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

This manuscript (permalink) was automatically generated from greenelab/clustermatch-gene-expr-manuscript@082cf81 on April 21, 2022.

Draft
This manuscript version is work-in-progress

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)

Abstract

Correlation coefficients are essential to identify intriguing relationships in data. In transcriptomics, genes with correlated expression can suggest that they share functions or are part of disease-relevant biological processes. Here we introduce the Clustermatch Correlation Coefficient (CCC), an efficient, easy-to-use and not-only-linear coefficient based on machine learning models. CCC efficiently captures general patterns in data by applying clustering algorithms and automatically adjusting the model's complexity. When applied to human gene expression data, CCC does not miss robust linear relationships while detecting more complex yet biologically meaningful patterns. CCC can detect non-linear patterns associated with sex as a biological variable missed by standard coefficients. Gene pairs highly ranked by CCC in expression data alone showed high probabilities of interaction in tissue-specific networks built from protein-interaction data, transcription factor regulation, and chemical and genetic perturbations. CCC is a highly-efficient, next-generation correlation coefficient that can readily be applied to genome-scale data and various domains.

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 that implement sufficiently complex models to discover underlying patterns. Correlation analysis is an essential statistical technique to discover 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 morpho-agronomic traits in crop plans [3]. Correlation methods are also used in supervised tasks, for example, for feature selection to improve prediction accuracy [4,5]. The Pearson correlation coefficient is ubiquitously deployed across application domains and diverse scientific areas. Even minor and significant improvements in these techniques could have enormous consequences in industry and research.

In transcriptomics, many analyses start with estimating the correlation between genes. More sophisticated approaches built on correlation analysis can suggest gene function [6], aid in discovering common and cell lineage-specific regulatory networks [7], and capture important interactions in a living organism that can uncover molecular mechanisms in other species [8,9]. The analysis of large RNA-seq datasets [10,11] can also reveal complex transcriptional mechanisms underlying human diseases [2,12,13,14,15]. Since the introduction of the omnigenic model of complex traits [16,17], gene-gene relationships are playing an increasingly important role in genetic studies of human diseases [18,19,20,21], even in specific fields such as polygenic risk scores [22]. In this context, recent approaches combine disease-associated genes from genome-wide association studies (GWAS) with gene co-expression networks to prioritize "core" genes directly affecting diseases [19,20,23]. These core genes are not captured by standard statistical methods but are believed to be part of disease-relevant and highly-interconnected regulatory networks. Therefore, more advanced correlation coefficients could dramatically improve the identification of more attractive 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 can only capture linear or monotonic patterns, missing complex yet essential relationships. The Maximal Information Coefficient (MIC) [24] or Distance Correlation (DC) [25] were proposed as metrics that capture non-linear patterns. However, their computational complexity makes them impractical for big data and even moderately sized datasets. We previously developed Clustermatch, a method for cluster analysis on highly diverse datasets that significantly outperformed Pearson, Spearman, MIC and DC in detecting simulated linear

and non-linear relationships with varying levels of noise [3]. 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 balances the complexity of relationships found and computation time. To assess its performance in RNA-seq data, we applied our method to gene expression data from the Genotype-Tissue Expression v8 (GTEx) project across different tissues [26]. CCC captured both strong linear relationships and novel non-linear patterns associated with sex as a biological variable, which were completely missed by standard coefficients. We also show that the CCC is most similar to MIC, although it is much faster to compute. Gene pairs detected in expression data by CCC were found to have higher probabilities of interaction in tissue-specific gene networks from the Genome-wide Analysis of gene Networks in Tissues (GIANT) [27]. Furthermore, its ability to efficiently handle diverse data types (including numerical and categorical features) reduces preprocessing steps and makes it appealing to analyze large and heterogeneous repositories.

Results

A robust and efficient not-only-linear dependence coefficient

The Clustermatch Correlation Coefficient (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 partitioning of those n objects derived by clustering each individual variable should match. In the case of numerical values, CCC uses quantiles to efficiently separate data points into different clusters (e.g., the median separates numerical data into two clusters). Once all partitions are generated according to each variable, CCC is defined as the maximum adjusted Rand index (ARI) [28] between them, ranging between 0 and 1. Details of the CCC algorithm can be found in Methods.

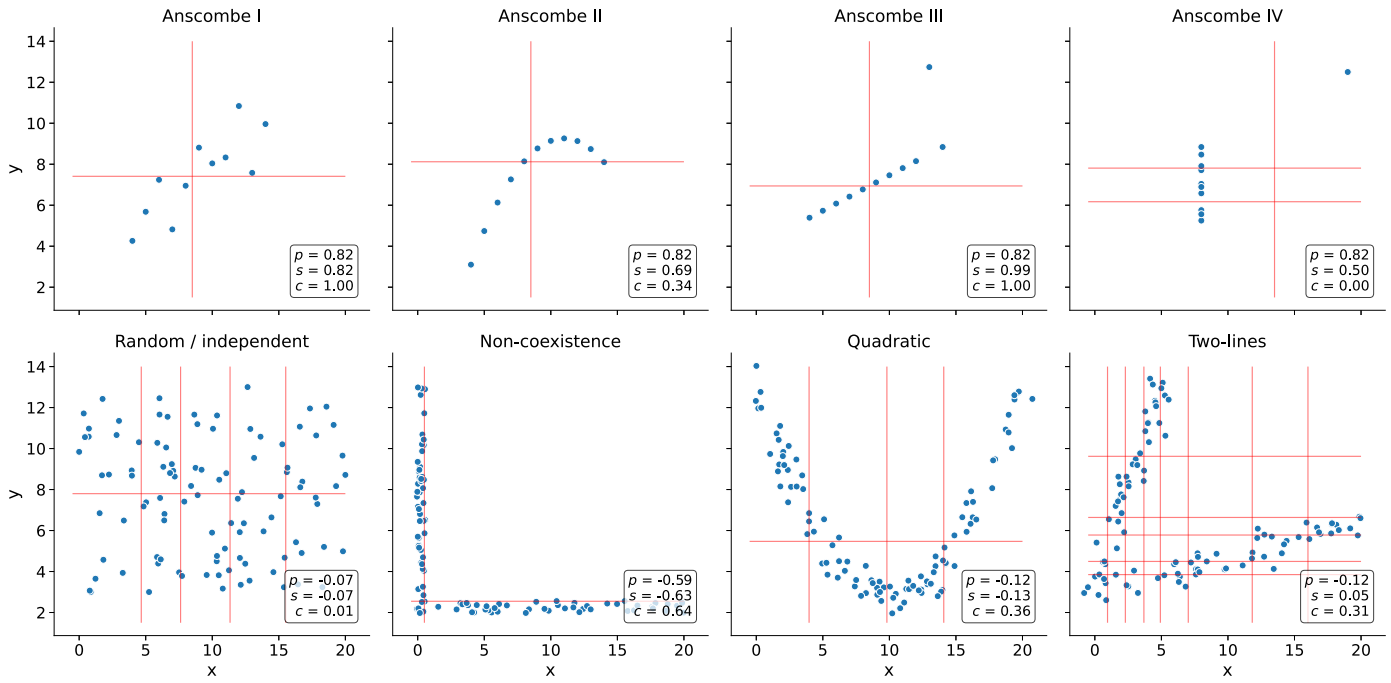

Figure 1: Different types of relationships in data. Each panel contains a set of simulated data points described by two variables: x and y. The first row shows Anscombe's quartet with four different datasets (from Anscombe I to IV) and 11 data points each. The second row contains a set of general patterns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

First, we examined how the Pearson (p), Spearman (s) and CCC (c) correlation coefficients behaved on different simulated data patterns. In the first row of Figure 1, we examine the classic Anscombe's

quartet [29], where red lines indicate how CCC clusters data points using each variable/feature individually (either x or y). Anscombe's quartet 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" [30,31,32], 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 non-linear relationships) can be masked by summary statistics alone. For example, Anscombe I contains a noisy but clear linear pattern, similar to Anscombe III where the linearity is perfect besides one outlier. In these two examples, CCC separates data points using two clusters (one red line for each variable x and y), yielding 1.0, indicating a strong relationship. Anscombe II seems to follow a quadratic relationship and is interpreted as linear by Pearson and Spearman. In contrast, CCC yields a lower yet non-zero value of 0.34, reflecting a more complex relationship than a linear pattern. Anscombe IV shows a vertical line where x values are almost constant except for one outlier. This outlier does not influence CCC as it does for Pearson or Spearman. Thus $c = 0.00$ (the minimum value) correctly indicates no association for this variable pair because, besides the outlier, for a single value of x there are ten different values for y. This variable pair does not fit the CCC assumption: the two clusters formed with x (approximately separated by $x = 13$) do not match the three clusters formed with y. The Pearson's correlation coefficient is the same across all these Anscombe's examples ($p = 0.82$), whereas Spearman is always above or equal to 0.50. These simulated datasets show that both Pearson and Spearman are very powerful in detecting linear patterns. However, any deviation in this assumption (like non-linear relationships or outliers) affects their robustness. One reason for this behavior is that these coefficients are based on data statistics alone, such as the mean, standard deviation or simple rankings, which seem to fall short in dealing with noisy data or more complex patterns.

