns and y rows, where each cell of this grid contains a portion of the data points, thereby defining a bivariate probability distribution. The MIC algorithm seeks an optimal grid configuration that maximizes the ratio of mutual information to $logmin\{x,y\}$, subject to the constraint that $xy < n^{0.6}$. This normalization process using $logmin\{x,y\}$ scales the MIC score between zero and one. The CCC, as defined in STAR Methods, also generates a symmetric, normalized score between zero and one. However, unlike MIC, which utilizes normalized mutual information, CCC employs the ARI. The ARI has an advantageous property: it consistently returns a baseline (zero) for independently drawn partitions, irrespective of the number of clusters (see Figure S2). This property is not inherent in mutual information, which can produce varied values for independent variables if the grid dimensions vary. MIC mitigates this by limiting the grid size with the constraint $xy < n^{0.6}$, which could also limit its ability to detect complex relationships.

Both CCC and MIC involve binning the input data vectors, aiming to maximize the mutual information and the ARI, respectively. However, their approaches differ in complexity and execution. MIC utilizes a sophisticated dynamic programming algorithm to identify the optimal grid. By contrast, CCC employs a more straightforward and faster method, partitioning the data points separately using the two vectors. While CCC might benefit from adopting MIC's more complex grid search approach, it remains uncertain if MIC could maintain its performance using CCC's simpler partitioning strategy.

Regarding their parameters, CCC's k_{max} (maximum number of clusters) and MIC's $B(n)$ (maximum grid size) serve similar purposes. They control both the complexity of the patterns detected and the computational time. For example, as illustrated in Figure 1 (Anscombe I and III), a k_{max} of 2 is adequate for identifying linear patterns but insufficient for more complex patterns like quadratic or two-lines patterns. A similar principle applies to MIC's $B(n)$. However, a critical distinction exists between the two: the constant baseline property of ARIs ensures that CCC returns a value close to zero for independent variables, regardless of k_{max}. By contrast, MIC may produce non-zero scores for independent data if $B(n)$ is set too high, as discussed in Section 2.2.1 of the supplementary material in Reshef et al.[23] The authors of MIC suggest that a value of $B(n) = n^{0.6}$ is generally effective in practice.

Comparison in gene expression data

We compared all the coefficients in this study with MIC, a popular nonlinear method that can find complex relationships in data, although it is very computationally intensive.[25] We ran MICe (see STAR Methods) on all possible pairwise comparisons of our 5,000 highly variable genes from whole blood in GTEx v8. 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 6C shows that these two coefficients relate almost linearly, and both compare very similarly with Pearson and Spearman.

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.[25,27,44–46] The original MIC implementation uses ApproxMaxMI, a computationally demanding heuristic estimator.[23] Recently, a more efficient implementation called MICe was proposed.[47] These two MIC estimators are provided by the minepy package,[44] 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 STAR Methods), so we also tested the cases using 1 and 3 central processing unit (CPU) cores. Figure 7 shows the time in seconds in log scale.

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[51], 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.

DISCUSSION

We introduce the CCC, an efficient, not-only-linear clustering-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 potentially complex relationships where a subset of samples or conditions are

