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

This manuscript (permalink) was automatically generated from greenelab/clustermatch-gene-expr-manuscript@b2593ea on March 15, 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 used across different research areas to identify intriguing relationships or answer critical questions. Knowing that genes have correlated expression can suggest that they share functions or are part of networks that influence different traits. Indeed, gene regulatory networks have been playing an increasingly important role in precision medicine since the introduction of more advanced models of the genetic architecture of complex traits. However, the strategies to detect these gene-gene connections are usually deployed with linear correlation coefficients, which are not well suited to discovering more complex, nonlinear patterns. Here we introduce the Clustermatch Correlation Coefficient (CCC), an efficient, easy-to-use and not-only-linear coefficient based on machine learning. Clustermatch derives a similarity value between numerical and categorical features by applying clustering algorithms on objects. Applying Clustermatch to human gene expression data reveals both linear and biologically meaningful nonlinear gene-gene relationships. We show that Clustermatch significantly improves the detection of different types of relationships, it is robust to outliers, and its single parameter can easily balance between pattern complexity and computation time. Clustermatch is most similar to the previously described Maximal Information Coefficient, although our method runs in a fraction of time and can be practically applied to genome-scale data. We anticipate that Clustermatch will dramatically improve the detection of crucial molecular patterns completely missed by standard coefficients.

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 nonlinear patterns. However, they are impractical not only for big data but also for 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 nonlinear 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. Clustermatch 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]. Clustermatch captured both known linear relationships and novel nonlinear and biologically meaningful gene-gene patterns completely missed by standard coefficients. The CCC is most similar to MIC, although it is much faster to compute. We found that a combined analysis of linear-only coefficients and CCC can help highlight complex, yet promising relationships. 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

Clustermatch is a dependence coefficient that can compute a similarity measure between any pair of variables, either with numerical or categorical values [3]. The method assumes that if there is a relationship between two variables/features describing n data points/objects, then the clusterings of those n objects derived using each variable individually should match (see Methods). Although different clustering algorithms can be applied to one-dimensional data [27], we used quantiles to efficiently separate data points into different clusters (i.e., the median separates numerical data into two clusters). In Clustermatch, the data is internally categorized into clusters, therefore numerical and categorical variables can be naturally integrated since clusters do not need an order. Once all clusterings from each variable are generated, the Clustermatch coefficient is defined as the maximum adjusted Rand index (ARI) [28] between them. We previously compared Clustermatch [3] with Pearson, Spearman, and two nonlinear correlation coefficients: the Maximal Information Coefficient (MIC) [29] and Distance Correlation (DC) [25]. Clustermatch outperformed these two methods in a simulated scenario with several relationship types (linear and nonlinear) and noise levels. Clustermatch was also significantly superior in computational complexity, making it the only practical not-only-linear coefficient for real and large datasets such as gene expression compendia. This study focused on RNA-seq data from GTEx v8 and compared which patterns were detected or missed by these coefficients.

