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

This manuscript (permalink) was automatically generated from greenelab/clustermatch-gene-expr-manuscript@83e12ce on February 7, 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

Introduction

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 [1]. The method assumes that if there is a relationship between two variables/features describing n data points/objects, then the clusterings on those n objects derived using each variable individually should match (see Methods). Although different clustering algorithms can be applied to one-dimensional data [2], we used quantiles to efficiently separate data points into different clusters (i.e., the median separates numerical data into two clusters). Since in Clustermatch the data is categorized into clusters, numerical and categorical data 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) [3] between them. We previously compared Clustermatch [1] with the Maximal Information Coefficient (MIC) [4] and Distance Correlation (DC) [5], two popular nonlinear correlation coefficients. 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.

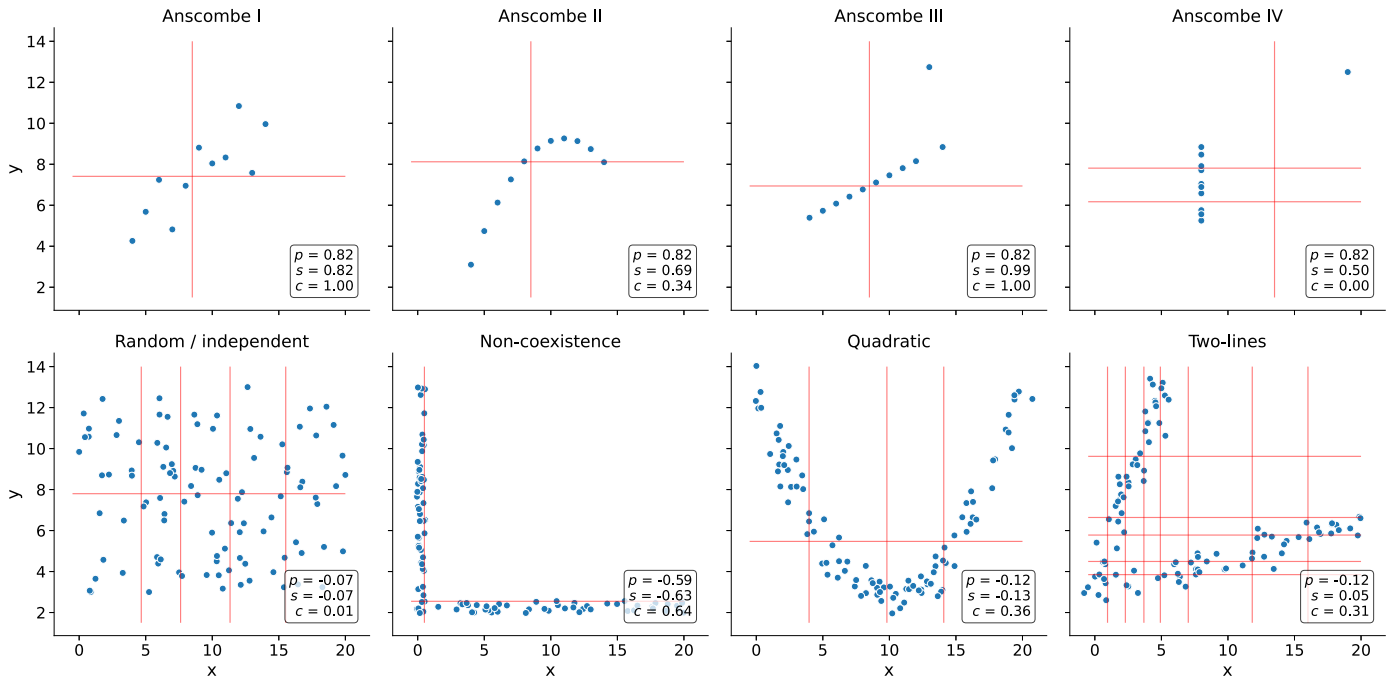

Figure 1: Different types of relationships in data. Each panel contains a set of simulated data points described by two variables: x and y. The first row shows Anscombe's quartet with four different datasets (from Anscombe I to IV) and 11 data points each. The second row contains a set of general patterns with 100 data points each. Each panel shows the correlation value using the Pearson (p), Spearman (s) and Clustermatch (c) coefficients. 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 [6] 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" [7,8,9], 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 nonlinear relationships reflecting real and complex biological relationships) can be masked by summary statistics. 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. For these two patterns, 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 [4,10,11]. 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 [1] 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 [12].