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

CellPress
OPEN ACCESS

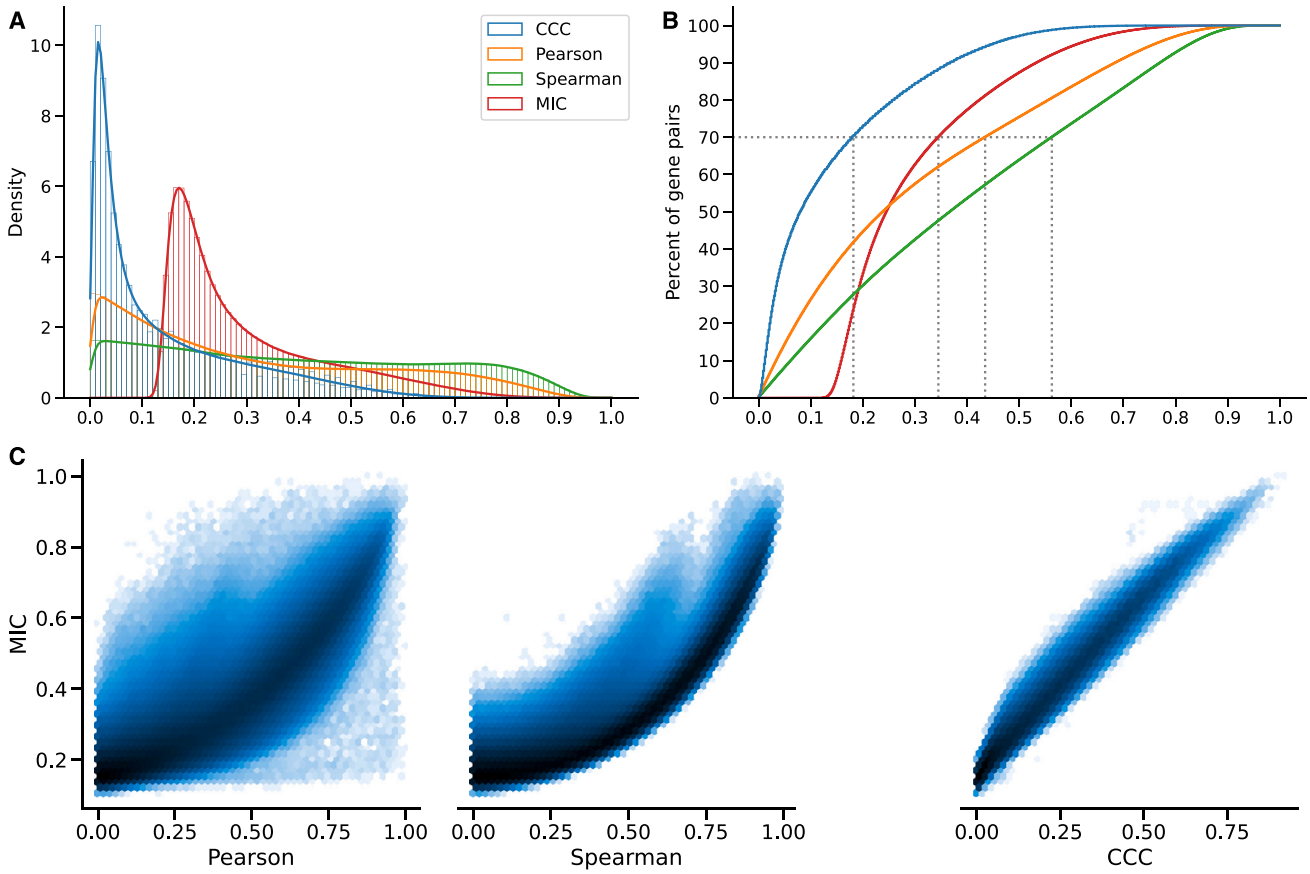


Figure 6. Distribution of MIC values on gene expression (GTEx v8, whole blood) and comparison with other methods

(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.

explained by other factors (such as differences between health and disease). We found that the 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 and standard coefficients 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, not-only-linear correlation coefficients that support categorical variables 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 toward a small fraction of human genes.[51,52] Some genes highlighted in CCC-ranked pairs (Figure 3B), such as *SDS* (12q24) or *PRSS36* (16p11), 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 co-expression patterns. A beyond-linear gene co-expression analysis on large compendia might shed light on the

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

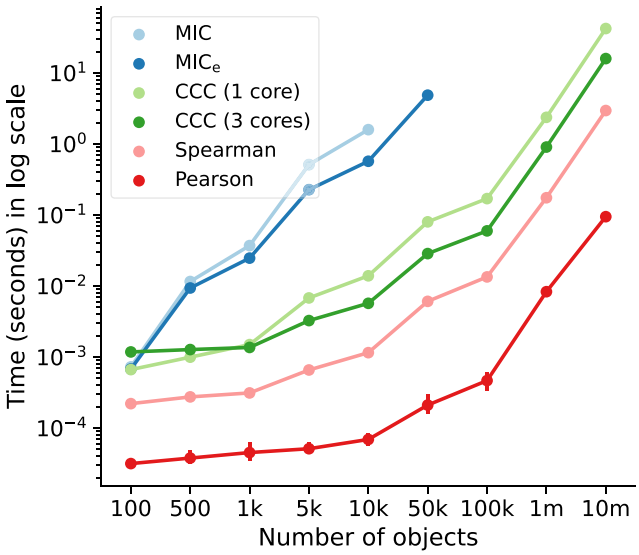


Figure 7. Computational complexity of all correlation coefficients on simulated data