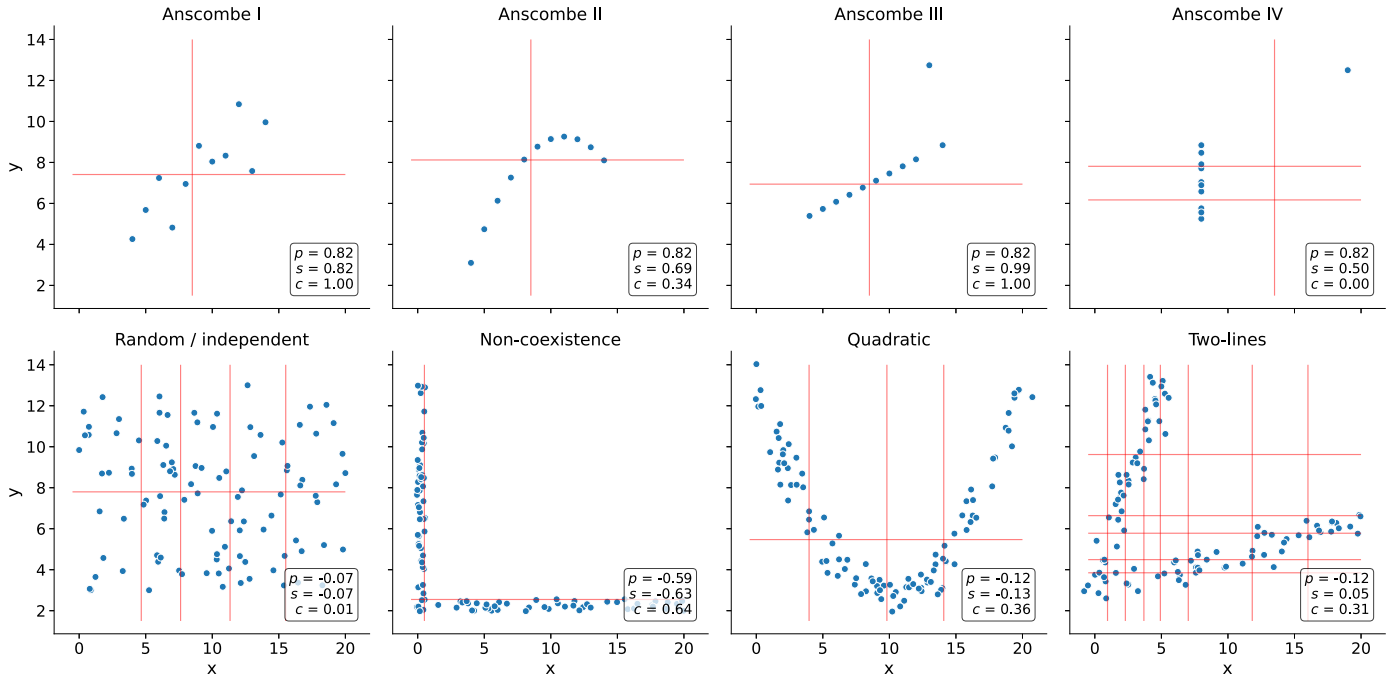

Figure 1: Different types of relationships in data. Each panel contains a set of simulated data points described by two variables: x and y. The first row shows Anscombe's quartet with four different datasets (from Anscombe I to IV) and 11 data points each. The second row contains a set of general patterns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

In Figure 1, we show how Pearson (p), Spearman (s) and Clustermatch (c) behave on different data patterns, where red lines indicate how Clustermatch clusters data points using each feature individually (either x or y). In the first row of the figure, the classic Anscombe's quartet [30] is shown, which comprises four synthetic datasets with entirely different patterns but the same data statistics (mean, standard deviation and Pearson's correlation). This kind of simulated data, recently revisited with the "Datasaurus" [31,32,33], are frequently 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 meaningfull nonlinear relationships) can be masked by summary statistics alone. For example, Anscombe I seems to show a noisy but clear linear pattern, similar to Anscombe III where the linearity is perfect besides one outlier. Here Clustermatch separates data points using two clusters (one red line for each variable x and y), yielding 1.0, the maximum value, correctly identifying the relationship. Anscombe II seems to follow a quadratic relationship interpreted as a linear pattern by Pearson and Spearman, whereas Clustermatch yields a lower yet non-zero value of 0.34. Anscombe IV shows a vertical line where x values are almost constant except for one outlier. This outlier does not influence Clustermatch 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 Clustermatch 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. The reason for this behavior is that these coefficients are based on data statistics such as the mean, standard deviation and, in the case of Spearman, data rankings, and these fall short in dealing with noisy data.

The second row of Figure 1 shows other simulated relationships with general nonlinear patterns, some of which were previously observed in gene expression data [29,34,35]. For the random/independent pair of variables, all coefficients correctly agree with a value close to zero. In this case, Clustermatch separates data points into different numbers of clusters which, when compared to each other, all give an ARI very close to zero (in fact, the maximum value $c = 0.01$, is reached with five clusters using x and two using y). For the other three examples (quadratic, non-coexistence and two-

lines), Pearson and Spearman generally fail to capture a clear pattern between variables x and y. These patterns also show how Clustermatch uses different degrees of complexity to capture the relationships. For the non-coexistence pattern where, for instance, one gene (x) might be expressed while the other one (y) is inhibited, Clustermatch only needs two clusters for both variables, similarly to a linear relationship (Anscombe I and III). For the quadratic pattern, Clustermatch 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 either Pearson or Spearman does not detect ($p = -0.12$ and $s = 0.05$, respectively). Here, Clustermatch 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 need for visualization before drawing conclusions on summary statistics alone. Although extra steps such as visual analyses are always necessary, larger datasets make it impossible to perform a manual assessment on each, for example, gene pair. More advanced yet interpretable techniques, such as Clustermatch, could reduce the number of false positives/negatives to focus human validation on patterns that are more likely to be real. Clustermatch has only one parameter: $k_{\max}$, the maximum number of internal clusters that the algorithm will use when partitioning data points. As we showed in the examples above, this parameter can control the level of complexity the end-user desires to capture. A value of $k_{\max} = 2$ makes the coefficient more leaned towards linear patterns, whereas higher values can detect other, more complex kinds of relationships. We found that $k_{\max} = 10$ (the default value) approximates well the coefficient values for different types of patterns [3] while balancing computing time and always keeping a close-to-zero value for random data, which is guaranteed by the adjusted-for-chance property of ARI [36].

The following sections compare the coefficients on real gene expression data and highlight some complex and biologically interesting relationships that only Clustermatch detects.

Clustermatch detects linear and nonlinear patterns in human transcriptomic data

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 pairwise correlation matrix using Pearson, Spearman and Clustermatch (see Methods).