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

Clustermatch detects linear and nonlinear patterns in human transcriptomic data

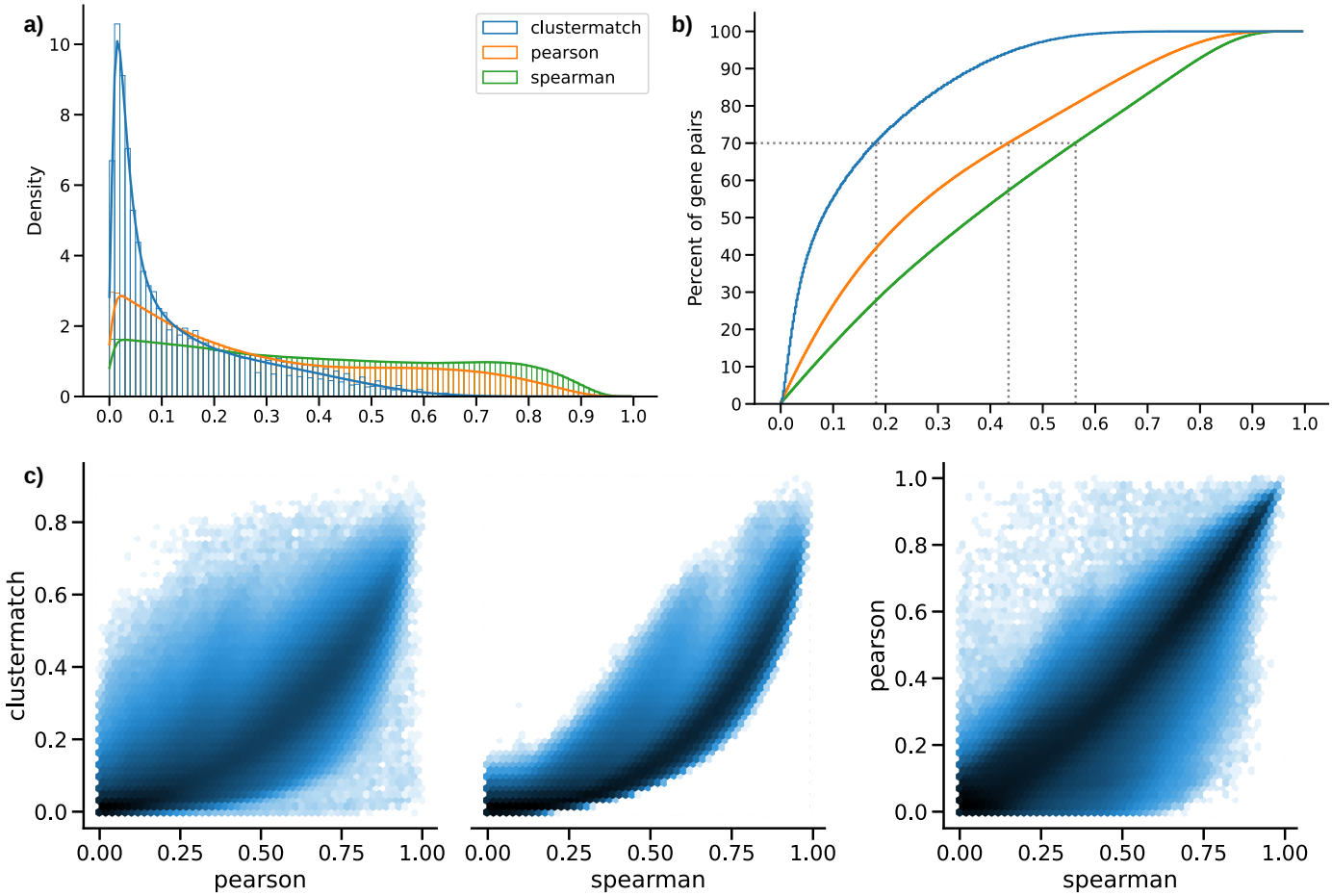

Figure 2: Distribution of coefficient values on gene expression (GTEx v8, whole blood). a) Histogram of coefficient values. **b)** Corresponding cumulative histogram. The dotted line maps the coefficient value that accumulates 70% of gene pairs. **c)** 2D histogram plot with hexagonal bins between all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

We 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. 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 value for Clustermatch (Figures 2 c).