The second row of Figure 1 shows other simulated relationships with general non-linear patterns, some previously observed in gene expression data [33,34,35]. For the random/independent pair of variables, all coefficients correctly agree with a value close to zero. The non-coexistence pattern, correctly captured by all coefficients, represents a case where one gene (x) might be expressed while the other one (y) is inhibited, highlighting a potentially strong biological relationship (such as a microRNA negatively regulating another gene). For the other two examples (quadratic and two-lines), Pearson and Spearman do not capture the non-linear pattern between variables x and y. These patterns also show how CCC uses different degrees of complexity to capture the relationships. For the quadratic pattern, for example, CCC separates x into more clusters (four in this case) to reach the maximum ARI. The two-lines example shows two embedded linear relationships with different slopes, which neither Pearson nor Spearman detect ($p = -0.12$ and $s = 0.05$, respectively). Here, CCC increases the complexity of the model by using eight clusters for x and six for y, resulting in $c = 0.31$.

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 can focus human interpretation on patterns that are more likely to reflect real biology. Those that capture non-linear patterns can reveal relationships missed by the linear-only coefficients that are often deployed.

The CCC reveals linear and non-linear patterns in human transcriptomic data

Here we compare the coefficients on real gene expression data and highlight some complex and biologically interesting relationships that only CCC detects. We used gene expression data from GTEx v8 across different tissues. We selected the top 5,000 genes with the largest variance for our initial analyses on whole blood and then computed the correlation matrix between genes using Pearson, Spearman and CCC (see Methods).

In Figure 2 a, we show how the correlation values distribute, where 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 cumulative histogram in Figure 2 b shows that coefficients reach 70% of gene pairs at $c = 0.18$, $p = 0.44$ and $s = 0.56$. A 2D histogram plot comparing each coefficient is shown in Figures 2 c, where CCC and Spearman tend to agree more than any of these with Pearson.

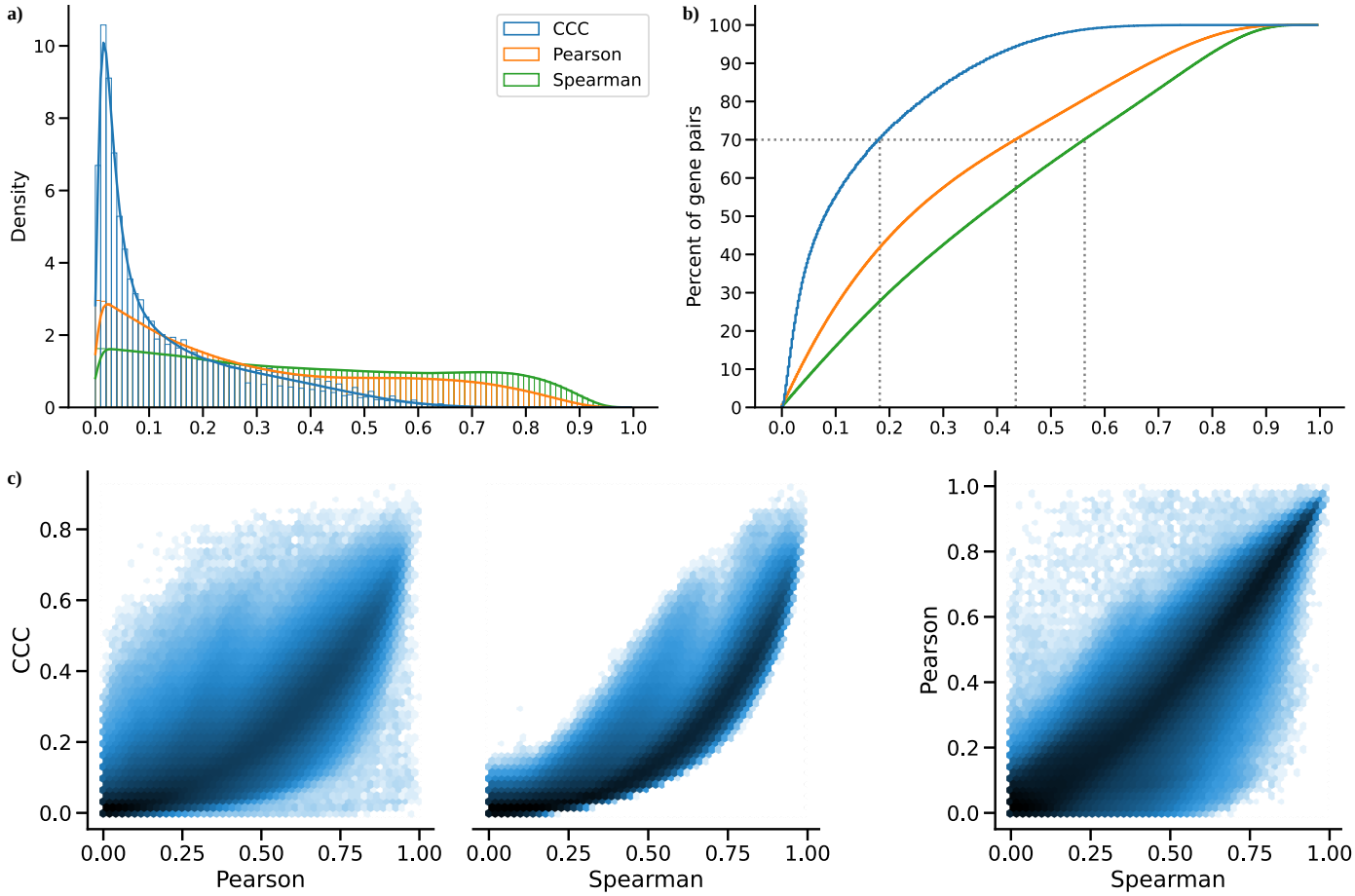

Figure 2: Distribution of coefficient values on gene expression (GTEx v8, whole blood). a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all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

A closer inspection of gene pairs that were either prioritized or disregarded by these coefficients revealed their ability to capture different patterns. For this, 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. An UpSet plot [36] is shown in Figure 3 a, where the intersections of these six categories allowed to precisely identify the gene expression patterns captured and missed by each coefficient.