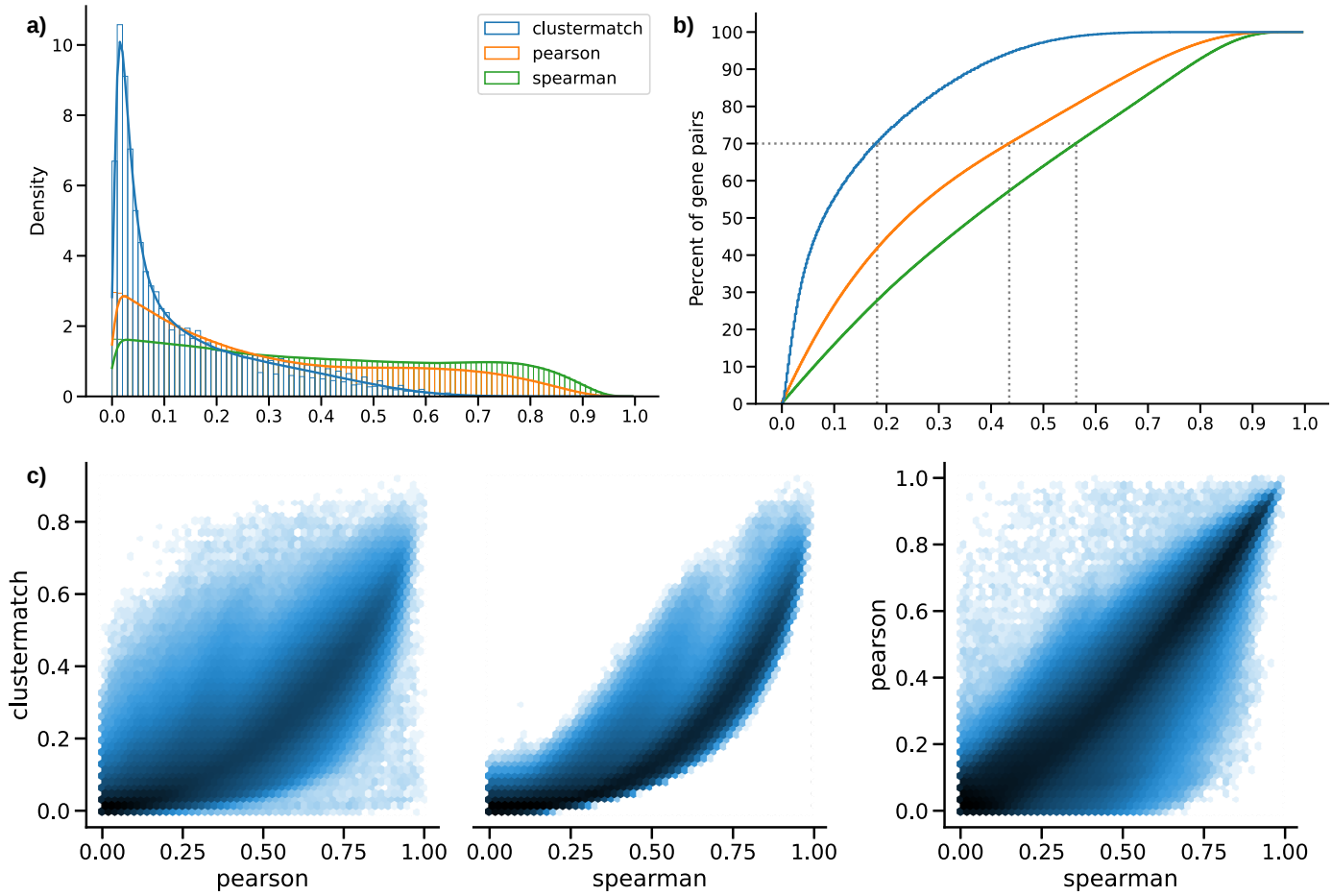

Figure 2: Distribution of coefficient values on gene expression (GTEx v8, whole blood). a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all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

In Figure 2 a, we show how the pairwise correlation values distribute, where Clustermatch (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 especially Spearman (mean=0.39, median=0.37, sd=0.26). Coefficients reach 70% of gene pairs at $c = 0.18$, $p = 0.44$ and $s = 0.56$ (Figure 2 b). Clustermatch and Spearman tend to agree more than any of these with Pearson, although many gene pairs seem to have a relatively higher correlation value only with Clustermatch (Figures 2 c).