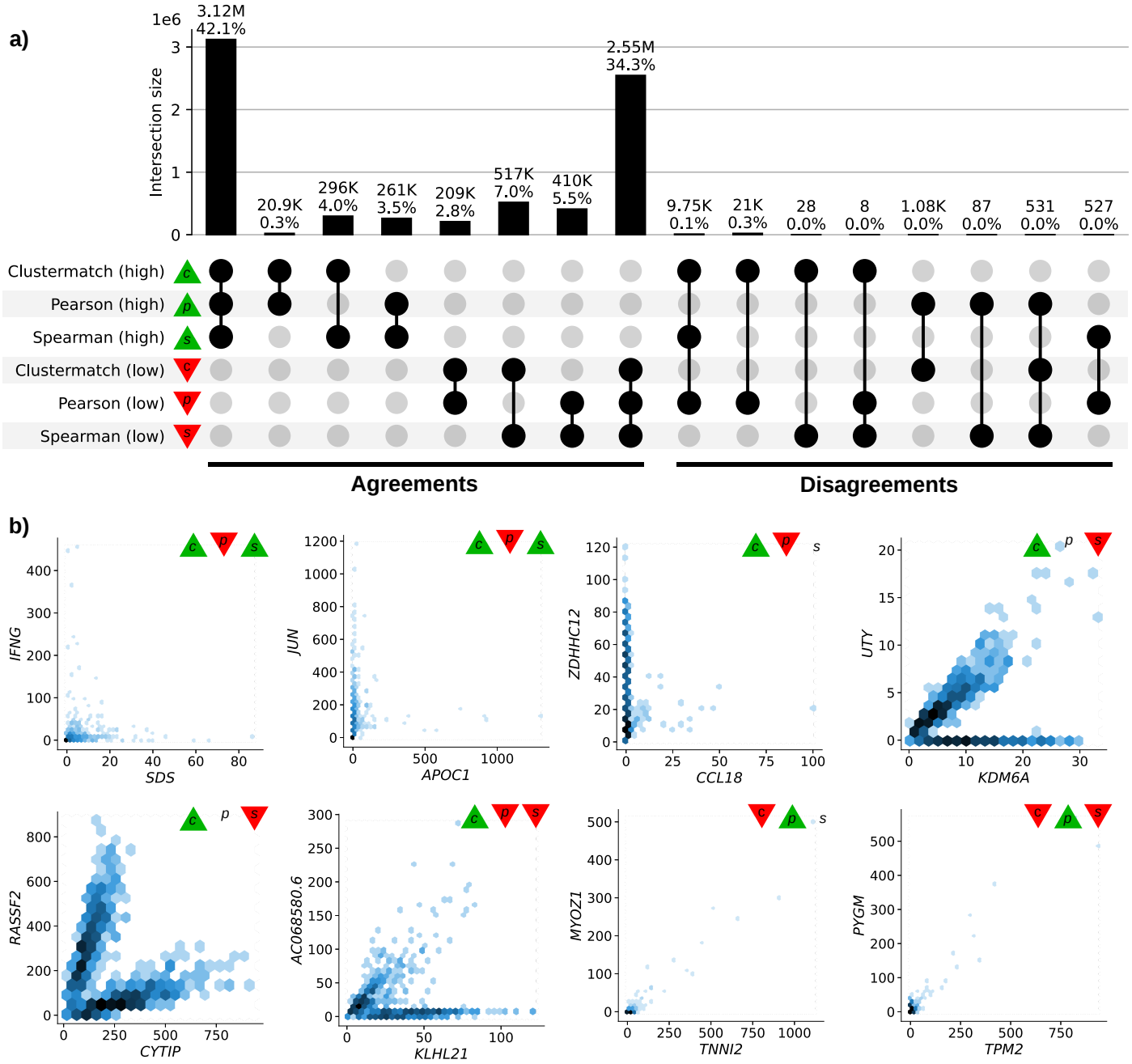

Figure 3: Intersection of gene pairs with high and low coefficient values (GTEx v8, whole blood). a) UpSet plot with six categories (rows) grouping the 30% of the highest (green triangle) and lowest (red triangle) correlation values for each method. Columns show different intersections of categories 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 obtained a correlation 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 [13] 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 agree on whether there is a strong linear correlation (42.1%) and whether there is no relationship (34.3%). This 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, 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, and 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).