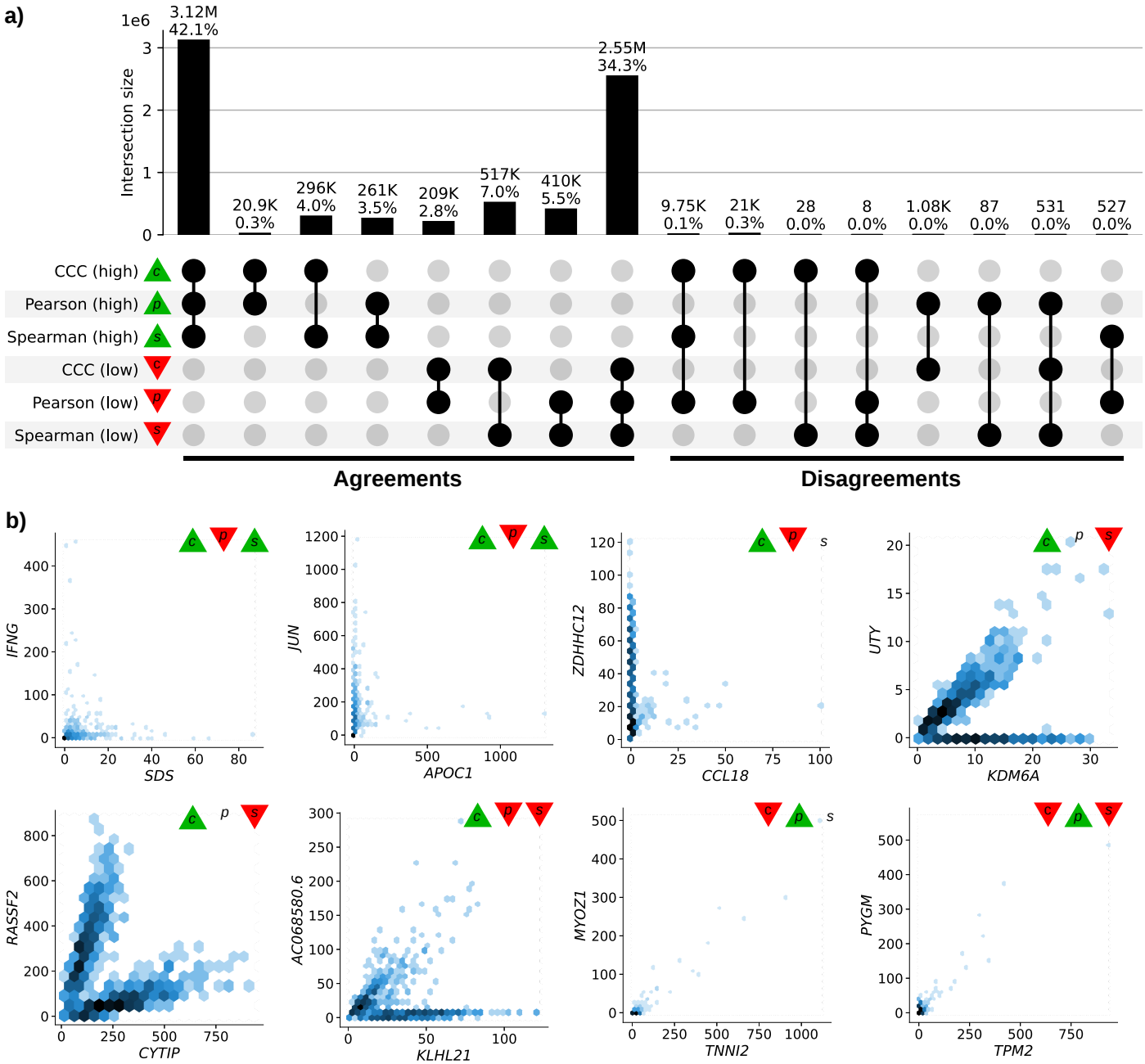

Figure 3: Intersection of gene pairs with high and low correlation coefficient values (GTEx v8, whole blood). a) UpSet plot with six categories (rows) grouping the 30% of the highest (green triangle) and lowest (red triangle) values for each coefficient. Columns show different intersections of categories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

For most cases, the three coefficients agreed on whether there is a strong correlation (42.1%) and whether there is no relationship (34.3%). Since Pearson and Spearman are linear-only, and CCC can also capture these patterns, it would be relatively safe to assume that these highly correlated gene pairs represent strong linear patterns. These results are crucial because it suggests that the user will not miss important linear patterns in expression data when using CCC. CCC and Spearman agree more on either highly or poorly correlated pairs (4.0% in "high", and 7.0% in "low") than any of these with Pearson (all between 0.3%-3.5% for "high", and 2.8%-5.5% for "low"). In summary, CCC agrees with either Pearson or Spearman in 90.5% of gene pairs, either by assigning a high or a low correlation value. Regarding disagreements (right part of Figure 3 a), more than 20 thousand gene pairs (20,987) with a high CCC value are not highly ranked by any other coefficient. There are also gene pairs with a high Pearson value and either low CCC (1,075), low Spearman (87) or both low CCC and low Spearman values (531). However, these cases mostly seem to be driven by potential outliers (Figure 3 b, analyzed

below). CCC does not miss gene pairs highly ranked by Spearman. We also compared the Maximal Information Coefficient (MIC) in a small random sample of this data (see Supplementary Material), given its higher computational complexity. We found that CCC behaves very similarly to MIC, although CCC performs up to two orders of magnitude faster than this coefficient (see Supplementary Material). This result is relevant because MIC, an advanced correlation coefficient able to capture general patterns beyond linear relationships, represented a significant step forward in correlation analysis research, and it has been successfully used in various application domains [4,37,38]. CCC, however, runs in a fraction of the time, making it a practical alternative for large data sets.

Next, we selected individual gene pairs where CCC disagreed with the rest. We analyzed gene pairs among the top five of each intersection in the "Disagreements" group (Figure 3 b, right) where CCC disagrees with Pearson, Spearman or both. The first three gene pairs at the top (*IFNG* - *SDS*, *JUN* - *APOC1*, and *ZDHHC12* - *CCL18*), with high CCC and low Pearson values, seem to follow a non-coexistence relationship: in samples where one of the genes is highly (slightly) expressed, the other is slightly (highly) activated, suggesting a potential inhibiting effect. The following three gene pairs (*UTY* - *KDM6A*, *RASSF2* - *CYTIP*, and *AC068580.6* - *KLHL21*) follow patterns combining either two linear or one linear and one independent relationships. In particular, genes *UTY* and *KDM6A* (paralogs) show a non-linear relationship where a subset of samples follows a robust linear pattern and another subset has a constant expression of one gene. This relationship is explained by the fact that *UTY* is in chromosome Y (Yq11) whereas *KDM6A* is in chromosome X (Xp11), and samples with a linear pattern are males, whereas those with no expression for *UTY* are females. This combination of linear and independent patterns is captured by CCC ($c = 0.29$, above the 80th percentile) but not by Pearson ($p = 0.24$, below the 55th percentile) or Spearman ($s = 0.10$, below the 15th percentile). Furthermore, the same gene pair pattern is highly ranked by CCC in all other tissues in GTEx, except for female-specific organs (Figure 4).

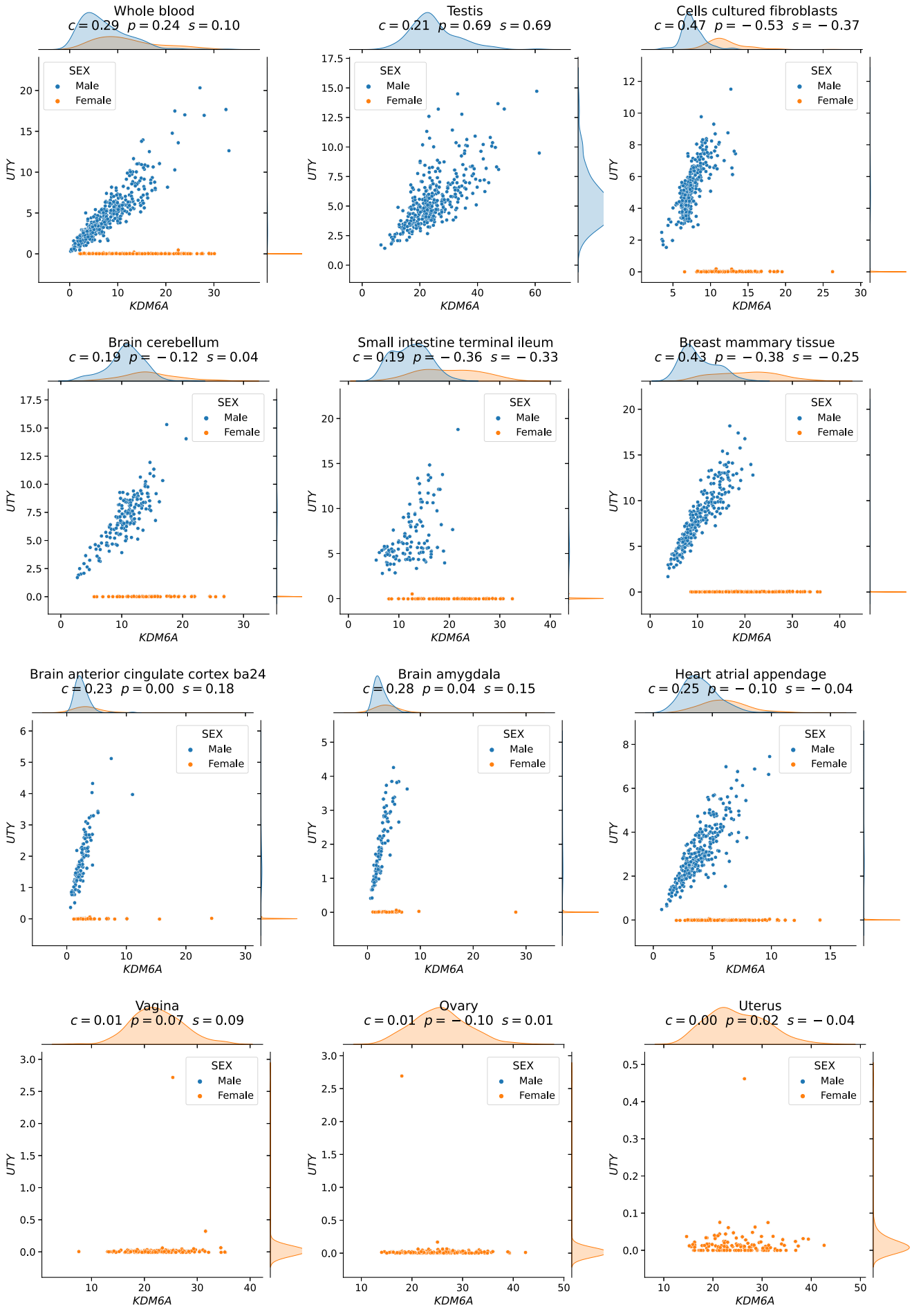

Figure 4: Scatter plots of genes *KDM6A* and *UTY* across different GTEx tissues. CCC captures this non-linear relationship in all GTEx tissues (nine examples are shown in the first three rows), except in female-specific organs (last row).