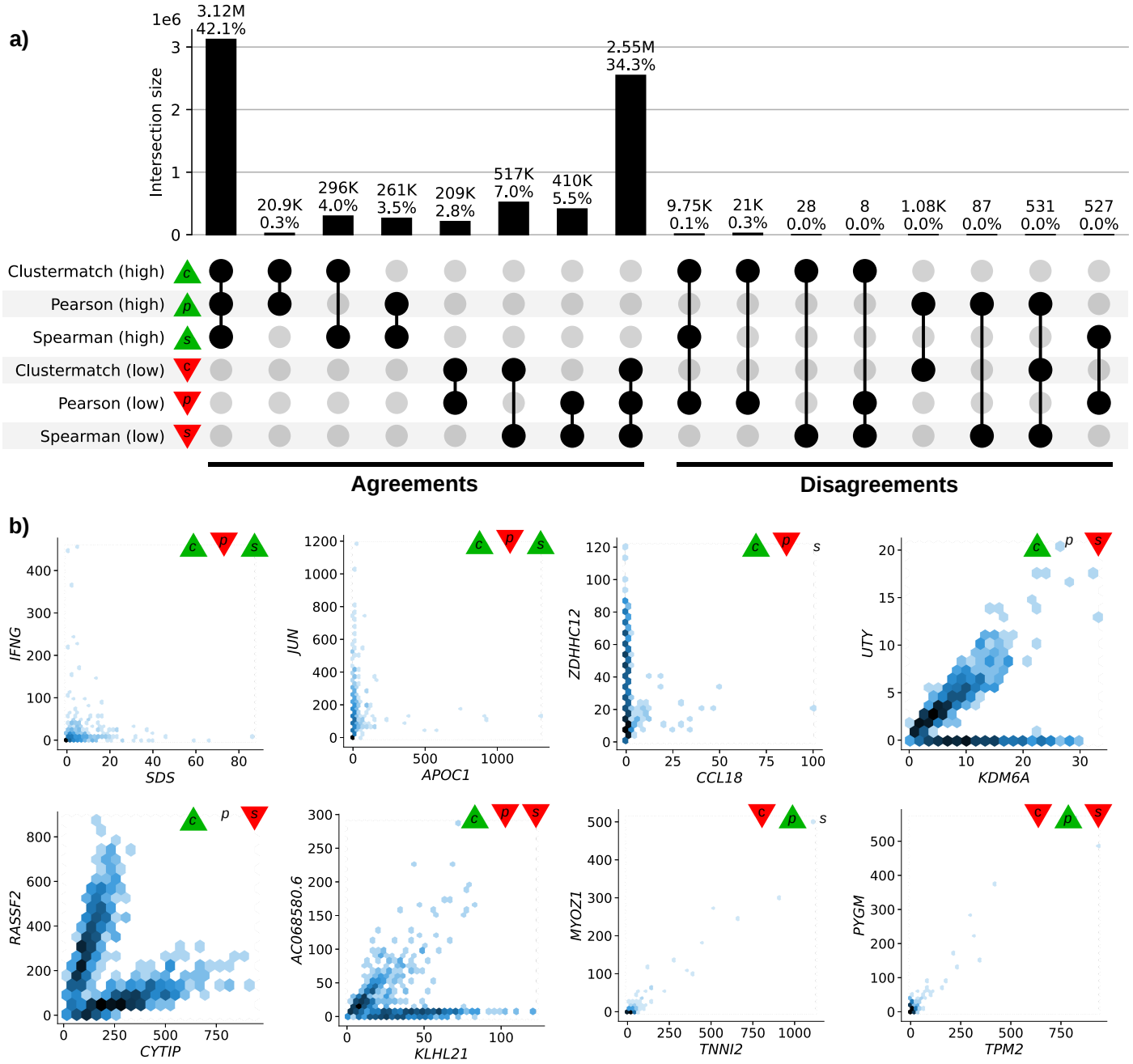

Figure 3: Intersection of gene pairs with high and low correlation coefficient values (GTEx v8, whole blood). a) UpSet plot with six categories (rows) grouping the 30% of the highest (green triangle) and lowest (red triangle) correlation values for each method. Columns show different intersections of categories grouped by agreements and disagreements. **b)** Hexagonal binning plots with examples of gene pairs where Clustermatch (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 correlation 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.

A closer inspection of gene pairs detected and missed by these coefficients revealed the ability of Clustermatch to capture more complex yet biologically meaningful 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 [37] 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. For most cases, the three coefficients agreed on whether there is a strong linear correlation (42.1%) and whether there is no relationship (34.3%). This result is crucial because it implies that the user will not miss important linear patterns in expression data when using Clustermatch. The figure also confirms that Clustermatch and Spearman agree more on highly 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”). Regarding disagreements (right part of the figure), more than 20 thousand gene pairs (20,987) with a high Clustermatch value are not highly ranked by any other coefficients. There are also gene pairs with a high Pearson value with either low Clustermatch (1,075) or low Clustermatch and low Spearman values (531). However, these cases mostly seem to be driven by outliers (Figure 3 b). Clustermatch does not miss gene pairs highly ranked by Spearman.

Figure 3 b shows individual gene pairs among the top five of each intersection category in the “Disagreements” group where Clustermatch disagrees with either Pearson, Spearman or both. The first three gene pairs at the top (*IFNG* / *SDS*, *JUN* / *APOC1*, and *ZDHHC12* / *CCL18*), with a high Clustermatch and low Pearson coefficient, 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 next three gene pairs (*UTY* / *KDM6A*, *RASSF2* / *CYTIP*, and *AC068580.6* / *KLHL21*) follow patterns combining either two linear or one linear and one constant relationships. In particular, genes *UTY* and *KDM6A*, paralogs, show a nonlinear relationship with a subset of samples following a robust linear pattern and another subset having 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 constant patterns is captured by Clustermatch ($c = 0.29$, 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 Clustermatch in all other tissues in GTEx, except for female-specific organs (Figure 4) as expected.