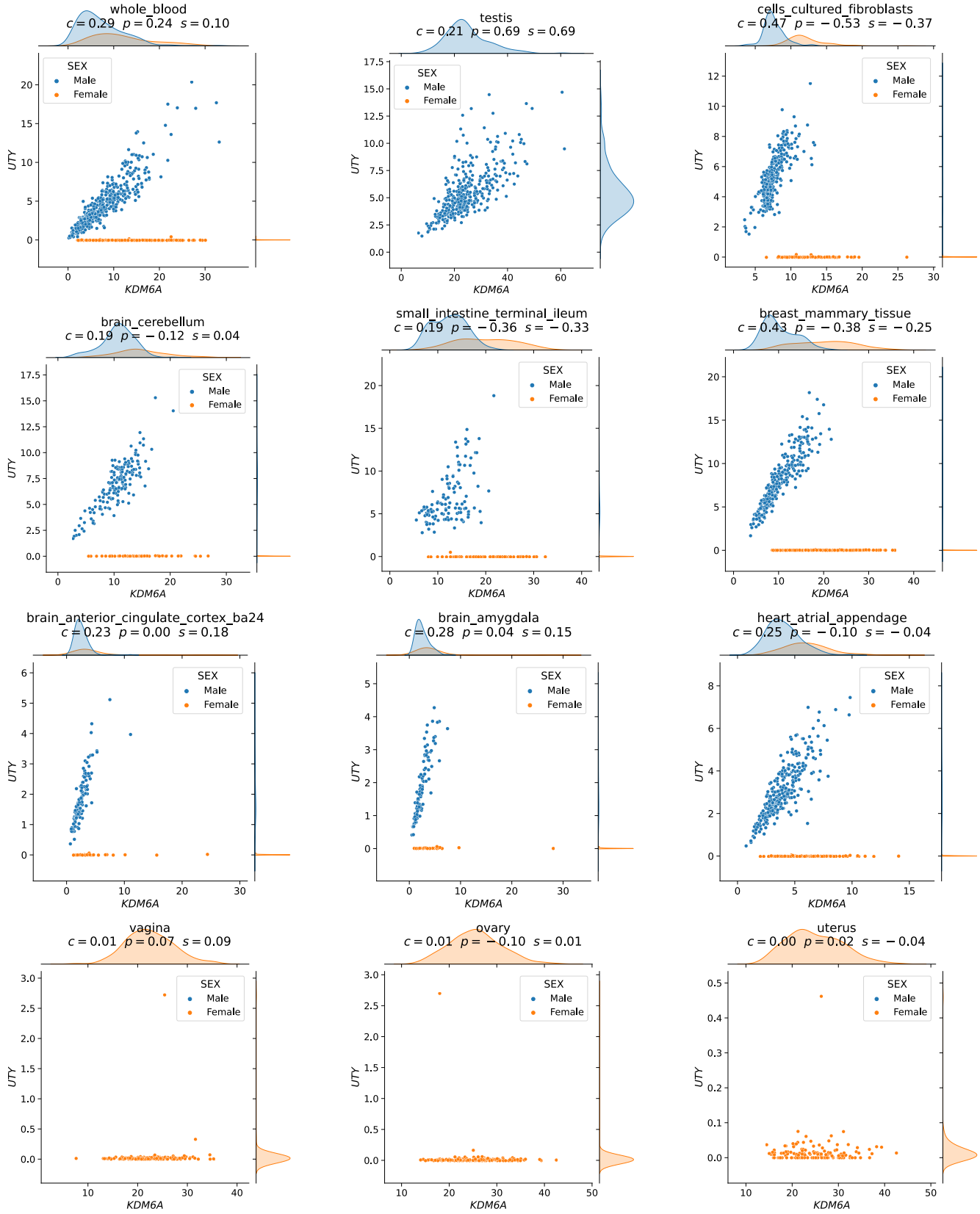

Figure 4: Scatter plots of genes *KDM6A* and *UTY* across different GTEx tissues. Clustermatch correctly captures the relationship in all GTEx tissues, and here we show nine of them in the first three rows. The last row shows three female-specific organs, where Clustermatch correctly finds no association.

To study the other gene pairs found by the correlation coefficients, we used tissue-specific gene networks from GIANT [14], 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 [15]). Figure 5 shows blood-specific networks for each gene pair (Figure 3 b) for which genes are present in GIANT models. Two large black nodes in the 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. Interestingly, gene pairs highly ranked by Clustermatch are part of very cohesive networks. 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 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.