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 (1,000 repetitions were performed for data size 100). CCC was run using 1 and 3 CPU cores. MIC and MICe did not finish running in a reasonable amount of time for data sizes of 10,000 and 100,000, respectively.

function of understudied genes. For example, genes *KLHL21* (1p36) and *AC068580.6* (a long non-coding RNA gene 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[37] and other cancers.[54,55] 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 3B).

Not-only-linear correlation coefficients might also be helpful in the field of genetic studies. In this context, GWASs have been successful in understanding the molecular basis of common diseases by estimating the association between genotype and phenotype.[56] However, the estimated effect sizes of genes identified with GWASs are generally modest, and they explain only a fraction of the phenotype variance, hampering the clinical translation of these findings.[57] 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[19,20,22] have shown that integrating gene co-expression networks with genetic studies could potentially identify core genes that are missed by linear-only models alone like GWASs. Our results suggest that building these networks with the latest approaches[58] and 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 in a single tissue from GTEx 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 are 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. We compute p values using computationally intensive permutation tests; in the future, we plan to explore efficient permutation approaches such as those based on extreme value theory.[59] 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 into how to adjust it for different applications or data types. Finally, computing the correlation between a gene pair represents only the first step of the analysis. Controlling for known confounders, integrating with other data types, and replicating in independent datasets are some of the other important steps to ensure the biological relevance of the detected patterns.

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 SABV 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.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Milton Pividori (milton.pividori@cuanschutz.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at https://github.com/greenelab/ccc and Zenodo: 13304247, and is publicly available as of the date of publication. DOIs are listed in the key resources table.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work has been supported by the Gordon and Betty Moore Foundation (GBMF 4552 to C.S.G.), the National Human Genome Research Institute (R01 HG010067 to C.S.G., K99 HG011898 and R00 HG011898 to M.P.), the National Cancer Institute (R01 CA237170 to C.S.G.), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 HD109765 to C.S.G.).

AUTHOR CONTRIBUTIONS

C.S.G. and D.H.M. supervised the project. M.P. conceived the study, developed the methodology, performed the analysis, and drafted the manuscript. C.S.G., D.H.M., and M.D.R. provided critical feedback about the methodology, analyses performed, and manuscript writing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data preprocessing
  - The Clustermatch Correlation Coefficient (CCC)
  - Tissue-specific network analyses using GIANT
  - Maximal Information Coefficient (MIC)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2024.08.005.

Received: September 9, 2022
Revised: June 18, 2024
Accepted: August 15, 2024
Published: September 6, 2024