This particular example highlights the importance of considering sex as a biological variable (SABV) [39] to avoid overlooking important differences between men and women, for example, in disease manifestations [40,41]. In transcriptome studies, a not-only-linear correlation coefficient like CCC could identify sex-specific differences in gene expression that, like the *UTY* - *KDM6A* example, are entirely missed by linear-only coefficients.

Some of these genes, such as *SDS* (12q24) and *ZDHHC12* (9q34), were previously found to have a relatively lower number of publications that were explained by a small set of chemical, physical and biological features [42]. A gene co-expression analysis on large compendia and beyond linear patterns could shed light on the function of understudied genes. On the other hand, gene *KLHL21* (1p36) and the novel RNA gene *AC068580.6* (*ENSG00000235027*, in 11p15) have a high CCC value and are entirely missed by the other methods. *KLHL21* was suggested as a potential therapeutic target for hepatocellular carcinoma [43] and other cancers [44,45], and its non-linear correlation with *AC068580.6* and potentially other genes might unveil other important players in cancer initiation or progression.

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

We analyzed the other gene pairs in Figure 3 b 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) [46,47], 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 [48], making it an ideal case for replication. These networks were built from expression and different interaction measurements, including gene co-expression (using Pearson correlation after several preprocessing steps), protein-interaction, transcription factor regulation, and chemical/genetic perturbations and microRNA target profiles from the Molecular Signatures Database (MSigDB [49]). We reasoned that our highly-ranked gene pairs using three different coefficients in a single tissue (whole blood in GTEx, Figure 3 b) should replicate across the multi-tissue, multi-cell type functional interaction networks in GIANT. First, for each pair in Figure 3 b for which genes were present in GIANT models (this excluded *AC068580.6* - *KLHL21*, which we comment below), we ran the tissue-specific interaction analysis selecting “blood”, which was the most equivalent to whole blood in GTEx. The predicted blood-specific networks for each of these six gene pairs is shown in Figure 5. 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 two genes.

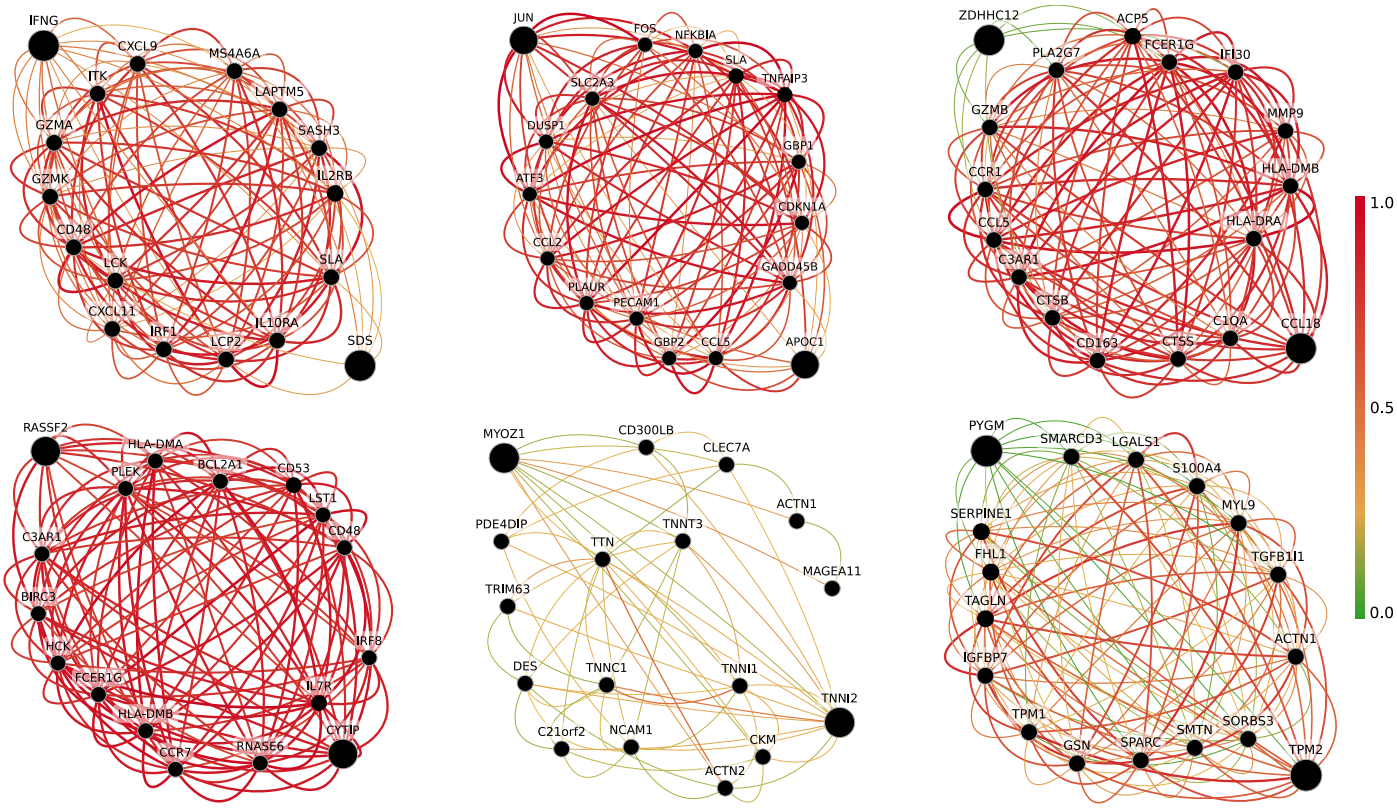

Figure 5: Predicted blood-specific networks from GIANT for six gene pairs prioritized by correlation coefficients. A node represents a gene, and an edge the probability that two genes are part of the same biological process in blood. A maximum of 15 genes are shown for each subfigure. The GIANT web application automatically determined a minimum interaction confidence (edges' weights) to be shown. All these analyses can be performed online using the following links: *IFNG - SDS* [50], *JUN - APOC1* [51], *ZDHHC12 - CCL18* [52], *RASSF2 - CYTIP* [53], *MYOZ1 - TNNI2* [54], *PYGM - TPM2* [55]. The GIANT web-server was accessed on April 4, 2022.

In general, gene pairs highly ranked by CCC were part of more cohesive networks compared with Pearson and Spearman. For example, the average probability of *IFNG* and *SDS* with other genes in the network was 0.42 and 0.29, respectively (Supplementary Table 1). A similar pattern was found for *JUN* and *APOC1*, where the interaction probability was 0.68 and 0.47, respectively. The evidence of interaction between *ZDHHC12* and *CCL18* was less evident, with interaction averages of 0.07 and 0.79, respectively. For this gene pair, a dense network was connected with *CCL18*, although the predicted interaction of *ZDHHC12* with this potential process was weak. *RASSF2* and *CYTIP*, with a high CCC value ($c = 0.20$) and low Pearson and Spearman ($p = 0.16$ and $s = 0.11$), were both strongly connected to the network, with interaction scores of at least 0.63 for both and an average of 0.75 and 0.84, respectively. The predicted networks for the two gene pairs exclusively prioritized by Pearson were much less cohesive. *MYOZ1* and *TNNI2* had low average interaction scores of 0.17 and 0.22 with the rest of the genes. On the other hand, the average interaction probability of *PYGM* with the network was very low (0.04), where only *TPM2* showed a strong connection to the network. This lack of replication for these two gene pairs in GIANT suggests that their high Pearson value in GTEx is driven by outliers (Figure 3 b).

We also assessed whether top gene pairs detected in whole blood from GTEx had evidence of being specifically expressed in blood cell lineages. For this, we let GIANT automatically predict a relevant tissue- or cell type-specific network (see Methods). All six gene pairs prioritized by CCC were predicted to be specific to a blood-relevant cell lineage (Supplementary Figure 9): natural killer cells, mononuclear phagocytes, macrophages and leukocytes. Additionally, the interaction of these gene pairs with the cell lineage-specific networks improved over the blood-specific network (Supplementary Table 1). Conversely, the top two Pearson gene pairs were predicted to be specific to skeletal muscle, and their interaction decreased. To confirm whether these patterns were globally present, we performed a systematic evaluation using the top 100 gene pairs from each intersection where CCC disagrees with Pearson and Spearman. We automatically predicted a relevant tissue or cell type

network for each gene pair using GIANT. Then, we took the top 5 more common predicted tissues/cell types in each coefficient intersection (shown in Figure 3 a). The results are in Figure 6, where each row shows 1) the number of gene pairs predicted to be specific to a tissue/cell type (bar plot on the left) and 2) the average probability of gene interaction in each tissue-specific network (box plot on the right).