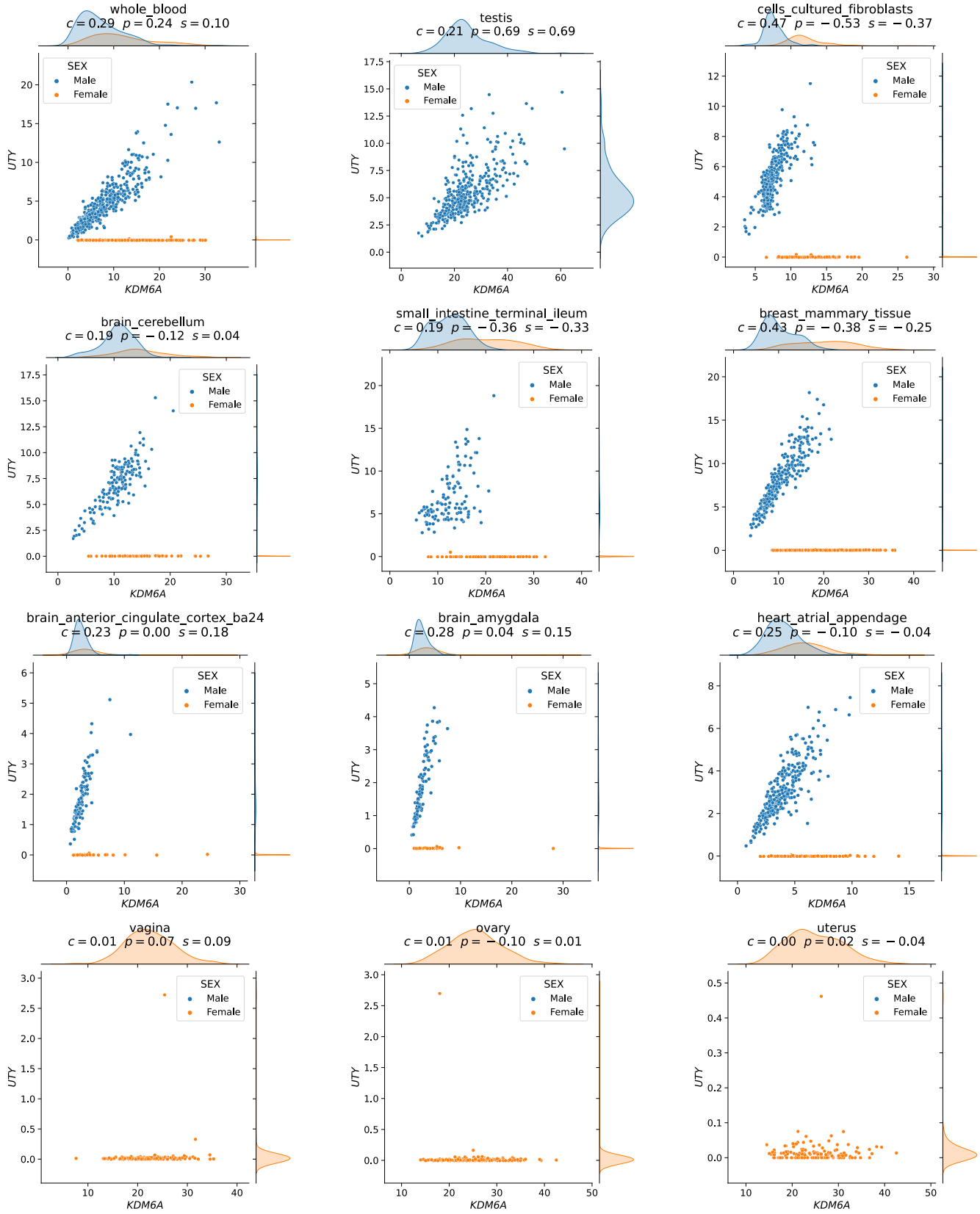

Figure 4: Scatter plots of genes *KDM6A* and *UTY* across different GTEx tissues. Clustermatch correctly captures the relationship in all GTEx tissues, and here we show nine of them in the first three rows. The last row shows three female-specific organs, where Clustermatch correctly finds no association.

To study the other gene pairs found by the correlation coefficients, we used tissue-specific gene networks from GIANT [38], where nodes represent genes and each edge a functional relationship weighted with a probability of interaction between two genes. GIANT networks were built from 987 genome-scale data sets across approximately 38,000 conditions, including expression and different interaction measurements such as gene co-expression (using Pearson correlation), protein-interaction, transcription factor regulation, and chemical and genetic perturbations and microRNA

target profiles from the Molecular Signatures Database (MSigDB [39]). Figure 5 shows cell type-specific networks for each gene pair (Figure 3 b) for which genes are present in GIANT models. 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. The tissue was automatically selected in each network according to the predicted tissue expression of gene pairs. Interestingly, gene pairs highly ranked by Clustermatch are part of very cohesive networks and strongly related to blood. For example, the average probability of gene connections with *IFNG* / *SDS* is very high, at least 0.79 for all other genes shown. This minimum average with *JUN* / *APOC1* is 0.56, for *ZDHHC12* / *CCL18* is 0.34 (where *ZDHHC12* shows the weakest links although *CCL18* is strongly connected), and for *RASSF2* / *CYTIP* is 0.76. Predicted networks for the two gene pairs prioritized by Pearson are much less cohesive, suggesting that the high correlation in whole blood is mostly driven by outliers. For example, the minimum/maximum average of interaction probabilities with *MYOZ1* / *TNNI2* is only 0.11/0.12, and for *PYGM* / *TPM2* is 0.13/0.24.