REFERENCES

1. Hanson, B., Sugden, A., and Alberts, B. (2011). Making data maximally available. Science *331*, 649. https://doi.org/10.1126/science.1203354.
2. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. *19*, 1454–1462. https://doi.org/10.1038/nn.4353.
3. Kong, J., Klein, B.E.K., Klein, R., Lee, K.E., and Wahba, G. (2012). Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. Proc. Natl. Acad. Sci. USA *109*, 20352–20357. https://doi.org/10.1073/pnas.1217269109.
4. Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., Wang, G., and Zhou, F. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. BMC Bioinformatics *17*, 142. https://doi.org/10.1186/s12859-016-0990-0.
5. Song, X.-F., Zhang, Y., Gong, D.-W., and Gao, X.-Z. (2022). A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data. IEEE Trans. Cybern. *52*, 9573–9586. https://doi.org/10.1109/tcyb.2021.3061152.
6. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309. https://doi.org/10.1016/j.cell.2011.01.004.
7. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576. https://doi.org/10.1038/ng.3259.
8. Ficklin, S.P., and Feltus, F.A. (2011). Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol. *156*, 1244–1256. https://doi.org/10.1104/pp.111.173047.
9. Tsaparas, P., Mariño-Ramírez, L., Bodenreider, O., Koonin, E.V., and Jordan, I.K. (2006). Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol. Biol. *6*, 70. https://doi.org/10.1186/1471-2148-6-70.
10. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
11. Wilks, C., Zheng, S.C., Chen, F.Y., Charles, R., Solomon, B., Ling, J.P., Imada, E.L., Zhang, D., Joseph, L., Leek, J.T., et al. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biol. *22*, 323. https://doi.org/10.1186/s13059-021-02533-6.
12. Taroni, J.N., Grayson, P.C., Hu, Q., Eddy, S., Kretzler, M., Merkel, P.A., and Greene, C.S. (2019). MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Syst. *8*, 380–394.e4. https://doi.org/10.1016/j.cels.2019.04.003.
13. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. *15*, e1007889. https://doi.org/10.1371/journal.pgen.1007889.
14. Yao, D.W., O'Connor, L.J., Price, A.L., and Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat. Genet. *52*, 626–633. https://doi.org/10.1038/s41588-020-0625-2.
15. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.
16. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.
17. Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell *177*, 1022–1034.e6. https://doi.org/10.1016/j.cell.2019.04.014.
18. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and Regev, A. (2021). Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics. Nat. Genet. *54*, 1479–1492. https://doi.org/10.1038/s41588-022-01187-9.
19. Pividori, M., Lu, S., Li, B., Su, C., Johnson, M.E., Wei, W.-Q., Feng, Q., Namjou, B., Kiryluk, K., Kullo, I.J., et al. (2023). Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms. Nat. Commun. *14*, 5562. https://doi.org/10.1038/s41467-023-41057-4.
20. Bakker, O.B., Claringbould, A., Westra, H.-J., Wiersma, H., Boulogne, F., Võsa, U., Symmons, S.M., Jonkers, I.H., Franke, L., and Deelen, P. (2021). Linking common and rare disease genetics through gene regulatory networks. Preprint at medRxiv. https://doi.org/10.1101/2021.10.21.21265342.
21. Mathieson, I. (2021). The omnigenic model and polygenic prediction of complex traits. Am. J. Hum. Genet. *108*, 1558–1563. https://doi.org/10.1016/j.ajhg.2021.07.003.

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

22. Lee, H.-C., Ichikawa, O., Glicksberg, B.S., Divaraniya, A.A., Becker, C.E., Agarwal, P., and Dudley, J.T. (2020). Identification of therapeutic targets from genetic association studies using hierarchical component analysis. BioData Min. *13*, 6. https://doi.org/10.1186/s13040-020-00216-9.

23. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. (2011). Detecting novel associations in large data sets. Science *334*, 1518–1524. https://doi.org/10.1126/science.1205438.

24. Székely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. Ann. Statist. *35*, 2769–2794. https://doi.org/10.1214/009053607000000505.

25. Cao, D., Chen, Y., Chen, J., Zhang, H., and Yuan, Z. (2021). An improved algorithm for the maximal information coefficient and its application. R. Soc. Open Sci. *8*, 201424. https://doi.org/10.1098/rsos.201424.

26. Liang, T., Zhang, Q., Liu, X., Lou, C., Liu, X., and Wang, H. (2020). Time-Frequency Maximal Information Coefficient Method and its Application to Functional Corticomuscular Coupling. IEEE Trans. Neural Syst. Rehabil. Eng. *28*, 2515–2524. https://doi.org/10.1109/tnsre.2020.3028199.

27. Chen, Y., Zeng, Y., Luo, F., and Yuan, Z. (2016). A New Algorithm to Optimize Maximal Information Coefficient. PLoS One *11*, e0157567. https://doi.org/10.1371/journal.pone.0157567.

28. Pividori, M., Cernadas, A., de Haro, L.A., Carrari, F., Stegmayer, G., and Milone, D.H. (2019). Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization. Bioinformatics *35*, 1931–1939. https://doi.org/10.1093/bioinformatics/bty899.

29. Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218. https://doi.org/10.1007/bf01908075.

30. Anscombe, F.J. (1973). Graphs in Statistical Analysis. Am. Stat. *27*, 17–21. https://doi.org/10.1080/00031305.1973.10478966.