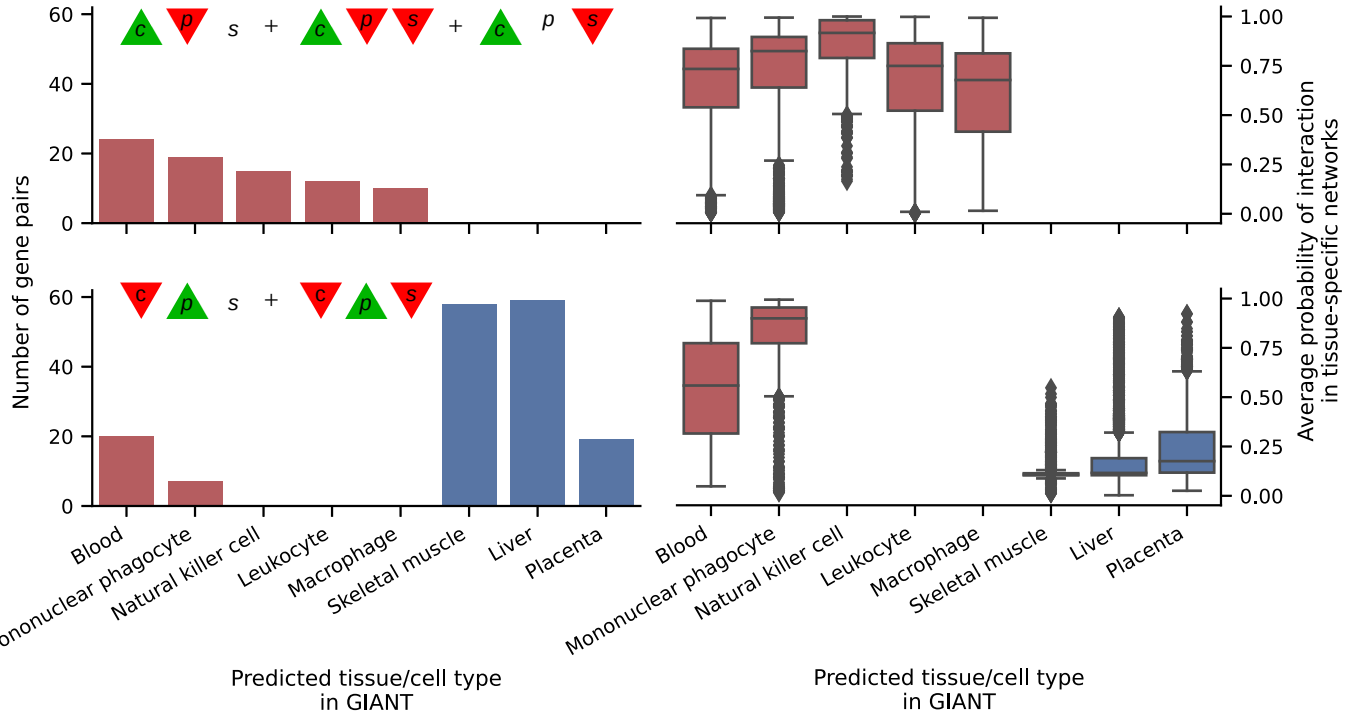

Figure 6: Summary of predicted tissue/cell type networks for gene pairs exclusively prioritized by CCC and Pearson. The first row combines all gene pairs where CCC is high and Pearson or Spearman are low. The second row combines all gene pairs where Pearson is high and CCC or Spearman are low. Bar plots (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.

Notably, most gene pairs highly prioritized by CCC and not by the rest were predicted to be strongly blood-specific (indicated with red bars/boxes), including several cell lineages in this tissue like natural killer cells, macrophages, and other leukocytes. The average probability of interaction between genes in these CCC-ranked networks was also very high, with all medians close to 75% and first quartiles above 40%. In contrast, most Pearson's gene pairs were predicted to be specific to tissues unrelated to blood. The probabilities of interaction in these Pearson-ranked networks were also generally lower than in CCC, except for blood-specific gene pairs. The gene-gene relationships exclusively detected by CCC in whole blood from GTEx were replicated in more sophisticated network models trained without GTEx. These results suggest that CCC can disentangle intricated cell lineage-specific transcriptional patterns missed by linear-only coefficients.

Discussion

We introduced the Clustermatch Correlation Coefficient (CCC), an efficient not-only-linear machine learning-based method that captures more general patterns in data. We applied it to gene expression data from GTEx v8 and found that our coefficient is robust to outliers and does not miss strongly linear relationships in gene-gene patterns. Moreover, CCC also captured complex and biologically meaningful relationships completely missed by standard coefficients. For instance, CCC exclusively detected gene pairs from the sex chromosomes following complex non-linear patterns. These examples, in particular, highlight the importance of considering sex as a biological variable and the critical role that not-only-linear coefficients can play in capturing sex-specific differences. We also

replicated gene pairs prioritized by CCC in GTEx using tissue-specific gene networks from GIANT trained without GTEx samples. We found that top CCC-ranked gene pairs in whole blood from GTEx were predicted in GIANT to be part of the same biological mechanism and specifically expressed in cell lineages from blood, even though CCC did not have access to any cell lineage-specific information. CCC derives scores very well aligned with MIC while being much more computationally efficient and thus practical for use in large modern datasets. Our approach has a single parameter that controls the complexity of the detected relationships while also balancing compute time. Finally, CCC is conceptually easy to interpret and robust to outliers, making it an advanced coefficient that can focus human interpretation on patterns more likely to resemble biological processes.

It is well-known that biomedical research is biased towards a small fraction of human genes [56,57]. Researching genes with well-known functions is easier, although this observational bias seems anachronic with current high-throughput technologies. Several factors explaining this behavior have been identified [42], such as RNA and protein abundance or gene length. Another potential explanation could also be a bias towards linear-only statistical methods. For example, genome-wide association studies (GWAS), which in most cases employ a linear model, have been successful in identifying thousands of genetic variants associated with complex human diseases [58,59]. These findings have been beneficial to understand the molecular basis of common diseases [60] and, more recently, design polygenic scores to estimate an individual's genetic predisposition for a particular trait or disorder [61]. However, the estimated effects of genes identified with GWAS are generally modest, and they explain only a fraction of the trait variance, hampering the clinical translation of these findings [62]. Recent theories, like the omnigenic model for complex traits [16,17], argue that these observations in genetic studies are explained by highly-interconnected gene regulatory networks, with some core genes having a more direct effect on the phenotype than others. Using this omnigenic perspective, we and others [19,20,23] have shown that integrating gene co-expression networks in genetic studies could potentially identify core genes that are missed by linear-only models alone like GWAS. Our results strongly suggest that building these networks with more advanced and efficient correlation coefficients could better estimate gene co-expression profiles and thus more accurately identify these core genes. We anticipate that approaches like CCC will play a significant role in the precision medicine field by providing the computational tools to focus on less-studied yet potentially more promising genes.

Our analyses have some limitations. We worked on a sample with the top variable genes to keep computation time feasible. Although CCC is much faster than MIC, Pearson and Spearman are still the most computationally efficient since they only rely on simple data statistics. However, as we have shown, this significantly limits their ability beyond linear patterns. Even with this small sample of genes, our results confirm that the advantages of using more advanced coefficients like CCC can help detect and study more intricate molecular mechanisms. Although we only used GTEx, we could still find complex and meaningful patterns within this homogeneous set of samples. The application of CCC on larger compendia, such as recount3 [11] with thousands of heterogeneous samples across different conditions, can reveal other potentially meaningful gene interactions. The single parameter of CCC, $k_{\max}$, controls the complexity of patterns found and also the compute time. We used $k_{\max} = 10$ in our analyses which, given our results, seems to be enough to find both linear and more complex patterns. A more comprehensive analysis of the most optimal values for this parameter could provide insights into better adjusting it for different applications.

While linear and rank-based correlation coefficients are exceptionally fast to calculate, not all relevant patterns in biological datasets are linear. For example, patterns associated with sex as a biological variable are not apparent to the linear-only coefficients that we evaluated but are revealed by not-only-linear methods. Not-only-linear coefficients can also disentangle intricate, yet relevant patterns from gene expression data alone 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. We found that the CCC is an efficient, next-generation correlation coefficient that is highly effective in transcriptome analyses and potentially useful in a broad range of other domains.