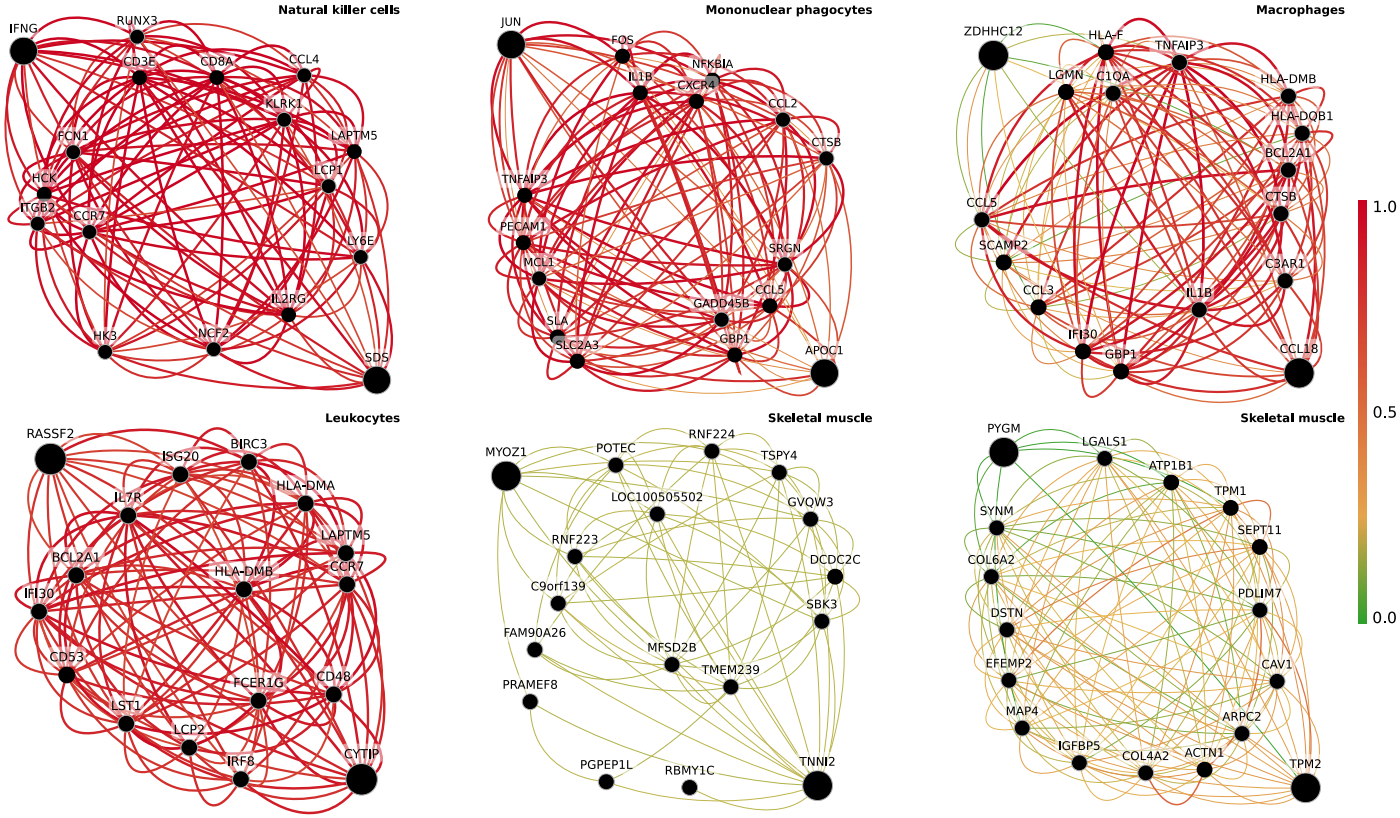

Figure 5: Predicted tissue-specific networks from GIANT for each gene pairs prioritized by correlation coefficients. A node represents a gene and an edge the probability that two genes are part of the same biological process in a specific cell type (indicated at the top-right corner). A maximum of 15 genes are shown for each subfigure. The GIANT web application automatically determined a minimum weight for edges to be shown. These analyses can be performed online using the following links: *IFNG* / *SDS* [40]; *JUN* / *APOC1* [41] *ZDHHC12* / *CCL18* [42] *RASSF2* / *CYTIP* [43] *MYOZ1* / *TNNI2* [44] *PYGM* / *TPM2* [45]

Some of these genes, such as *SDS* (12q24) and *ZDHHC12* (9q34), were previously found to have a relatively lower number of publications that correlated well with a small set of chemical, physical and biological features [46]. 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 Clustermatch correlation and are entirely missed by the other methods. *KLHL21* was suggested as a potential therapeutic target for hepatocellular carcinoma [47] and other cancers [48,49], and its nonlinear correlation with *AC068580.6* and potentially other genes might unveil other important players in cancer initiation or progression.