31. Matejka, J., and Fitzmaurice, G. (2017). Same Stats, Different Graphs. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (ACM), pp. 1290–1294. https://doi.org/10.1145/3025453.3025912.

32. Murray, L.L., and Wilson, J.G. (2021). Generating data sets for teaching the importance of regression analysis. Decis. Sci. J. Innov. Edu. *19*, 157–166. https://doi.org/10.1111/dsji.12233.

33. Wang, Q., Zhang, H., Liang, Y., Jiang, H., Tan, S., Luo, F., Yuan, Z., and Chen, Y. (2019). A Novel Method to Efficiently Highlight Nonlinearly Expressed Genes. Front. Genet. *10*, 1410. https://doi.org/10.3389/fgene.2019.01410.

34. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. Mol. Biol. Cell *9*, 3273–3297. https://doi.org/10.1091/mbc.9.12.3273.

35. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. *20*, 1983–1992. https://doi.org/10.1109/tvcg.2014.2346248.

36. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

37. Shi, L., Zhang, W., Zou, F., Mei, L., Wu, G., and Teng, Y. (2016). KLHL21, a novel gene that contributes to the progression of hepatocellular carcinoma. BMC Cancer *16*, 815. https://doi.org/10.1186/s12885-016-2851-7.

38. Flatiron Institute. HumanBase. data-driven predictions of gene function and interactions. https://hb.flatironinstitute.org/.

39. Flatiron Institute. HumanBase Data sources. https://hb.flatironinstitute.org/data.

40. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

41. Ju, W., Greene, C.S., Eichinger, F., Nair, V., Hodgin, J.B., Bitzer, M., Lee, Y.S., Zhu, Q., Kehata, M., Li, M., et al. (2013). Defining cell-type specificity at the transcriptional level in human disease. Genome Res. *23*, 1862–1873. https://doi.org/10.1101/gr.155697.113.

42. Flatiron Institute. RASSF2, CYTIP - HumanBase. https://hb.flatironinstitute.org/gene/9770+9595.

43. Flatiron Institute. MYOZ1, TNNI2 - HumanBase. https://hb.flatironinstitute.org/gene/58529+7136.

44. Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., and Furlanello, C. (2013). minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics *29*, 407–408. https://doi.org/10.1093/bioinformatics/bts707.

45. Tang, D., Wang, M., Zheng, W., and Wang, H. (2014). RapidMic: Rapid Computation of the Maximal Information Coefficient. Evol. Bioinform. Online *10*, 11–16. https://doi.org/10.4137/ebo.s13121.

46. Zhang, Y., Jia, S., Huang, H., Qiu, J., and Zhou, C. (2014). A Novel Algorithm for the Precise Calculation of the Maximal Information Coefficient. Sci. Rep. *4*, 6662. https://doi.org/10.1038/srep06662.

47. Reshef, Y., Reshef, D., Finucane, H., Sabeti, P., and Mitzenmacher, M. (2016). Measuring Dependence Powerfully and Equitably. J. Mach. Learn. Res. *17*, 1–63.

48. Clayton, J.A., and Collins, F.S. (2014). Policy: NIH to balance sex in cell and animal studies. Nature *509*, 282–283. https://doi.org/10.1038/509282a.

49. Bhargava, A., Arnold, A.P., Bangasser, D.A., Denton, K.M., Gupta, A., Hilliard Krause, L.M., Mayer, E.A., McCarthy, M., Miller, W.L., Raznahan, A., et al. (2021). Considering Sex as a Biological Variable in Basic and Clinical Studies: An Endocrine Society Scientific Statement. Endocr. Rev. *42*, 219–258. https://doi.org/10.1210/endrev/bnaa034.

50. Shansky, R.M., and Murphy, A.Z. (2021). Considering sex as a biological variable will require a global shift in science culture. Nat. Neurosci. *24*, 457–464. https://doi.org/10.1038/s41593-021-00806-8.

51. Pfeiffer, T., and Hoffmann, R. (2007). Temporal patterns of genes in scientific publications. Proc. Natl. Acad. Sci. USA *104*, 12052–12056. https://doi.org/10.1073/pnas.0701315104.