Methods

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

The CCC algorithm

The Clustermatch Correlation Coefficient (CCC) computes a similarity value $c \in [0, 1]$ between any pair of numerical or categorical features/variables $\mathbf{x}$ and $\mathbf{y}$ measured on n objects. CCC assumes that if two features $\mathbf{x}$ and $\mathbf{y}$ are similar, then the partitioning of the n objects using each feature separately should match. For example, given $\mathbf{x} = (1.1, 2.7, 3.2, 4.0)$ and $\mathbf{y} = 10x = (11, 27, 32, 40)$ for $n = 4$, partitioning each variable into two clusters ($k = 2$) using their medians (2.95 for $\mathbf{x}$ and 29.5 for $\mathbf{y}$) would result in partition $\pi^{\mathbf{x}} = (1, 1, 2, 2)$ for $\mathbf{x}$, and partition $\pi^{\mathbf{y}} = (1, 1, 2, 2)$ for $\mathbf{y}$. If we compute the agreement between $\pi^{\mathbf{x}}$ and $\pi^{\mathbf{y}}$ using any measure of similarity between partitions, like the adjusted Rand index (ARI) [28], it will return the maximum value (1.0 in the case of ARI). For CCC, a given value of k might not be the right one to find a relationship between two features. For instance, in the quadratic example in Figure 1, CCC returns a value of 0.36 (grouping objects in four clusters using one feature and two using the other). If we used only two clusters instead, CCC would return a similarity value of 0.02. The CCC algorithm (shown below) searches for this optimal number of clusters given a maximum k, which is its single parameter $k_{\max}$.

Algorithm 1: CCC algorithm

1 **Function** `get_partitions`($\mathbf{v}$, $k_{\max}$)**:**
  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 $\Omega_{rc} \leftarrow \{i \mid v_i = \mathcal{C}_c\}, \forall c \in [1, r]$
  `// Remove singleton partitions`
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 main function of the algorithm, `ccc`, generates a list of partitionings $\Omega^{\mathbf{x}}$ and $\Omega^{\mathbf{y}}$ (lines 14 and 15), for each feature $\mathbf{x}$ and $\mathbf{y}$. Then, it computes the ARI between each partition in $\Omega^{\mathbf{x}}$ and $\Omega^{\mathbf{y}}$ (line 16), and then it keeps the pair that generates the maximum ARI. Finally, since ARI does not have a lower bound (it could return negative values, which in our case are not meaningful), CCC always returns a value between 0 and 1 (line 17).

Since CCC only needs a pair of partitions to compute a similarity value, any type of feature that can be used to perform clustering/grouping of the n objects is supported. If the feature is numerical (lines 2 to 5 in the `get_partitions` function), then quantiles are used for clustering (for example, the median generates $k = 2$ clusters of objects), from $k = 2$ to $k = k_{\max}$. If the feature is categorical (lines 7 to 9), the categories are used to group objects together. Consequently, since features are internally categorized into clusters, numerical and categorical variables can be naturally integrated since clusters do not need an order. Although not developed in this study, CCC provides a framework where not only 1-dimensional variables can be compared (such as genes across n samples) but, in theory, also random vectors (multivariate random variables) such as an image.

For all our analyses we used $k_{\max} = 10$. This means that for each gene pair, 20 partitions are generated (10 for each gene), and 100 ARI comparisons are performed. Smaller values of $k_{\max}$ can reduce computation time, although at the expense of missing more complex, general relationships. Our examples in Figure 1 suggest that using $k_{\max} = 2$ would force CCC to find linear-only patterns, which could be a valid use case scenario where only this kind of relationships are desired. In addition, $k_{\max} = 2$ implies that only two partitions are generated, and only one ARI comparison is performed.

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

A Python implementation of CCC (optimized with `numba` [63]) can be found in our Github repository [64], as well as a package published in the Python Package Index (PyPI) that can be easily installed.

Maximal Information Coefficient (MIC)

We used the Python package `minepy` [65,66] (version 1.2.5) to estimate the MIC coefficient. In GTEx v8 (whole blood), we ran MIC (the original implementation using the heuristic estimator ApproxMaxMI [33]) with the default parameters `alpha=0.6`, `c=15` and `estimator='mic_approx'`. For our computational complexity analyses (see Supplementary Material), we also ran a new optimized implementation called MIC_e [67] provided by `minepy` (using parameter `estimator='mic_e'`).

Gene expression data, preprocessing and sampling

We downloaded GTEx v8 data for all tissues, normalized using TPM (transcripts per million), and focused our primary analysis on whole blood, which has a good sample size (755). We selected the top 5,000 genes from whole blood with the largest variance after standardizing with $log(x+1)$ to avoid a bias towards highly-expressed genes. We then computed Pearson, Spearman and CCC on these 5,000 genes across all 755 samples on the TPM-normalized data, generating a pairwise similarity matrix of size 5,000 x 5,000. To reduce the time to compute MIC and compare it with the other coefficients, we randomly sampled 100,000 gene pairs from all possible combinations in the set of 5,000 top variable genes ($n * (n-1)/2 = 12497500$).

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 [47/]. The GIANT version used in this study included 987 genome-scale datasets with approximately 38,000 conditions from an estimated number of 14,000 publications. Details on how these networks were built are described in [27]. Briefly, tissue-specific gene networks were built using gene expression data (without GTEx samples [48]) from the NCBI's Gene Expression Omnibus (GEO) [68], protein-protein interaction (BioGRID [69], IntAct [70], MINT [71] and MIPS [72]), transcription factor regulation using binding motifs from JASPAR [73], and chemical and genetic perturbations from MSigDB [74]. 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 functional relationships were built using expert curation and experimentally derived gene annotations from the Gene Ontology. Then, one naive Bayesian classifier 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 [75] 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 additional 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.

References

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

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

3. **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 (2019-06-01) https://doi.org/gfg4bt
DOI: 10.1093/bioinformatics/bty899 · PMID: 30357313

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

5. **A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data.**
Xian-Fang Song, Yong Zhang, Dun-Wei Gong, Xiao-Zhi Gao
IEEE transactions on cybernetics (2021-03-17) https://www.ncbi.nlm.nih.gov/pubmed/33729976
DOI: 10.1109/tcyb.2021.3061152 · PMID: 33729976

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

23. **Identification of therapeutic targets from genetic association studies using hierarchical component analysis**
Hao-Chih Lee, Osamu Ichikawa, Benjamin S Glicksberg, Aparna A Divaraniya, Christine E Becker, Pankaj Agarwal, Joel T Dudley
BioData Mining (2020-12) https://doi.org/gjp5pf
DOI: 10.1186/s13040-020-00216-9 · PMID: 32565911 · PMCID: PMC7301559

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

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

26. **The GTEx Consortium atlas of genetic regulatory effects across human tissues**
The GTEx Consortium
Science (2020-09-11) https://doi.org/ghbnhr
DOI: 10.1126/science.aaz1776 · PMID: 32913098 · PMCID: PMC7737656

27. **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://doi.org/f7dvkv
DOI: 10.1038/ng.3259 · PMID: 25915600 · PMCID: PMC4828725

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

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

30. **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

31. **Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing**
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 · ISBN: 9781450346559

32. **Generating data sets for teaching the importance of regression analysis**
Lori L Murray, John G Wilson
Decision Sciences Journal of Innovative Education (2021-04) https://doi.org/gjmgqt
DOI: 10.1111/dsji.12233

33. **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

34. **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

35. **Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by 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

36. **UpSet: Visualization of Intersecting Sets**
Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister
IEEE Transactions on Visualization and Computer Graphics (2014-12-31) https://doi.org/f3ssr5

DOI: 10.1109/tvcg.2014.2346248 · PMID: 26356912 · PMCID: PMC4720993

37. **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

38. **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

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

40. **Considering Sex as a Biological Variable in Basic and Clinical Studies: An Endocrine Society Scientific Statement**
Aditi Bhargava, Arthur P Arnold, Debra A Bangasser, Kate M Denton, Arpana Gupta, Lucinda M Hilliard Krause, Emeran A Mayer, Margaret McCarthy, Walter L Miller, Armin Raznahan, Ragini Verma
Endocrine Reviews (2021-05-25) https://doi.org/gm642r
DOI: 10.1210/endrev/bnaa034 · PMID: 33704446 · PMCID: PMC8348944

41. **Considering sex as a biological variable will require a global shift in science culture**
Rebecca M Shansky, Anne Z Murphy
Nature Neuroscience (2021-04) https://doi.org/gjhkx8
DOI: 10.1038/s41593-021-00806-8 · PMID: 33649507

42. **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

43. **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

44. **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

45. **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

46. **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

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

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

49. **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

50. **SDS, IFNG - HumanBase** https://hb.flatironinstitute.org/gene/10993+3458

51. **JUN, APOC1 - HumanBase** https://hb.flatironinstitute.org/gene/3725+341

52. **CCL18, ZDHHC12 - HumanBase** https://hb.flatironinstitute.org/gene/6362+84885

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

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

55. **PYGM, TPM2 - HumanBase** https://hb.flatironinstitute.org/gene/5837+7169

56. **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

57. **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

58. **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

59. **15 years of genome-wide association studies and no signs of slowing down**
Ruth JF Loos