Clustermatch and Maximal Information Coefficient strongly agree on gene pair prioritization

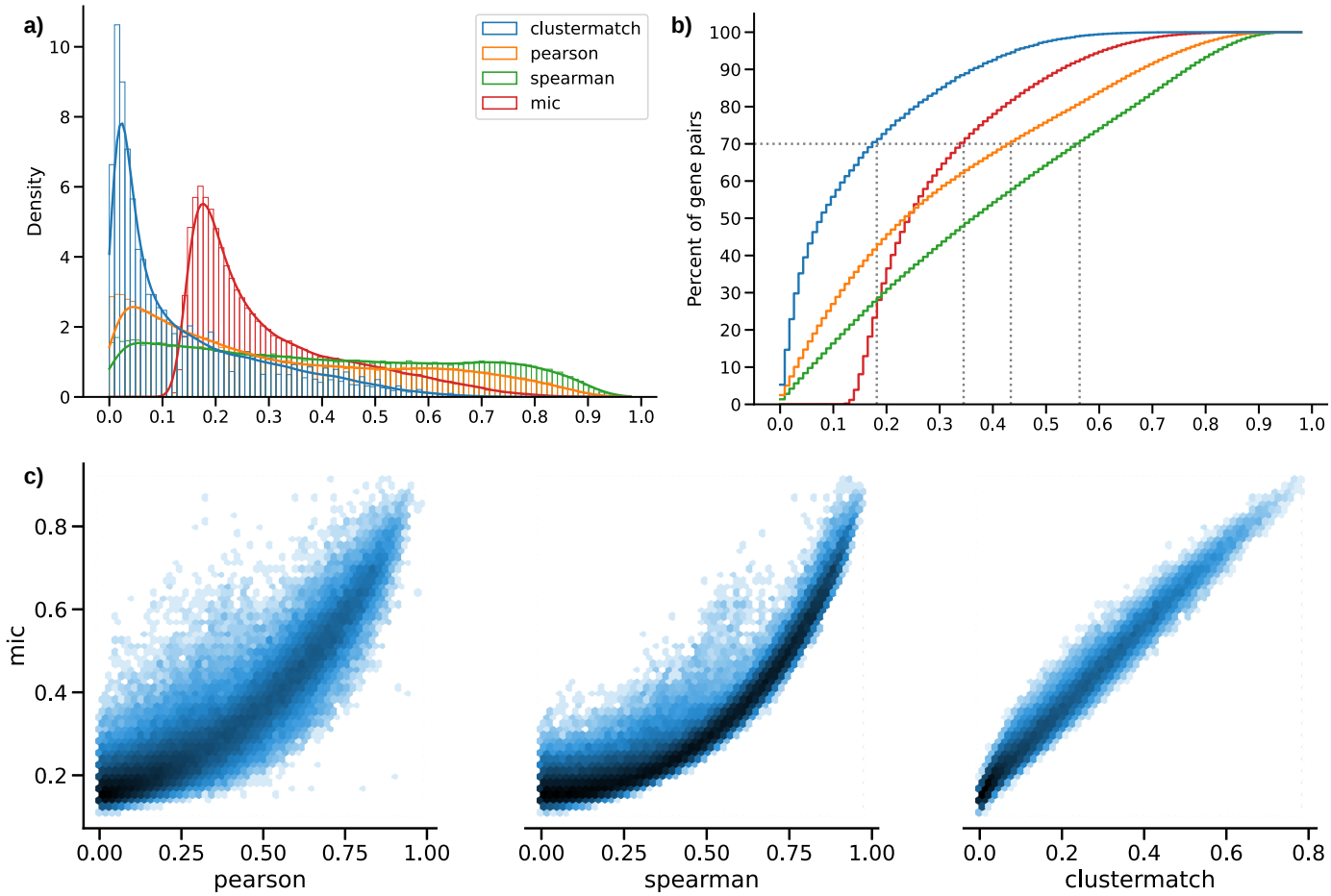

Figure 6: Distribution of Maximal Information Coefficient (MIC) and comparison with other methods. Given the high time complexity of MIC, approximately 1% of gene pairs were sampled from our previous set of 5,000 genes (GTEx v8, whole blood). **a)** Histogram of coefficient values. **b)** Corresponding 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.

Finally, we compared all the coefficients with the Maximal Information Coefficient (MIC [24]), a popular nonlinear method that can find complex relationships in data, although very expensive in computational terms. 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 same analysis on the distribution of coefficients, shown in Figure 6. We verified that Clustermatch and MIC behave similarly in this dataset, with essentially the same distribution but only shifted. Figure 6 c shows that these two coefficients relate almost linearly and compare very similarly with Pearson and Spearman. This result is relevant because MIC represented a significant step forward in correlation analysis research, and it has been successfully used in various application domains [4,50,51]. However, the use of MIC in large datasets remains limited due to its very long computation time. Our work and analyses suggest that Clustermatch could be an equally effective but much more efficient next-generation correlation coefficient.