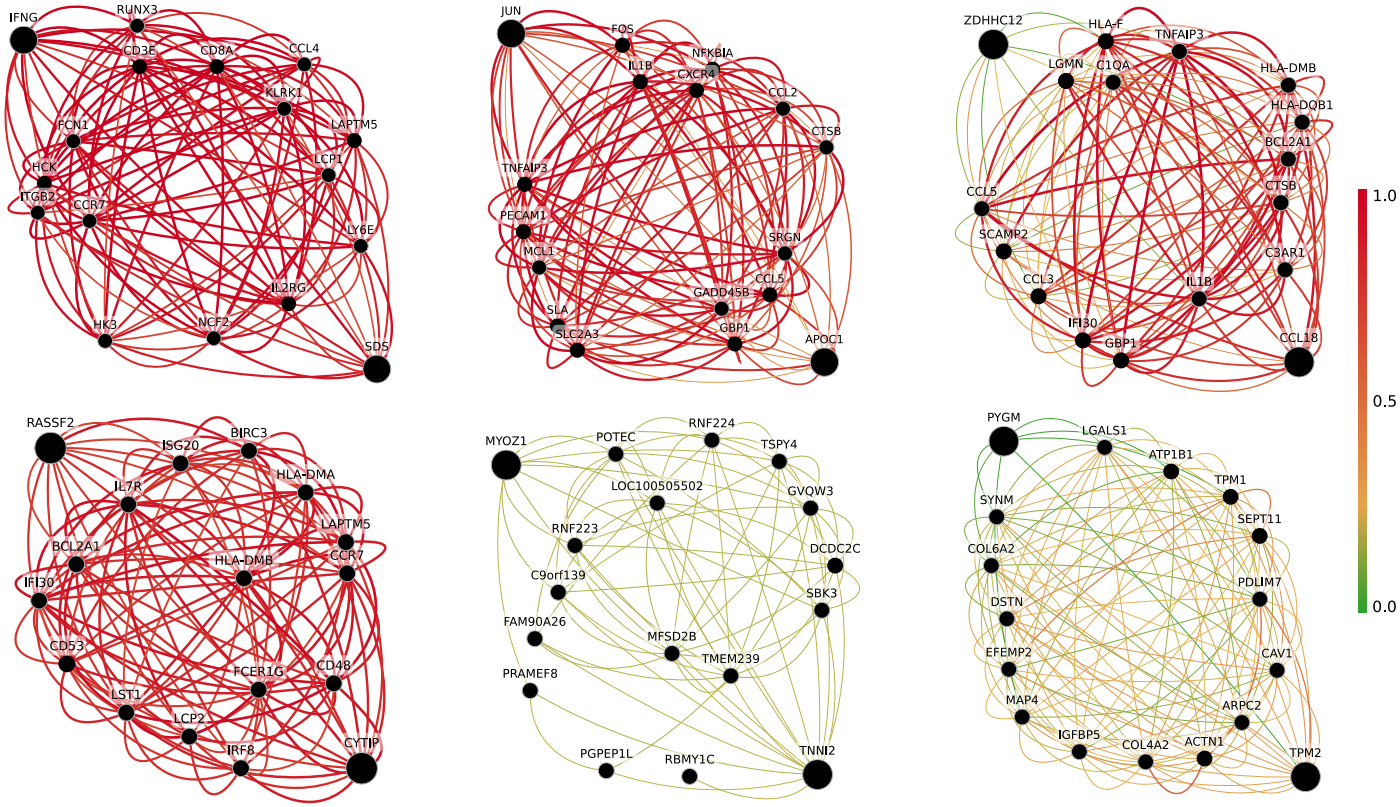

Figure 5: Predicted tissue-specific networks from GIANT for each gene pairs prioritized by correlation coefficients. A node represents a gene and an edge the probability that two genes are part of the same biological process in a blood-related cell type (indicated at the top of each subfigure). A maximum 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* [16]; *JUN* / *APOC1* [17] *ZDHHC12* / *CCL18* [18] *RASSF2* / *CYTIP* [19] *MYOZ1* / *TNNI2* [20] *PYGM* / *TPM2* [21]

Discussion

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

    `// TODO: not implemented yet in optimized version`

    $\mathcal{C} \leftarrow \cup_j \{v_i\}$

    $r \leftarrow |\mathcal{C}|$

    $\Omega_{rc} \leftarrow \{i \mid v_i = \mathcal{C}_c\}, \forall c \in [1, r]$

  `// TODO: remove singletons`

  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

References

1. **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

2. **The Data Model Concept in Statistical mapping**
George F Jenks
International Yearbook of Cartography (1967)

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

4. **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

5. **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

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

7. **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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

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

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

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

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

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

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

Acknowledgements

Supplementary material