Nature Communications (2020-12) https://doi.org/gpt8k5
DOI: 10.1038/s41467-020-19653-5 · PMID: 33214558 · PMCID: PMC7677394

60. **Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps**
Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R Robertson, Jason M Torres, NWilliam Rayner, Anthony J Payne, Valgerdur Steinthorsdottir, Robert A Scott, Niels Grarup, … Mark I McCarthy
Nature Genetics (2018-11) https://doi.org/gfb68d
DOI: 10.1038/s41588-018-0241-6 · PMID: 30297969 · PMCID: PMC6287706

61. **Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations**
Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, Sekar Kathiresan
Nature Genetics (2018-09) https://doi.org/gdz64f
DOI: 10.1038/s41588-018-0183-z · PMID: 30104762 · PMCID: PMC6128408

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

63. **Numba: a LLVM-based Python JIT compiler**
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 · ISBN: 9781450340052

64. **GitHub - greenelab/clustermatch-gene-expr**
GitHub
https://github.com/greenelab/clustermatch-gene-expr

65. **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 (2013-02-01) https://doi.org/f4nxg6
DOI: 10.1093/bioinformatics/bts707 · PMID: 23242262

66. **GitHub - minepy/minepy: minepy - Maximal Information-based Nonparametric Exploration**
GitHub
https://github.com/minepy/minepy

67. **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

68. **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

69. **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

70. **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 (2012-01-01) https://doi.org/bpmrdk
DOI: 10.1093/nar/gkr1088 · PMID: 22121220 · PMCID: PMC3245075

71. **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 (2012-01) https://doi.org/cqvx3b
DOI: 10.1093/nar/gkr930 · PMID: 22096227 · PMCID: PMC3244991

72. **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

73. **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 (2010-01) https://doi.org/ddwfqp
DOI: 10.1093/nar/gkp950 · PMID: 19906716 · PMCID: PMC2808906

74. **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 (2005-10-25) https://doi.org/d4qbh8
DOI: 10.1073/pnas.0506580102 · PMID: 16199517 · PMCID: PMC1239896

75. **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 (2013-11) https://doi.org/f5g4hm
DOI: 10.1101/gr.155697.113 · PMID: 23950145 · PMCID: PMC3814886

76. **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

77. **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

78. **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

79. **A Novel Algorithm for the Precise Calculation of the Maximal Information Coefficient**
Yi Zhang, Shili Jia, Haiyun Huang, Jiqing Qiu, Changjie Zhou
Scientific Reports (2015-05) https://doi.org/gpt7c7
DOI: 10.1038/srep06662 · PMID: 25322794 · PMCID: PMC4200418

Supplementary material

Comparison with the Maximal Information Coefficient (MIC) on gene expression data

We compared all the coefficients in this study with the MIC [24], a popular non-linear method that can find complex relationships in data, although very computationally intensive [76]. To circumvent this limitation of MIC, we took a small random sample of 100,000 gene pairs from 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 7. We verified that CCC and MIC behave very similarly in this dataset, with essentially the same distribution but only shifted. Figure 7 c shows that these two coefficients relate almost linearly and both compare very similarly with Pearson and Spearman.

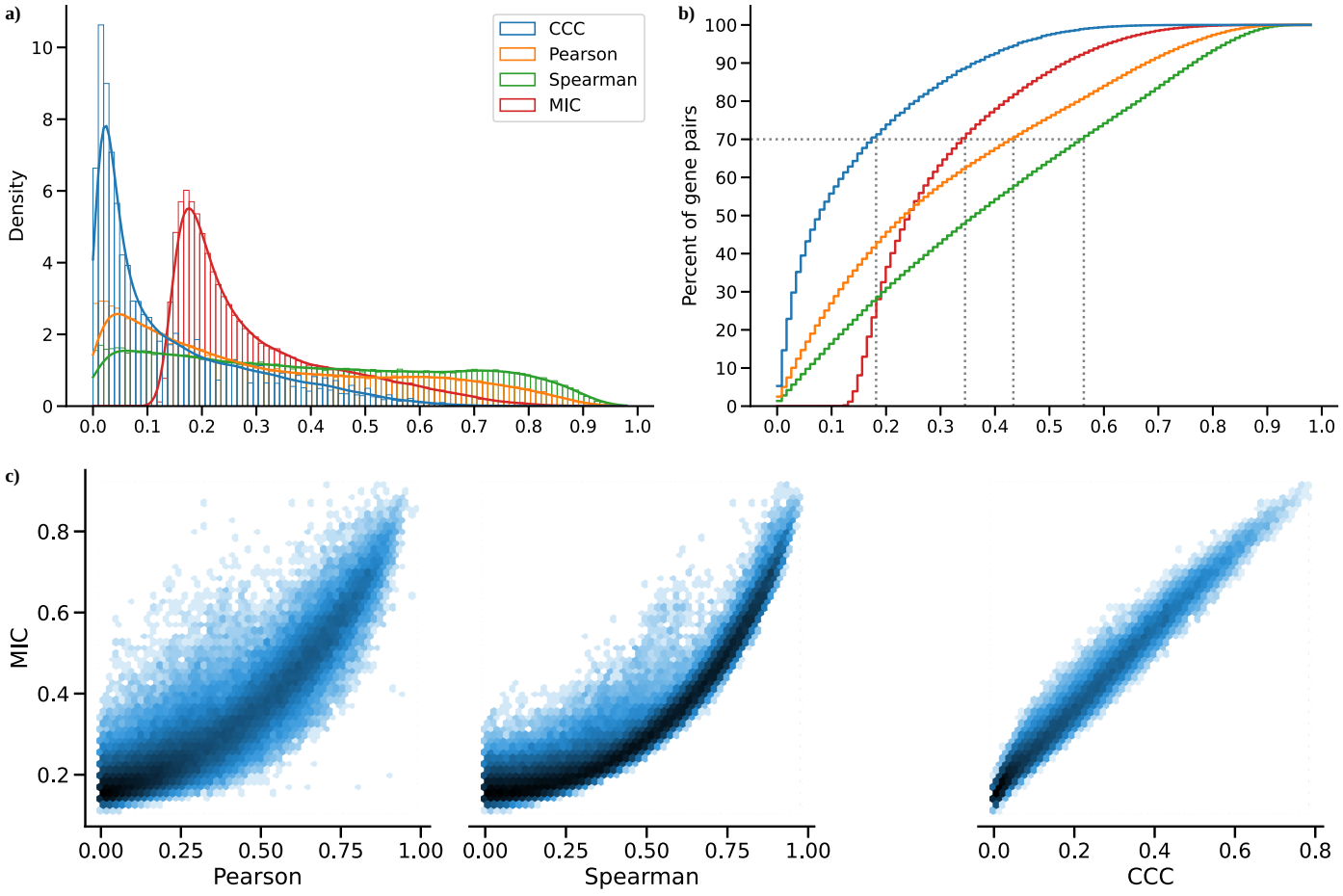

Figure 7: Distribution of MIC values on a small sample from whole blood (GTEx v8) and comparison with other methods. We randomly sampled 100,000 gene pairs (approximately 1% of the total) from our set of 5,000 genes. **a)** Histogram of coefficient values. **b)** Corresponding cumulative histogram. The 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.

Computational complexity of coefficients

We also compared CCC with the other coefficients in terms of computational complexity. Although CCC and MIC identify the same gene pairs in the gene expression data (see here), the use MIC in large datasets remains limited due to its very long computation time, despite some methodological/implementation improvements [65,76,77,78,79]. The original MIC implementation uses ApproxMaxMI, a computationally demanding heuristic estimator [33]. Recently, a more efficient implementation called MIC_e was proposed [67]. These two MIC estimators are provided by the `minepy` package [65], a C implementation available for Python, which we used here. We compared all these coefficients in terms of computation time on randomly generated variables of differente sizes, which simulates an scenario of gene expression data with different numbers of conditions. Differently than the rest, our method CCC allows to easily parallelize the computation of a single gene pair (see Methods), so we also tested the cases using 1 and 3 CPU cores. The results in Figure 8 show the time in seconds (left) and its log scale (right).