52. Su, A.I., and Hogenesch, J.B. (2007). Power-law-like distributions in biomedical publications and research funding. Genome Biol. *8*, 404. https://doi.org/10.1186/gb-2007-8-4-404.

53. Stoeger, T., Gerlach, M., Morimoto, R.I., and Nunes Amaral, L.A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. *16*, e2006643. https://doi.org/10.1371/journal.pbio.2006643.

54. Chen, J., Song, W., Du, Y., Li, Z., Xuan, Z., Zhao, L., Chen, J., Zhao, Y., Tuo, B., Zheng, S., et al. (2018). Inhibition of KLHL21 prevents cholangiocarcinoma progression through regulating cell proliferation and motility, arresting cell cycle and reducing Erk activation. Biochem. Biophys. Res. Commun. *499*, 433–440. https://doi.org/10.1016/j.bbrc.2018.03.152.

55. Li, C., Li, R., Hu, X., Zhou, G., and Jiang, G. (2022). Tumor-promoting mechanisms of macrophage-derived extracellular vesicles-enclosed microRNA-660 in breast cancer progression. Breast Cancer Res. Treat. *192*, 353–368. https://doi.org/10.1007/s10549-021-06433-y.

56. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

57. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.

58. Burns, J.J.R., Shealy, B.T., Greer, M.S., Hadish, J.A., McGowan, M.T., Biggs, T., Smith, M.C., Feltus, F.A., and Ficklin, S.P. (2022). Addressing noise in co-expression network construction. Brief. Bioinform. *23*, bbab495. https://doi.org/10.1093/bib/bbab495.

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

59. Knijnenburg, T.A., Wessels, L.F.A., Reinders, M.J.T., and Shmulevich, I. (2009). Fewer permutations, more accurate *P*-values. Bioinformatics *25*, i161–i168. https://doi.org/10.1093/bioinformatics/btp211.

60. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

61. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

62. Vinh, N.X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. J. Mach. Learn. Res. *11*, 2837–2854.

63. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

64. Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., et al. (2013). The BioGRID interaction database: 2013 update. Nucleic Acids Res. *41*, D816–D823. https://doi.org/10.1093/nar/gks1158.

65. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. *40*, D841–D846. https://doi.org/10.1093/nar/gkr1088.

66. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857–D861. https://doi.org/10.1093/nar/gkr930.

67. Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1999). MIPS: a database for genomes and protein sequences. Nucleic Acids Res. *27*, 44–48. https://doi.org/10.1093/nar/27.1.44.

68. Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*, D105–D110. https://doi.org/10.1093/nar/gkp950.

69. Huttenhower, C., Schroeder, M., Chikina, M.D., and Troyanskaya, O.G. (2008). The Sleipnir library for computational functional genomics. Bioinformatics *24*, 1559–1561. https://doi.org/10.1093/bioinformatics/btn237.

70. GitHub (2024). minepy - Maximal Information-based Nonparametric Exploration (MINE) in C and Python (GitHub). https://github.com/minepy.

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GTEx v8	The GTEx Consortium[10]	https://gtexportal.org/
All pairwise gene correlations in GTEx whole blood	This paper	https://doi.org/10.5281/zenodo.10472273
Coefficients percentiles in GTEx whole blood	This paper	https://doi.org/10.5281/zenodo.10472273
Correlations for *UTY - KDM6A* and *DDX3Y - KDM6A* in all GTEx tissues	This paper	https://doi.org/10.5281/zenodo.10472273
Software and algorithms		
Python 3.9.16	Python Software Foundation	https://python.org/
scikit-learn-0.24.2	Pedregosa et al.[60]	https://github.com/scikit-learn/scikit-learn
SciPy	Virtanen et al.[61]	https://github.com/scipy/scipy
MIC	Reshef et al.[47]	https://github.com/minepy/minepy
GIANT	Greene et al.[7]	https://hb.flatironinstitute.org/
CCC	This paper	GitHub: https://github.com/greenelab/ccc or Zenodo: https://doi.org/10.5281/zenodo.13304247

METHOD DETAILS

Data preprocessing

We downloaded gene expression data from GTEx v8 (https://gtexportal.org/) 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 toward highly expressed genes. We then computed Pearson,[60,61] Spearman,[60,61] the Maximal Information Coefficient (MIC)[47] and CCC on these 5,000 genes across all 755 samples, generating a pairwise similarity matrix of size 5,000 x 5,000.