Discussion

We previously showed that Clustermatch outperformed all other coefficients in a simulated cluster analysis scenario with linear and nonlinear patterns and varying noise levels. Here we introduced the Clustermatch correlation coefficient, an efficient machine learning-based method and an optimized

Python implementation. 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. Clustermatch also captured complex and biologically meaningful relationships completely missed by standard coefficients. We also showed that directly comparing Clustermatch with linear-only coefficients highlighted the most complex and potentially promising gene pairs. Finally, Clustermatch derives scores very well aligned with the Maximal Information Coefficient while being much more computationally efficient and thus practical for use in large modern datasets.

It is well-known that biomedical research is biased towards a small fraction of human genes [52,53]. 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 [46], such as RNA and protein abundance or gene length. Another potential explanation could also be a bias towards linear-only statistical methods. Recent computational approaches such as deep learning have revolutionized several areas in academia and industry. However, their interpretability, which is essential in biology and medicine, is still a significant limitation. In these health-related fields, another breakthrough theoretical advance was the omnigenic model of complex traits, where gene regulatory networks are becoming first-class players in genetic studies. This model shifted our attention to other, potentially less studied genes that are part of disease-relevant networks while also being helpful to explain why polygenic risk scores perform so poorly across different population ancestries [22]. The streetlight effect [54], where we only search where it is easiest to look, is widespread in areas beyond biology. We anticipate that advanced, efficient and interpretable approaches like Clustermatch will play a significant role in 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 for Clustermatch. Although Clustermatch is much faster than MIC or DC, Pearson and Spearman are still the most efficient since they only rely on simple data statistics, which significantly limits their ability beyond linear patterns and are susceptible to outliers. Even with this small sample of genes, our results confirm that the advantages of using more sophisticated yet efficient methods like Clustermatch can help detect and study more intricate molecular mechanisms. Although we only used GTEx, with a relatively homogeneous set of samples, we could still find complex and meaningful patterns, suggesting that the application of Clustermatch on larger compendia, such as recount3 [11] with thousands of samples across different conditions, can reveal other potentially meaningful gene interactions.

Computing a correlation coefficient based on simple data summaries is tempting fast. However, this study shows that those methods can miss crucial, not-only-linear patterns to understand the big picture. We provide an efficient, next-generation correlation coefficient based on machine learning techniques that can process heterogeneous data types seamlessly, dramatically easing preprocessing steps for the end-user.

Methods

Clustermatch algorithm

Algorithm 1: Clustermatch algorithm

Function `get_partitions`($\mathbf{v}$, $k_{\max}$):
  Output:
    Ω_r: clustering with r clusters over n objects
  if $\mathbf{v} \in \mathbb{R}^n$ **then**
    for $r \leftarrow 2$ **to** $\min\{k_{\max}, |\mathbf{v}| - 1\}$ **do**
      $\boldsymbol{\rho} \leftarrow (\rho_\ell \mid \Pr(v_i < \rho_\ell) \leq (\ell - 1)/r), \forall \ell \in [1, r+1]$
      $\Omega_{r\ell} \leftarrow \{i \mid \rho_\ell < v_i \leq \rho_{\ell+1}\}, \forall \ell \in [1, r]$
  else
    $\mathcal{C} \leftarrow \cup_j \{v_i\}$
    $r \leftarrow |\mathcal{C}|$
    $\Omega_{rc} \leftarrow \{i \mid v_i = \mathcal{C}_c\}, \forall c \in [1, r]$
  `// Remove singleton partitions`
  $\Omega \leftarrow \{\Omega_r \mid |\Omega_r| > 1\}, \forall r$
  return Ω

Function `clustermatch`($\mathbf{x}$, $\mathbf{y}$, $k_{\max}$):
  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 $\mathbf{y}$ ($c \in [0, 1]$)
  $\Omega^{\mathbf{x}} = \text{get\_partitions}(\mathbf{x}, k_{\max})$
  $\Omega^{\mathbf{y}} = \text{get\_partitions}(\mathbf{y}, k_{\max})$
  $c \leftarrow \max\{\mathcal{A}(\Omega^{\mathbf{x}}_p, \Omega^{\mathbf{y}}_q)\}, \forall p, q$
  return c

An optimized Python implementation of Clustermatch can be found in our Github repository [55], as well as the code and data needed to reproduce all analyses.

GTEx v8 data and sampling approach

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 Clustermatch on these 5,000 genes across all 755 samples, 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 this set of 5,000 genes ($n * (n - 1)/2 = 12497500$).

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. **The Data Model Concept in Statistical mapping**

George F Jenks
International Yearbook of Cartography (1967)

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

29. **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

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

31. **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

32. **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

33. **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

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. **Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance**
Nguyen Xuan Vinh, Julien Epps, James Bailey
Journal of Machine Learning Research (2010) http://www.jmlr.org/papers/v11/vinh10a.html

37. **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

38. **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

39. **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

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

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

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

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

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

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

46. **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

47. **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

48. **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

49. **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

50. **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

51. **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

52. **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

53. **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

54. **Why Scientific Studies Are So Often Wrong: The Streetlight Effect**
Discover Magazine
https://www.discovermagazine.com/the-sciences/why-scientific-studies-are-so-often-wrong-the-streetlight-effect

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

Acknowledgements