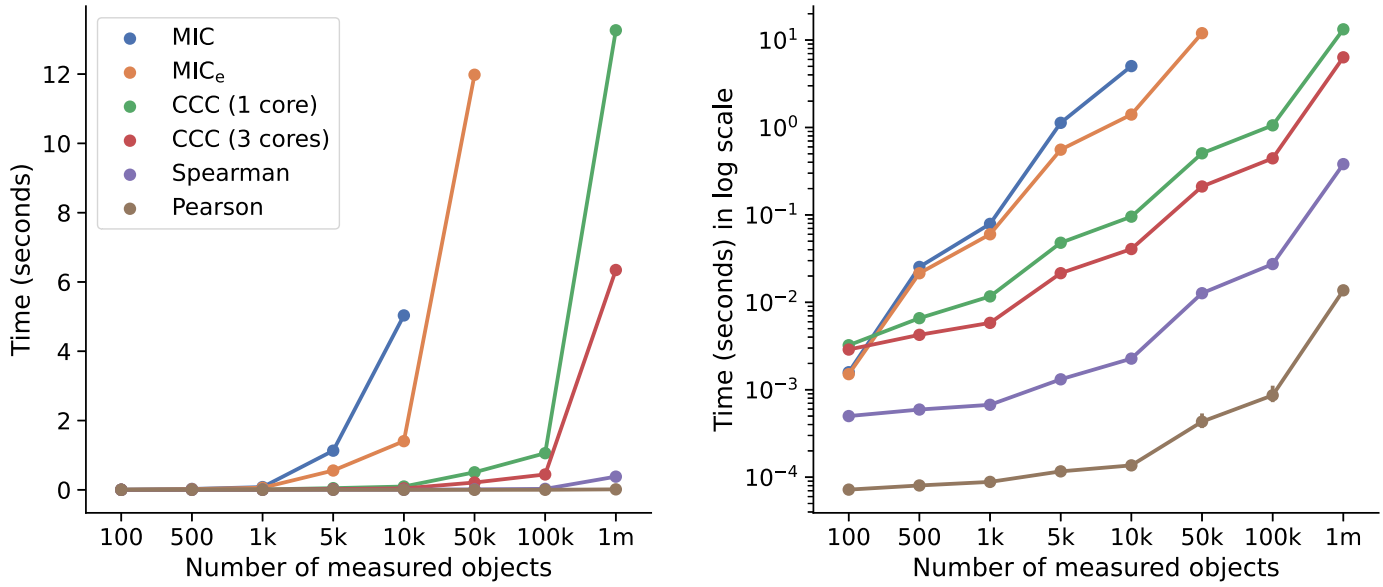

Figure 8: Computational complexity of all correlation coefficients on simulated data. We simulated variables/features with varying data sizes (from 100 to a million, x-axis). The plots show the average time taken for each coefficient (in seconds on the left, and log scale on the right) on 10 repetitions (1000 repetitions were performed for data size 100). CCC was run using 1 and 3 CPU cores. MIC and MIC_e did not finish running in a reasonable amount of time for data sizes of 10,000 and 100,000, respectively.

As we already expected, Pearson and Spearman were the fastest, given that they only need to compute basic summary statistics from the data. For example, Pearson is three orders of magnitude faster than CCC. Among the non-linear 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 [65], 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 at the single-core CCC. This suggests that new implementations using more advaned processing units (such as GPUs) are feasible and could make CCC reach speeds closer to Pearson.

Tissue-specific gene networks with GIANT

Table 1: Network statistics of six gene pairs (prioritized by correlation coefficients) for blood and predicted cell types.

The minimum, average and maximum interaction coefficients for genes in each network are shown. For each gene pair, also the direct interaction coefficient in each cell type is shown.

	Interaction confidence							
	Blood				Predicted cell type			
Gene	Direct	Min.	Avg.	Max.	Direct	Min.	Avg.	Max.
IFNG	0.11	0.19	0.42	0.54	0.50	0.74	0.90	0.99
SDS		0.18	0.29	0.41		0.65	0.81	0.94
JUN	0.07	0.26	0.68	0.97	0.07	0.36	0.73	0.94
APOC1		0.22	0.47	0.77		0.29	0.50	0.80
ZDHHC12	0.04	0.05	0.07	0.10	0.02	0.03	0.12	0.33
CCL18		0.74	0.79	0.86		0.36	0.70	0.90
RASSF2	0.63	0.63	0.75	0.90	0.65	0.69	0.75	0.88
CYTIP		0.63	0.84	0.91		0.76	0.84	0.91
MYOZ1	0.10	0.09	0.17	0.37	0.01	0.11	0.11	0.12
TNNI2		0.10	0.22	0.44		0.10	0.11	0.12
PYGM	0.05	0.02	0.04	0.14	0.01	0.01	0.02	0.04
TPM2		0.05	0.56	0.80		0.01	0.28	0.47

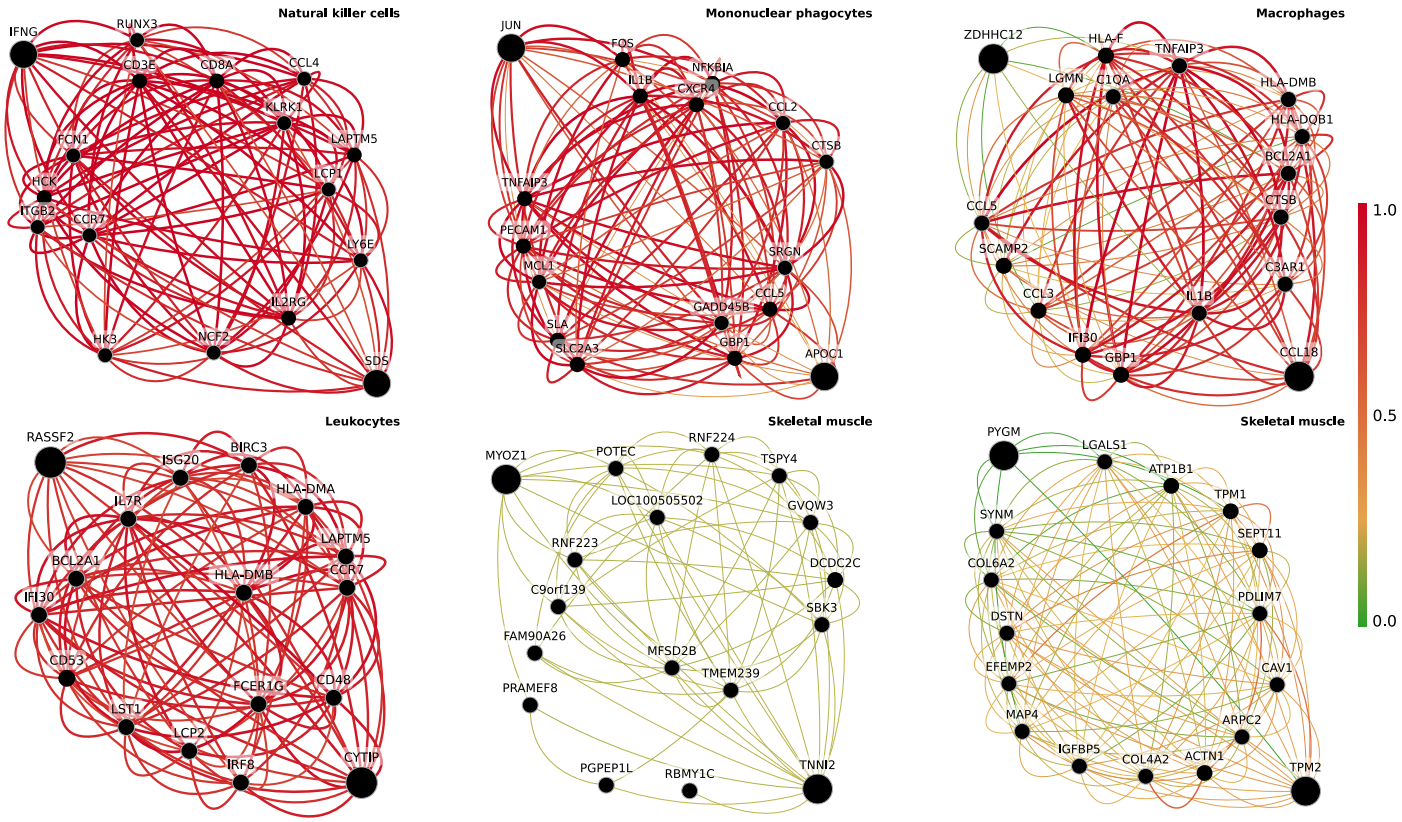

Figure 9: Predicted tissue-specific networks from GIANT for six gene pairs prioritized by correlation coefficients. Gene pairs are from Figure 3 b. A node represents a gene and an edge the probability that two genes are part of the same biological process in a specific cell type. The cell type for each gene network was automatically predicted using [75] and it is indicated at the top-right corner of each network. A maximum of 15 genes are shown for each subfigure. The GIANT web application automatically determined a minimum interaction confidence (edges's weights) to be shown.

All these analyses can be performed online using the following links: *IFNG* - *SDS* [50], *JUN* - *APOC1* [51], *ZDHHC12* - *CCL18* [52], *RASSF2* - *CYTIP* [53], *MYOZ1* - *TNNI2* [54], *PYGM* - *TPM2* [55]. The GIANT web-server was accessed on April 4, 2022.