The Clustermatch Correlation Coefficient (CCC)

Definitions

Definition 1.1. Given a data vector $x = (x_1, x_2, \ldots, x_n) \in R^n$ then define

$$\pi_\ell = \{i | \rho_\ell < x_i \leq \rho_{\ell+1}\}, \forall \ell \in [1, k]$$

as a *partition* of the n objects of x into $|\pi| = k$ clusters, where ρ is a set of $k+1$ cutpoints (e.g., quantiles) that define the clusters, with $\rho_1 = min(x)$ and $\rho_{k+1} = max(x)$. If x is a categorical vector with no intrinsic ordering, then a partition is defined as

$$\pi_\ell = \{i | x_i = C_\ell\}, \forall \ell \in [1, |C|]$$

where $C = \{c_1, c_2, \ldots, c_m\}$ is a set of unique values in x corresponding to the $m = |C|$ categorical values that define the clusters.

Definition 1.2. Given two partitions π and π' of n objects, the *adjusted Rand Index (ARI)*[29] is given by

$$ARI(\pi, \pi') = \frac{2(n_0 n_1 - n_2 n_3)}{(n_0+n_2)(n_2+n_1)+(n_0+n_3)(n_3+n_1)},$$

where n_0 is the number of object pairs that are in the same cluster in both partitions π and π', n_1 is the number of object pairs that are in different clusters, n_2 is the number of object pairs that are in the same cluster in π but in different clusters in π', and n_3 is the number of object pairs that are in different clusters in π but in the same cluster in π'. Intuitively, $n_0 + n_1$ reflects the number of object pairs where both partitions agree, and $n_2 + n_3$ are those in which they disagree.

Definition 1.3. The *Clustermatch Correlation Coefficient (CCC)* between x and y is defined as the maximum ARI between all possible partitions of x and y

$$CCC(x, y) = max\left\{0, \max_{\pi_j \in \Pi^x} \max_{\pi_l \in \Pi^y} \{ARI(\pi_j, \pi_l)\}\right\}, \forall |\pi| \in [2, k_{max}]$$

where Π^x is a set of partitions derived from x, Π^y is a set of partitions derived from y, and k_{max} specifies the maximum number of clusters allowed. The ARI has an upper bound of 1 (achieved when both partitions are identical), and although it does not have a

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

well-defined lower bound, values equal or less than zero are achieved when partitions are independent. Therefore, $CCC(x,y) \in [0,1]$. In the special case where all n objects in either x or y have the same value, the CCC is undefined.

The CCC has the following basic properties derived from the ARI: 1) symmetry, since $ARI(\pi,\pi') = ARI(\pi',\pi)$; 2) normalization, since it takes a minimum value of zero and a maximum of one since $ARI(\pi,\pi) = 1$; 3) constant baseline, since the ARI is adjusted-for-chance,[29] it returns a value close to zero for independently drawn partitions, and this also holds when partitions have different number of clusters.[62] This is an important property, since CCC compares partitions with different numbers of clusters, and relationships between two variables (such as linear or quadratic) might be better represented with different numbers of clusters as shown in Figure 1.

The maximum number of clusters k_{max}

The parameter k_{max} is the maximum number of clusters k allowed for any partition derived from x or y. On one hand, note that the same value of k might not be the right one to find a relationship between any two variables. For instance, in the quadratic example in Figure 1, CCC returns a value of 0.36 (grouping objects in four clusters using one variable and two using the other). If we used only two clusters instead, CCC would return a similarity value of 0.02. On the other hand, computational time increases quadratically with k_{max}. In addition, it is important to note that given the constant baseline property of the ARI, the CCC returns a value close to zero for independent variables regardless of the value of k_{max}. As shown in Figure S2, this holds even for very large values of k_{max}, approaching the number of objects n. Note that as k_{max} approaches n, the number of singleton clusters (i.e., clusters with only one object) increases, which would not be useful for finding relationships between variables. Therefore, given the constant baseline property, k_{max} only represents a tradeoff between the ability to capture complex patterns and the computational cost, with random/independent variables having a CCC value close to zero regardless of the value of k_{max}; we found that $k_{max} = 10$ works well in practice, and it was used as the default maximum number of clusters across all our analyses.

Statistical significance

Our null hypothesis is that the variables represented by x and y are independent. To compute a P-value, we perform a set of permutations of values in y and compute the CCC between x and each permuted vector. The P-value is the proportion of CCC values using the permuted data that are greater than or equal to the CCC value between x and y. We used 1 million permutations in all our analyses, and we adjusted the P-values using the Benjamini and Hochberg procedure[36] to control the false discovery rate (FDR); given the computational cost, we computed a P-value only for gene pairs from the "Disagreements" group in Figure 3, which contains gene pairs ranked differently by the correlation coefficients.

Strength of linear correlation

Figure 4 shows the relationships between *UTY* (chromosome Y) and *KDM6A* (chromosome X) across tissues in GTEx. For this gene pair, CCC can find a complex pattern where a subset of samples (males) follows a clear linear relationship, and there is no relationship in the rest of the samples (females). As it can be seen, there is a difference in the strength of the linear correlation between male samples across different tissues. For example, in brain cerebellum, the linear correlation among male samples is stronger than in small intestine (terminal ileum). As shown in Figure S3A, this difference is reflected by all coefficients when only male samples are considered.

However, when we consider all samples (males and females), there is no longer a linear relationship between *UTY* and *KDM6A*. Therefore, while a subset of the data displays linear relationships, overall, it is no longer true that there is a linear correlation. CCC assumes that if two variables (genes in our case) are similar, the clustering of objects (samples) using each variable separately should match. As shown in Figure S3A with red lines, this clustering of samples and their matching can be seen for the gene pair *UTY* / *KDM6A*: when we only consider male samples, CCC finds clusterings in brain cerebellum with a larger matching than in small intestine because the linear strength differs. But when we consider all samples together (males and females, as shown in Figure S3B), the pattern is nonlinear, the distribution of all the data is different, and so are the clusterings found by CCC.

The effect of analyzing *all* the data (males and females) in this nonlinear pattern (Figure 4) is also clear in the negative sign of Pearson and Spearman coefficients in small intestine or even other tissues with a very strong and clear linear pattern among male samples such as breast mammary tissue. This case indicates that Pearson and Spearman, although statistically significant, are capturing the wrong pattern. Therefore, the fact that CCC yields a similar value (0.19) for these nonlinear patterns in brain cerebellum and small intestine (Figure S3B) reflects a similar clustering matching when considering all the samples. When applied only to the data with linear relationships of varying strength, CCC performs consistently with other coefficients.

Presence of substructure in the data

Consider a scenario where there are known and undesirable substructures in the data. In the example in Figure S4, we have simulated two distinct clusters (normally distributed) placed diagonally, horizontally, and vertically. The only case where the CCC is close to 1.0 (Diagonal, left) is when the clusterings/partitions of objects using each variable (x and y) match, which coincides with a linear pattern. In the other two cases (Horizontal and Vertical), clusterings of objects do not match, leading to a CCC value of zero. We note that MIC has the same behavior.

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.[38] 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 Greene et al.[7] Briefly, tissue-specific gene networks were built using gene expression data (without GTEx samples[39]) from the NCBI's Gene Expression Omnibus (GEO),[63] protein-protein

Please cite this article in press as: Pividori et al., An efficient, not-only-linear correlation coefficient based on clustering, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.08.005

interaction (BioGRID,[64] IntAct,[65] MINT[66] and MIPS[67]), transcription factor regulation using binding motifs from JASPAR,[68] and chemical and genetic perturbations from MSigDB.[40] 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[69]) 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 Ju et al.[41] 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[44,70] (version 1.2.5) to estimate the MIC coefficient. In GTEx v8 (whole blood), we used MICe (an improved implementation of the original MIC introduced in Reshef et al.[47]) with the default parameters alpha=0.6, c=15 and estimator='mic_e'. We used the pairwise_distances function from scikit-learn[60] to parallelize the computation of MIC on GTEx. For our computational complexity analyses, we ran the original MIC (using parameter estimator='mic_approx') and MICe (estimator='mic_e').